NBER WORKING PAPER SERIES

TOO MUCH DATA: PRICES AND INEFFICIENCIES IN DATA MARKETS

Daron Acemoglu
Ali Makhdoumi
Azarakhsh Malekian
Asuman Ozdaglar

Working Paper 26296
http://www.nber.org/papers/w26296

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2019

We are grateful to Alessandro Bonatti and Hal Varian for useful conversations and comments. We gratefully acknowledge financial support from Google, Microsoft, the National Science Foundation, and the Toulouse Network on Information Technology. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Too Much Data: Prices and Inefficiencies in Data Markets
Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar
NBER Working Paper No. 26296
September 2019
JEL No. D62,D83,L86

ABSTRACT

When a user shares her data with an online platform, she typically reveals relevant information about other users. We model a data market in the presence of this type of externality in a setup where one or multiple platforms estimate a user's type with data they acquire from all users and (some) users value their privacy. We demonstrate that the data externalities depress the price of data because once a user's information is leaked by others, she has less reason to protect her data and privacy. These depressed prices lead to excessive data sharing. We characterize conditions under which shutting down data markets improves (utilitarian) welfare. Competition between platforms does not redress the problem of excessively low price for data and too much data sharing, and may further reduce welfare. We propose a scheme based on mediated data-sharing that improves efficiency.

Daron Acemoglu
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142-1347
and CIFAR
and also NBER
daron@mit.edu

Ali Makhdoumi
Duke University
100 Fuqua Drive
Durham, NC 27708
United States
ali.makhdoumi@duke.edu

Azarakhsh Malekian
Rotman School of Management
University of Toronto
and LIDS, MIT
azarakhsh.malekian@rotman.utoronto.ca

Asuman Ozdaglar
Dept of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
77 Massachusetts Ave, E40-130
Cambridge, MA 02139
asuman@mit.edu

1 Introduction

The data of billions of individuals are currently being utilized for personalized advertising or other online services.[1] The use and transaction of individual data are set to grow exponentially in the coming years with more extensive data collection from new online apps and integrated technologies such as the Internet of Things and with the more widespread applications of artificial intelligence (AI) and machine learning techniques. Most economic analysis emphasizes benefits from the use and sharing of data because this permits better customization, better information, and more input into AI applications. It is often claimed that because data enable a better allocation of resources and more or higher quality innovation, the market mechanism generates too little data sharing (e.g., Varian [2009], Jones and Tonetti [2018], Veldkamp et al. [2019], and Veldkamp [2019]). Economists have recognized that consumers might have privacy concerns (e.g., Stigler [1980], Posner [1981], and Varian [2009]), but have often argued that data markets could appropriately balance privacy concerns and the social benefits of data (e.g., Laudon [1996] and Posner and Weyl [2018]). In any case, the willingness of the majority of users to allow their data to be used for no or very little direct benefits is argued to be evidence that most users place only a small value on privacy.[2]

This paper, in contrast, argues that there are major forces that will make individual-level data underpriced and the market economy generate too much data. The reason is simple: when an individual shares her data, she compromises not only her own privacy but also the privacy of other individuals whose information is correlated with hers. This negative externality tends to create excessive data sharing. Moreover, when there is excessive data sharing, each individual will overlook her privacy concerns and part with her own information because others' sharing decisions will have already revealed much about her.

Some of the issues we emphasize are highlighted by the *Cambridge Analytica* scandal. The company acquired the private information of millions of individuals from data shared by 270,000 Facebook users who voluntarily downloaded an app for mapping their personality traits, called "This is your digital life". The app accessed users' news feed, timeline, posts and messages, and revealed information about other Facebook users. *Cambridge Analytica* was ultimately able to infer valuable information about more than 50 million Facebook users, which it deployed for designing personalized political messages and advertising in the Brexit referendum and 2016 US presidential election.[3] Though some of the circumstances of this scandal are unique, the issues are general. For example, when an individual shares information about his own behavior, habits and preferences, this contains considerable information about the behavior, habits and preferences of not only his friends but also other people with similar characteristics (e.g., the routines and choices of a highly-

[1] Just Facebook has almost 2.5 billion monthly (active) individual users.

[2] Consumers often report valuing privacy (e.g., Westin [1968]; Goldfarb and Tucker [2012]), but do not take much action to protect their privacy (e.g., "Why your inbox is crammed full of privacy policies", WIRED, May 24, 2018 and Athey et al. [2017]).

[3] See New York Times, March, 19 2018 and New York Times, March, 13, 2018, and The Guardian, April, 13, 2018

educated gay from Central America in his early 20s in Somerville, Massachusetts is informative about others with the same profile and residing in the same area).

The following example illustrates the nature of the problem, introduces some of our key concepts, and clarifies why there will be excessive data sharing and very little willingness to protect privacy on the part of users.

Consider a platform with two users, $i = 1, 2$. Each user owns her own personal data, which we represent with a random variable X_i (from the viewpoint of the platform). The relevant data of the two users are related, which we capture by assuming that their random variables are jointly normally distributed with mean zero and correlation coefficient ρ. The platform can acquire or buy the data of a user in order to better estimate her preferences or actions. Its objective is to minimize the mean square error of its estimates of user types, or maximize the amount of *leaked information* about them. Suppose that the valuation (in monetary terms) of the platform for the users' leaked information is one, while the value that the first user attaches to her privacy, again in terms of leaked information about her, is $1/2$ and for the second user it is $v > 0$. We also assume that the platform makes take-it-or-leave-it offers to the users to purchase their data. In the absence of any restrictions on data markets or transaction costs, the first user will always sell her data (because her valuation of privacy, $1/2$, is less than the value of information to the platform, 1). But given the correlation between the types of the two users, this implies that the platform will already have a fairly good estimate of the second user's information. Suppose, for illustration, that $\rho \approx 1$. In this case, the platform will know almost everything relevant about user 2 from user 1's data, and this undermines the willingness of user 2 to protect her data. In fact, since user 1 is revealing almost everything about her, she would be willing to sell her own data for a very low price (approximately 0 given $\rho \approx 1$). But once the second user is selling her data, this also reveals the first user's data, so the first user can only charge a very low price for her data. Therefore in this simple example, the platform will be able to acquire both users' data at approximately zero price. Critically, however, this price does not reflect the users' valuation of privacy. When $v \leq 1$, the equilibrium is efficient because data are socially beneficial in this case (even if data externalities change the distribution of economic surplus between the platform and users). However, it can be arbitrarily inefficient when v is sufficiently high. This is because the first user, by selling her data, is creating a negative externality on the second user.

This simple example captures the most important economic ideas of the paper. Our formal analysis generalizes these insights by considering a community of users whose information are correlated in an arbitrary fashion and who have heterogeneous valuations of privacy. Finally, we analyze this model both under a monopoly platform and under competition between platforms trying to simultaneously attract users and acquire their data.

Our main results correspond to generalizations of the insights summarized by the preceding example. First, we introduce our general framework and characterize the first-best allocation which maximizes the sum of surplus of users and platforms. The first best typically involves considerable data transactions, but those individuals creating significant (negative) externalities on others should not share their data. Second, we establish the existence of an equilibrium and characterize the prices at which data will be transacted. This characterization clarifies how the market price of data for a user and the distribution of surplus depend on information leaked by other users. Third and more importantly, we provide conditions under which the equilibrium in the data market is inefficient as well as conditions for simple restrictions on markets to improve welfare. At the root of these inefficiencies are the economic forces already highlighted by our example: inefficiencies arise when a subset of users are willing to part with their data, which are informative about other users whose value of privacy is high. We show that these insights extend to environments with competing platforms and incomplete information as well.

We further investigate various policy approaches to data markets. Person-specific taxes on data transactions can restore the first best, but are impractical. We show in addition how uniform taxation on all data transactions might, under some conditions, improve welfare. Finally, we propose a new regulation scheme where data transactions are mediated in a way that reduces their correlation with the data of other users, thus minimizing leaked information about others. We additionally develop a procedure for implementing this scheme based on "*de-correlation*", meaning transforming users' data so that their correlation with others' data and types is removed.[4]

Our paper relates to the literature on privacy and its legal and economic aspects. The classic definition of privacy, proposed by justices Warren and Brandeis in 1890, is the protection of someone's personal space and the right to be let alone (Warren and Brandeis [1890]). Relatedly, and more relevant to our focus, Westin [1968] defines it as the control over and safeguarding of personal information, and this perspective has been explored from various angles in recent work (e.g., Pasquale [2015], Tirole [2019], Zuboff [2019]). More closely related to our paper are MacCarthy [2010], Fairfield and Engel [2015], Choi et al. [2019], and Bergemann et al. [2019]. MacCarthy [2010] and Fairfield and Engel [2015] are the first contributions we are aware of that emphasize externalities in data sharing, which play a central role in our analysis. More recently, Choi et al. [2019] develop a reduced-form model with a related informational externality and a number of results similar to our excessive information sharing finding. There are several important differences between this paper and ours, however. First, Choi et al. [2019] do not model how data sharing creates an externality and simply assume that consumer welfare depends negatively on the number of other consumers on an online platform. In contrast, much of our analysis is devoted to the study of how the correlation structure across different users jointly determines sharing decisions and the amount of leaked information. Second, they assume that consumers are identical, while our above example already illustrates that heterogeneity in privacy concerns plays a critical role in the inefficiencies in data markets. Indeed, our analysis highlights that there are only limited inefficiencies when all users are homogeneous (specifically, the equilibrium is efficient in this case

[4]This de-correlation procedure is different from anonymization of data because it does not hide information about the user sharing her data but about others who are correlated with this user.

when they have low or sufficiently high value of privacy). Third, their paper does not analyze the case with competing platforms. Finally, Bergemann et al. [2019] also study an environment with data externalities. Though there are some parallels between the two papers, their work is different from and largely complementary to ours. In particular, they analyze an economy with symmetric users where there is a monopolist platform and data are used by this monopolist or other downstream firms (such as advertisers) for price discrimination. They focus on the implications for market prices, profits, and efficiency of the structure of the downstream market and whether data are collected in an anonymized or non-anonymized form.

Our paper also relates to the growing literature on information markets. One branch of this literature focuses on the use of personal data for improved allocation of online resources (e.g., Bergemann and Bonatti [2015], Goldfarb and Tucker [2011], and Montes et al. [2018]). Another branch investigates how information can be monetized either by dynamic sales or optimal mechanisms (e.g., Anton and Yao [2002], Babaioff et al. [2012], Eső and Szentes [2007], Horner and Skrzypacz [2016], and Bergemann et al. [2018], and Admati and Pfleiderer [1986] and Begenau et al. [2018] for markets on financial data). A third branch focuses on optimal collection and acquisition of information, for example, Agarwal et al. [2018], Chen and Zheng [2018], and Chen et al. [2018]. Lastly, a number of papers investigate the question of whether information harms consumers, either because users are unaware of the data being collected about them (Taylor [2004]) or because of price discrimination related reasons (Acquisti and Varian [2005]). See Acquisti et al. [2016], Bergemann and Bonatti [2019], and Agrawal et al. [2018] for excellent surveys of different aspects of this literature.

The rest of the paper proceeds as follows. Section 2 presents our model, focusing on the case with a single platform for simplicity. Section 3 provides our main results, in particular, characterizing the structure of equilibria in data markets and highlighting their inefficiency due to data externalities. It also shows how shutting down data markets may improve welfare. Section 4 extends these results to a setting with competing platforms, while Section 5 presents analogous results when the value of privacy of each user is their private information. Section 6 shows that two types of policies, taxes and third-party-mediated information sharing, can improve welfare. Section 7 concludes, while Appendix A presents the proofs of some of the results stated in the text and the online Appendix B contains the remaining proofs and additional examples and results.

2 Model

In this section we introduce our model, focusing first on the case with a single platform. Competition between platforms is analyzed in Section 4.

2.1 Information and Payoffs

We consider n users represented by the set $\mathcal{V} = \{1, \ldots, n\}$. Each user $i \in \mathcal{V}$ has a type denoted by x_i which is a realization of a random variable X_i. We assume that the vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ has a joint normal distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix of $\mathbf{X}$. Let Σ_{ij} designate the (i, j)-th entry of Σ and $\Sigma_{ii} = \sigma_i^2 > 0$ denote the variance of individual i's type.

Each user has some personal data, S_i, which is informative about her type (for example, based on her past behavior, preferences, or contacts). We suppose that $S_i = X_i + Z_i$ where Z_i is an independent random variable with standard normal distribution, i.e., $Z_i \sim \mathcal{N}(0, 1)$.

For any user joining the platform, the platform can derive additional revenue if it can predict her type. This might be because of improved personalized services, targeted advertising, or price discrimination for some services sold on the platform. Since the exact source of revenue for the platform is immaterial for our analysis, we simply assume that the platform's revenue from each user is a(n inverse) function of the mean square error of its forecast of the user's type, minus what the platform pays to users to acquire their information. Namely, the objective of the platform is to minimize

$$\sum_{i \in \mathcal{V}} \left(\mathbb{E}\left[\left(\hat{x}_i\left(\mathbf{S}\right) - X_i \right)^2 \right] - \sigma_i^2 + p_i \right), \tag{1}$$

where $\mathbf{S}$ is the vector of data the platform acquires, $\hat{x}_i\left(\mathbf{S}\right)$ is the platform's estimate of the user's type given this information, $-\sigma_i^2$ is included as a convenient normalization, and p_i denotes payments to user i from the platform (we ignore for simplicity any other transaction costs incurred by the platform and discuss taxes and regulations in Section 6). This payment to the user may be a direct one in an explicit data market or it may be an implicit transfer, for example, in the form of some good or service the platform provides to the user in exchange for her data.

Users value their privacy, which we also model in a reduced-form manner as a function of the same mean square error.[5] This reflects both pecuniary and nonpecuniary motives, for example, the fact that a user may receive a greater consumer surplus when the platform knows less about her or she may have a genuine demand for keeping her preferences, behavior, and information private. There may also be political and social reasons for privacy, for example, for concealing dissident activities or behaviors disapproved by some groups. We assume, specifically, that user i's value of privacy is $v_i \geq 0$, and her payoff is

$$v_i \left(\mathbb{E}\left[\left(\hat{x}_i\left(\mathbf{S}\right) - X_i \right)^2 \right] - \sigma_i^2 \right) + p_i.$$

This expression and its comparison with (1) clarifies that the platform and users have potentially-opposing preferences over information about user type. We have again subtracted σ_i^2 as a normal-

[5] In this and the next section, we do not model the decision of whether to join the platform. Joining decisions are introduced in Section 4, where we assume that users receive additional services (unrelated to their personal data) from platforms encouraging them to join even in the presence of loss of privacy.

ization, which ensures that if the platform acquires no additional information about the user and makes no payment to her, her payoff is zero.

Critically, users with $v_i < 1$ value their privacy less than the valuation that the platform attaches to information about them, and thus reducing the mean square error of the estimates of their types is socially beneficial. In contrast, users with $v_i > 1$ value their privacy more, and reducing their mean square error is socially costly. In a world without data externalities (where data about one user have no relevance to the information about other users), the first group of users should allow the platform to acquire (buy) their data, while the second group should not. A simple market mechanism based on prices for data can implement this efficient outcome.

We will see that the situation is very different in the presence of data externalities.

2.2 Leaked Information

A key notion for our analysis is *leaked information*, which captures the reduction in the mean square error of the platform's estimate of the type of a user. When the platform has no information about user i, its estimate satisfies $\mathbb{E}\left[(\hat{x}_i - X_i)^2\right] = \sigma_i^2$. As the platform receives data from this and other users, its estimate improves and the mean square error declines. The notion of leaked information captures this reduction in mean square error.

Specifically, let $a_i \in \{0, 1\}$ denote the data sharing action of user $i \in \mathcal{V}$ with $a_i = 1$ corresponding to sharing. Denote the the profile of sharing decisions by $\mathbf{a} = (a_1, \ldots, a_n)$ and the decisions of agents other than i by $\mathbf{a}_{-i}$. We also use the notation $\mathbf{S_a}$ to denote the data of all individuals for whom $a_j = 1$, i.e., $\mathbf{S_a} = (S_j :\ j \in \mathcal{V} \text{ s.t. } a_j = 1)$. Given a profile of actions $\mathbf{a}$, the *leaked information* of (or about) user $i \in \mathcal{V}$ is the reduction in the mean square error of the best estimator of the type of user i:

$$\mathcal{I}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}\left[(X_i - \hat{x}_i(\mathbf{S_a}))^2\right].$$

Notably, because of data externalities, leaked information about user i depends not just on her decisions but also on the sharing actions taken by all users.

With this notion at hand, we can write the payoff of user i given the price vector $\mathbf{p} = (p_1, \ldots, p_n)$ as

$$u_i(a_i, \mathbf{a}_{-i}, \mathbf{p}) = \begin{cases} p_i - v_i \mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}), & a_i = 1 \\ \\ -v_i \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}), & a_i = 0, \end{cases}$$

where recall that $v_i \geq 0$ is user's value of privacy.

We also express the platform's payoff more compactly as

$$U(\mathbf{a}, \mathbf{p}) = \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathbf{a}) - \sum_{i \in \mathcal{V}:\ a_i = 1} p_i. \tag{2}$$

2.3 Equilibrium Concept

An action profile $\mathbf{a} = (a_1, \ldots, a_n)$ and a price vector $\mathbf{p} = (p_1, \ldots, p_n)$ constitute a pure strategy equilibrium if both users and the platform maximize their payoffs given other players' strategies. More formally, in the next definition we define an equilibrium as a *Stackelberg equilibrium* in which the platform chooses the price vector recognizing the *user equilibrium* that will result following this choice.

Definition 1. Given the price vector $\mathbf{p} = (p_1, \ldots, p_n)$, an action profile $\mathbf{a} = (a_1, \ldots, a_n)$ is user equilibrium if for all $i \in \mathcal{V}$,

$$a_i \in \text{argmax}_{a \in \{0,1\}} u_i(a_i = a, \mathbf{a}_{-i}, \mathbf{p}).$$

We denote the set of user equilibria at a given price vector $\mathbf{p}$ by $\mathcal{A}(\mathbf{p})$. A pair $(\mathbf{p}^{\text{E}}, \mathbf{a}^{\text{E}})$ of price and action vectors is a pure strategy Stackelberg equilibrium if $\mathbf{a}^{\text{E}} \in \mathcal{A}(\mathbf{p}^{\text{E}})$ and there is no profitable deviation for the platform, i.e.,

$$U(\mathbf{a}^{\text{E}}, \mathbf{p}^{\text{E}}) \geq U(\mathbf{a}, \mathbf{p}), \quad \text{for all } \mathbf{p} \text{ and for all } \mathbf{a} \in \mathcal{A}(\mathbf{p}).$$

In what follows, we refer to a pure strategy Stackelberg equilibrium simply as an equilibrium.

3 Analysis

In this section, we first study the first-best information sharing decisions that maximize the sum of users and platform payoffs and then proceed to characterizing the equilibrium and its efficiency properties.

3.1 First Best

We define the first best as the data sharing decisions that maximize utilitarian social welfare or social surplus given by the sum of the payoffs of the platform and users. Social surplus from an action profile $\mathbf{a}$ is

$$\text{Social surplus}(\mathbf{a}) = U(\mathbf{a}, \mathbf{p}) + \sum_{i \in \mathcal{V}} u_i(\mathbf{a}, \mathbf{p}) = \sum_{i \in \mathcal{V}} (1 - v_i) \mathcal{I}_i(\mathbf{a}).$$

Prices do not appear in this expression because they are transfers from the platform to users.[6] The first-best action profile, $\mathbf{a}^{\text{W}}$, maximizes this expression. The next proposition characterizes the first-best action profile.

[6]In including the platform's payoff in social surplus we are assuming that this payoff is not coming from shifting revenues from some other (perhaps off-line) businesses. If we do not include the payoff of the platform in our welfare measure, our inefficiency results would hold a fortiori.

Proposition 1. *The first best involves* $a_i^{\mathrm{W}} = 1$ *if*

$$\sum_{j \in \mathcal{V}} (1 - v_j) \frac{\left(\mathrm{Cov}\left(X_i, X_j \mid a_i = 0, \mathbf{a}_{-i}^{\mathrm{W}}\right)\right)^2}{1 + \sigma_j^2 - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i}^{\mathrm{W}})} \geq 0, \tag{3}$$

and $a_i^{\mathrm{W}} = 0$ *if (3) is negative.*

The proof of this proposition as well as all other proofs, unless otherwise stated, are presented in Appendix A.

To understand this result, consider first the case in which there are no data externalities so that the covariance terms in (3) are zero, except $\mathrm{Cov}\left(X_i, X_i \mid a_i = 0, \mathbf{a}_{-i}^{\mathrm{W}}\right) = \sigma_i^2$, so that the left-hand side is simply $\sigma_i^4/(1 + \sigma_i^2)$ times $1 - v_i$. This yields $a_i^{\mathrm{W}} = 1$ if $v_i \leq 1$. The situation is different in the presence of data externalities, because now the covariance terms are non-zero. In this case, an individual should optimally share her data only if it does not reveal too much about users with $v_j > 1$.

3.2 Equilibrium Preliminaries

The next lemma characterizes two important properties of the leaked information function $\mathcal{I}_i : \{0, 1\}^n \to \mathbb{R}$.

Lemma 1. *1.* Monotonicity: *for two action profiles* $\mathbf{a}$ *and* $\mathbf{a}'$ *with* $\mathbf{a} \geq \mathbf{a}'$*,*

$$\mathcal{I}_i(\mathbf{a}) \geq \mathcal{I}_i(\mathbf{a}'), \quad \forall i \in \{1, \ldots, n\}.$$

2. Submodularity: *for two action profiles* $\mathbf{a}$ *and* $\mathbf{a}'$ *with* $\mathbf{a}'_{-i} \geq \mathbf{a}_{-i}$*,*

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) \geq \mathcal{I}_i(a_i = 1, \mathbf{a}'_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}'_{-i}).$$

The monotonicity property states that as the set of users who share their information expands, the leaked information about each user (weakly) increases. This is an intuitive consequence of the fact that more information always facilitates the estimation problem of the platform and reduces the mean square error of its estimates. More important for the rest of our analysis is the submodularity property, which implies that the marginal increase in the leaked information from individual i's sharing decision is decreasing in the information shared by others. This too is intuitive and follows from the fact that when others' actions reveal more information, there is less to be revealed by the sharing decision of any given individual.

Using Lemma 1 we next show that for any price vector $\mathbf{p} \in \mathbb{R}^n$, the set $\mathcal{A}(\mathbf{p})$ is a (non-empty) complete lattice.

Lemma 2. *For any* $\mathbf{p}$*, the set* $\mathcal{A}(\mathbf{p})$ *is a complete lattice, and thus has a least and a greatest element.*

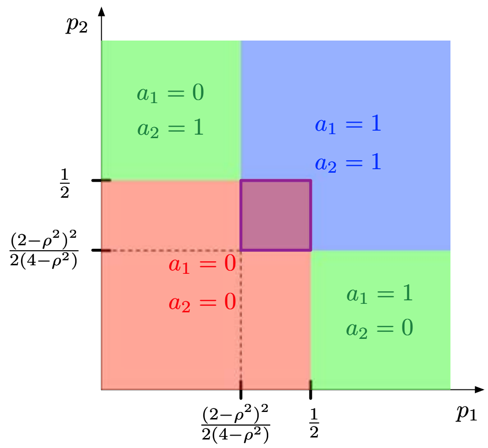


Figure 1: The user equilibrium as a function of price vector (p_1, p_2) in the setting of Example 1. For the prices in the purple area in the center, both $a_1 = a_2 = 0$ and $a_1 = a_2 = 1$ are user equilibria.

Lemma 2 implies that the set of user equilibria is always non-empty, but may not be singleton as we illustrate in the next example.

Example 1. Suppose there are two users 1 and 2 with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $v_1 = v_2 = v$. The set of user equilibria in this case is depicted in Figure 1 (see Appendix B.3 for details). When $p_1, p_2 \in \left[\frac{(2-\rho^2)^2}{2(4-\rho^2)}, \frac{1}{2}\right]$, both action profiles $a_1 = a_2 = 0$ and $a_1 = a_2 = 1$ are user equilibria. This is a consequence of the submodularity of the leaked information function (Lemma 1): when user 1 shares her data, she is also revealing a lot about user 2, and making it less costly for her to share her data. Conversely, when user 1 does not share, this encourages user 2 not to share. Despite this multiplicity of user equilibria, there exists a unique (Stackelberg) equilibrium for this game given by $a_1^{\mathrm{E}} = a_2^{\mathrm{E}} = 1$ and $p_1^{\mathrm{E}} = p_2^{\mathrm{E}} = \frac{(2-\rho^2)^2}{2(4-\rho^2)}$. This uniqueness follows because the platform can choose the price vector to encourage both users to share.

3.3 Existence of Equilibrium

The next theorem establishes the existence of a (pure strategy) equilibrium.

Theorem 1. *An equilibrium always exists. That is, there exist an action profile* $\mathbf{a}^{\mathrm{E}}$ *and a price vector* $\mathbf{p}^{\mathrm{E}}$ *such that* $\mathbf{a}^{\mathrm{E}} \in \mathcal{A}(\mathbf{p}^{\mathrm{E}})$, *and*

$$U(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) \geq U(\mathbf{a}, \mathbf{p}), \qquad \text{for all } \mathbf{p} \text{ and for all } \mathbf{a} \in \mathcal{A}(\mathbf{p}). \tag{4}$$

Note that the equilibrium may not be unique, but if there are multiple equilibria, all of them yield the same payoff for the platform (since otherwise (4) would not be satisfied for the equilibrium with lower payoff for the platform). The following example clarifies this point.

Example 2. Suppose there are three users with the same value of privacy and variance: $v_i = 1.18$ and $\sigma_i^2 = 1$ for $i = 1, 2, 3$. We let all off-diagonal entries of Σ to be 0.3. Any action profile where two out of three users share their information is an equilibrium, and thus there are three distinct equilibria. But it is straightforward to verify that they all yield the same payoff to the platform.

3.4 An Illustrative Example

In this subsection, we provide an illustrative example that shows how some of the key objects in our analysis are computed and highlights a few of the subtle aspects of the equilibrium. Consider the same setting as in Example 1 with two users with the same value of privacy, v, and a correlation coefficient ρ between their information. Given the action profile of users, the joint distribution of (X_1, S_1, S_2) in this example is

$$\begin{pmatrix} X_1 \\ S_1 \\ S_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 & \rho \\ 1 & 1+\gamma_1^2 & \rho \\ \rho & \rho & 1+\gamma_2^2 \end{pmatrix} \right),$$

where

$$\gamma_i^2 = \begin{cases} 1 & a_i = 1, \\ \infty & a_i = 0, \end{cases}$$

(see Appendix B.3). Suppose the platform has received the signals (S_1, S_2), then its optimal estimator for X_1, $\hat{x}_1(S_1, S_2)$, is simply the conditional expectation of X_1 given s_1 and s_2, and its mean square error is equal to this estimator's variance,

$$\frac{\gamma_1^2(1+\gamma_2^2-\rho^2)}{(1+\gamma_1^2)(1+\gamma_2^2)-\rho^2}.$$

Similarly, the mean square error of the platform's best estimator for X_2 is

$$\min_{\hat{x}_2} \mathbb{E}\left[(\hat{x}_2(S_1, S_2) - X_2)^2 \right] = \frac{\gamma_2^2(1+\gamma_1^2-\rho^2)}{(1+\gamma_1^2)(1+\gamma_2^2)-\rho^2}.$$

We first show that the total payment from the platform to users is non-monotone in the number of users sharing their information. When the platform induces both users to share ($a_1 = a_2 = 1$), it makes a total payment of $v\frac{(2-\rho^2)^2}{4-\rho^2}$. In contrast, when it only induces the first user to share ($a_1 = 1$, $a_2 = 0$), this will cost $\frac{v}{2}$. Therefore, when $\rho^2 \geq \frac{7-\sqrt{17}}{4} \approx 0.71$, the platform pays less to have both users share their data. Intuitively, this cost-saving for the platform is a consequence of the submodularity of leaked information (Lemma 1): when both users share, the data of each are less valuable in view of the information revealed by the other user. This finding reflects one of the

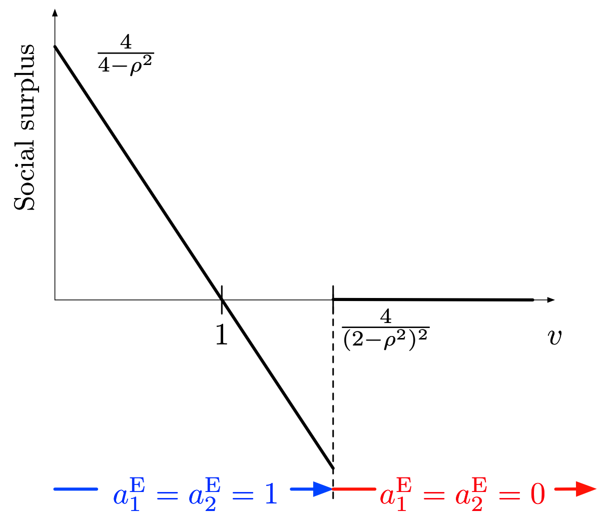


Figure 2: Equilibrium and social surplus as a function of the value of privacy v for a setting with two users with $\sigma_1^2 = \sigma_2^2 = 1$, $\Sigma_{12} = \rho$, and $v_1 = v_2 = v$.

claims made in the Introduction: market prices for data do not reflect the value that users attached to their privacy and may be depressed because of data externalities.

We next illustrate that equilibrium (social) surplus is non-monotonic in the users' value of privacy. Equilibrium surplus is depicted in Figure 2. For values of v larger than $\frac{4}{(2-\rho^2)^2}$, users do not share their data and equilibrium surplus is zero. When v is smaller than 1, users share their data and equilibrium surplus is positive. For intermediate values of v, in particular for $v \in [1, \frac{4}{(2-\rho^2)^2}]$, the platform chooses a price vector that induces both users to share their data, but in this case, the social surplus is negative. The intuition is related to the point already emphasized in the previous paragraph: when both users share their data, the externalities depress the market prices for data and this makes it profitable for the platform to acquire the users' data even though $v > 1$. More explicitly, when user 2 shares her data, this reveals sufficient information about user 1 that she becomes willing to accept a relatively low price for sharing her data, and this maintains an equilibrium with low prices for data even though both users attach a relatively high value to their privacy.

3.5 Equilibrium Prices

In this subsection, we characterize the equilibrium price vector. For any action profile $\mathbf{a} \in \{0, 1\}^n$, let $\mathbf{p}^{\mathbf{a}}$ denote the least (element-wise minimum) equilibrium price vector that sustains an action profile $\mathbf{a}$ in a user equilibrium. More specifically, $\mathbf{p}^{\mathbf{a}}$ is defined such that:[7]

$$\mathbf{p}^{\mathbf{a}} \leq \mathbf{p}, \quad \text{for all } \mathbf{p} \text{ such that } \mathbf{a} \in \mathcal{A}(\mathbf{p}).$$

[7]Prices for users not sharing their data are not well-defined.

Profit maximization by the platform implies that equilibrium prices must satisfy this property — since otherwise the platform could reduce prices and still implement the same action profile. We therefore refer to $\mathbf{p}^{\mathbf{a}}$ as "equilibrium price vector" or simply as "equilibrium prices" (with the understanding that these would be the equilibrium prices when the platform chooses to induce action profile $\mathbf{a}$).

The next theorem computes this price vector (and shows that it exists).

Theorem 2. *For any action profile $\mathbf{a} \in \{0,1\}^n$, we have*

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) = \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) + \frac{\left(\sigma_i^2 - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right)^2}{(\sigma_i^2 + 1) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})}, \tag{5}$$

and

$$\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) = \mathbf{d}_i^T \left(I + D_i\right)^{-1} \mathbf{d}_i, \quad \text{for all } a_i = 1,$$

where D_i is the matrix obtained by removing row and column i from matrix Σ as well as all rows and columns j for which $a_j = 0$, and $\mathbf{d}_i$ is $(\Sigma_{ij} : j$ s.t. $a_j = 1)$. The equilibrium price that sustains action profile $\mathbf{a}$ is

$$p_i^{\mathbf{a}} = \begin{cases} v_i \frac{\left(\sigma_i^2 - \mathcal{I}_i(a_i=0,\mathbf{a}_{-i})\right)^2}{(\sigma_i^2+1) - \mathcal{I}_i(a_i=0,\mathbf{a}_{-i})}, & a_i = 1, \\ 0, & a_i = 0. \end{cases}$$

The first part of Theorem 2 provides a decomposition of leaked information about user i in terms of leaked information about her when she does not share her data. In particular, the first term on the right-hand side of the equation (5), $\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$, is her leaked information resulting from the data sharing of other users and thus represents the data externality. The second term is the additional leakage when user i shares her data. The second part of Theorem 2 states that the equilibrium price offered to any user i who shares her information must make her indifferent between sharing and not sharing. This is because prices are determined by the platform's offers. This is also why the equilibrium price, $p_i^{\mathbf{a}}$, is equal to the value of privacy, v_i, multiplied by the second term in (5), which is the additional leakage of information and hence the loss of privacy resulting from the user's own data sharing.

The following is an immediate corollary of Theorem 2.

Corollary 1. *For any user i, the equilibrium price $p_i^{(a_i=1,\mathbf{a}_{-i})}$ (that induces $a_i = 1$ for any action profile $\mathbf{a}_{-i} \in \{0,1\}^{n-1}$) is increasing in σ_i^2 and decreasing in the data externality captured by $\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$. Moreover, leaked information $\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i})$ is increasing in σ_i^2 and in the data externality $\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$.*

The first part of Corollary 1 shows that a higher variance of user's type, σ_i^2, increases the equilibrium price. Intuitively, a higher variance makes the user's type more difficult to predict and thus her own information more valuable. This also explains why the price is decreasing in the data externality — represented by information leaked by others, $\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$. The last part of

Corollary 1 shows that a higher variance of individual type, as well as a greater data externality, increase the overall leakage of information about the user.

The next proposition establishes that greater correlation between users' data, and thus greater data externality, reduces equilibrium prices.

Proposition 2. *For any action profile $\mathbf{a} \in \{0,1\}^n$ and any $i \in \mathcal{V}$, the price $p_i^{\mathbf{a}}$ is nonincreasing in the absolute value of the covariance between any pair of users given the action profile $\mathbf{a}$.*

The equilibrium price for i is the difference between the information that platform has about i with and without i's own data (multiplied by her value of privacy v_i). Suppose first that the correlation between i's data and some other user j's data increases. Mathematically, this means that $|\mathrm{Cov}(X_i, X_j \mid \mathbf{a})|$ increases, and user j's information about i becomes more accurate. Therefore, provided that $a_j = 1$, this reduces the value of user i's information for predicting her type and thus depresses her benefits from protecting her data. The same result also applies when the correlation between two other users increases, that is, when $|\mathrm{Cov}(X_j, X_k \mid \mathbf{a})|$ increases. In this case, the platform obtains more accurate information about users j and k, and indirectly their data become more informative about user i, leading to the same conclusion.

Proposition 2 establishes that equilibrium prices are nonincreasing in the correlation between users' data. Yet the number of users who share their data in equilibrium is non-monotone in this correlation as we show in the next example.

Example 3. Consider a setting with three users with $\Sigma_{12} = \Sigma_{13} = 1/2$, $\Sigma_{23} = \rho$, $v_i = 3/2$, and $\sigma_i^2 = 1$ for all users. We have the following cases.

1. $\rho = 0$: the equilibrium decisions are $a_1 = 1$, and $a_2 = a_3 = 0$, the total payment is 0.75, and the total information leakage is $\sum_{i=1}^3 \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) = 0.75$.

2. $\rho = 1/2$: the equilibrium decisions are $a_1 = a_2 = a_3 = 1$, the total payment is 1.6, and the total information leakage is $\sum_{i=1}^3 \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) = 1.66$.

3. $\rho = 1$: the equilibrium decisions are $a_1 = 0$ and $a_2 = a_3 = 1$, the total payment is 0.5, and the total information leakage is $\sum_{i=1}^3 \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) = 1.5$.

Therefore, as we increase the correlation among users, the set of users that share information in equilibrium may increase (from case 1 to case 2) or decrease (from case 2 to case 3). The intuition is that as more information is leaked about a user (in this case about user 1), the value of her data both to the platform and to herself declines, and the platform may no longer find it worthwhile to compensate her for selling her data.

The next proposition, however, shows that equilibrium prices are decreasing in the set of users sharing their data.

Proposition 3. *For two action profiles $\mathbf{a}, \mathbf{a}'$ with $\mathbf{a}' \geq \mathbf{a}$, we have $p_i^{\mathbf{a}'} \leq p_i^{\mathbf{a}}$ for all $i \in \mathcal{V}$ for which $a_i = 1$.*

Proposition 3 follows from Theorem 2 and Lemma 1. In particular, using Theorem 2 the equilibrium price for user i is her additional loss of privacy (increase in the information leakage multiplied by v_i) if she shares her data. From the submodularity of information leakage (Lemma 1) the additional information the user leaks about herself decreases when more people share their data.[8]

3.6 Inefficiency

This subsection presents one of our main results, documenting the extent of inefficiency in data markets.

First note that all users with value of privacy less than 1 will always share their data in equilibrium. For future reference, we state this straightforward result as a lemma.

Lemma 3. *All users with value of privacy $v_i \leq 1$ share their data in equilibrium.*[9]

Motivated by this lemma, we partition users into two sets, those with value of privacy below 1 ("low-value users") and those above ("high-value users"):

$$\mathcal{V}^{(l)} = \{i \in \mathcal{V} : \ v_i \leq 1\} \text{ and } \mathcal{V}^{(h)} = \{i \in \mathcal{V} : \ v_i > 1\}.$$

We also denote by $\mathbf{v}^{(h)}$ and $\mathbf{v}^{(l)}$ the vectors of valuations of privacy for high-value and low-value users, respectively. Lemma 3 then implies that for all $i \in \mathcal{V}^{(l)}$ we have $a_i^{\mathrm{E}} = 1$.

The next theorem provides necessary and sufficient conditions for efficiency and inefficiency.

Theorem 3.
1. *Suppose every high-value user is uncorrelated with all other users. Then the equilibrium is efficient.*
2. *Suppose at least one high-value user is correlated (has a non-zero correlation coefficient) with a low-value user. Then there exists $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ such that for $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ the equilibrium is inefficient.*
3. *Suppose every high-value user is uncorrelated with all low-value users and at least one high-value user is correlated with another high-value user. Let $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ be the subset of high-value users correlated with at least one other high-value user. Then for each $i \in \tilde{\mathcal{V}}^{(h)}$ there exists $\bar{v}_i > 0$ such that if for any $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$, the equilibrium is inefficient*

Theorem 3 clarifies the source of inefficiency in our model. If high-value users are not correlated with others, the equilibrium is efficient. In this case, there may still be data externalities among low-value users and these may affect market prices (and the distribution of economic gains

[8] The proposition covers the prices for users sharing their data, since prices for those not sharing are not well-defined.

[9] The only subtlety here is about users with $v_i = 1$. If these users' information is correlated with others who are already sharing, their equilibrium price will be strictly less than 1, and this will make it strictly beneficial for the platform to purchase their data. If they are correlated with others who are not sharing, then the platform would still like to purchase these data because of the additional reduction in the mean square error of its estimates of others' types they enable. When such an individual is uncorrelated with anybody else, then the platform would be indifferent between purchasing her data and not. In this case, for simplicity of notation, we suppose that it still purchases.

between the users and the platform). But they do not create a loss of privacy for users who prefer not to share their data.

However, the second part of the theorem shows that if high-value users are correlated with low-value users, the equilibrium is typically inefficient. The additional condition $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ is not a restrictive one as highlighted in Example 4 below and rules out cases in which high-value users suffer only little loss of privacy but generate socially valuable information about low-value users. In general, the inefficiency identified in this part of the theorem can take one of two forms: either high-value users do not share their data, but because of information leaked about them, they suffer a loss of privacy. Or given the amount of leaked information about them, high-value users decide to share themselves (despite their initial reluctance to do so).

Finally, the third part of the theorem covers the remaining case, where high-value users are uncorrelated with low-value users but are correlated among themselves. The equilibrium is again inefficient, because the platform can induce some of them to share their data (even though individually each would prefer not to). This is because when a subset of them share, this compromises the privacy of others, depresses data prices, and may incentivize them to share too (in turn further depressing data prices). This inefficiency applies when some high-value users have intermediate values of privacy (i.e., $v_i \in (1, \bar{v}_i)$), since those with sufficiently high value of privacy cannot be induced to share their data.

Overall, this theorem highlights that inefficiency in data markets originates from the combination of sufficiently high value attached to privacy by some users and their correlation with other users. It therefore emphasizes that inefficiency in our model is tightly linked to data externalities.

3.7 Are Data Markets Beneficial?

Theorem 3 focuses on the comparison of the market equilibrium to the first best. This is a tough comparison for the market because in the first best some users share their data and benefit from market transactions, while others do not share. A lower bar for data markets is whether they achieve positive social surplus so that any inefficiencies they create are (partially) compensated by benefits for other agents. We next show that this is not necessarily the case and provide a sufficient condition for the equilibrium (social) surplus to be negative — so that shutting down data markets all together would improve social surplus (and thus utilitarian welfare). Let us also introduce the following notation: for any action profile $\mathbf{a} \in \{0,1\}^n$, we let $\mathcal{I}_i(T)$ denote the leaked information about user i where $T = \{i \in \mathcal{V} : \ a_i = 1\}$.

Proposition 4.

$$\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) \leq \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}) - \sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}).$$

The first term is an upper bound on the gain in social surplus from the sharing decisions of low-value users (even if these gains do not necessarily accrue to the users themselves and are mainly captured by the platform). This expression is an upper bound because we are evaluating this term under the assumption that all users share their data, thus maximizing the amount of socially beneficial information about low-value users. The second term is a lower bound on the loss of privacy from high-value users. It is a lower bound because the loss of privacy is evaluated for the minimal set of agents, the low-value ones, who always share their data (in equilibrium a superset of $\mathcal{V}^{(l)}$ will share their data).

We also add that leaked information in this proposition is only a function of the matrix Σ as shown in Theorem 2, so the right-hand side is in terms of model parameters and does not depend on equilibrium objects.

An important and immediate implication of this proposition is contained in the next corollary.

Corollary 2. *If*

$$\sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}) > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}), \tag{6}$$

then the equilibrium surplus is negative and utilitarian welfare improves if data markets are shut down.

The next proposition provides a sufficient condition in terms of values of privacy and correlations between data that ensures condition (6) and implies that the equilibrium necessarily has negative social surplus.

Proposition 5. *Suppose $\sigma_i^2 = 1$ for all $i \in \mathcal{V}$. If*

$$\sum_{i \in \mathcal{V}^{(h)}} \left((v_i - 1) \frac{\sum_{j \in \mathcal{V}^{(l)}} \Sigma_{ij}^2}{|\mathcal{V}^{(l)}| + 1} \right) > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i),$$

then the equilibrium surplus is negative. When the correlation between any pair of high-value and low-value users is greater than ρ, this condition holds if

$$\frac{\rho^2 |\mathcal{V}^{(l)}|}{|\mathcal{V}^{(l)}| + 1} \sum_{i \in \mathcal{V}^{(h)}} (v_i - 1) > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i).$$

Proposition 5 provides explicit sufficient conditions in terms of the values of privacy and the correlation between high and low-value users for negative equilibrium surplus.

Example 4. We consider a setting with two communities, each of size 10. Suppose that all users in community 1 are low-value and have a value of privacy equal to 0.9, while all users in community 2 are high-value (with $v_h > 1$). We also take the variances of all user data to be 1, the correlation between any two users who belong to the same community to be $1/20$, and the correlation between any two users who belong to different communities to be ρ. Figure 3 depicts equilibrium surplus as a function of v_h and ρ. The curve in the figure represents the combinations of these two variables for which the social surplus is equal to zero. Moving in the northeast direction reduces

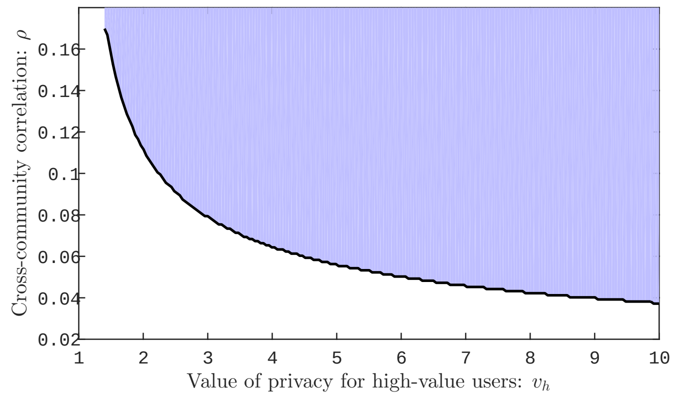


Figure 3: Shaded area shows the pairs of (ρ, v_h) with negative equilibrium surplus in the setting of Example 4.

equilibrium surplus and hence the shaded area has negative surplus. Consequently, in this part of the parameter space, shutting down data markets improves utilitarian social welfare.

Two points are worth noting. First, relatively small values of the correlation coefficient ρ are sufficient for social surplus to be negative. Second, when v_h is very close to 1, the social surplus is always positive because the negative surplus from high-value users is compensated by the social benefits their data sharing creates for low-value users.

4 Competition Among Platforms

In this section we generalize the main results from the previous section to a setting in which multiple platforms compete for (the data of) users. For simplicity we focus on the case in which there are two platforms. Formally, we are dealing with a three-stage game in which first users decide which platform to join (if any), then platforms simultaneously offer prices for data, and then finally all users simultaneously decide whether to share their data. We start by assuming that data externalities are only among users joining the same platform, but then generalize this to the case in which the platform learns valuable information about users that are not on its platform as well. At the end of the section, we also discuss the case in which platforms first set prices for data in order to attract users.

4.1 Information and Payoffs

For any $i \in \mathcal{V}$, we denote by $b_i \in \{0, 1, 2\}$ the joining decision of user i in the first-stage game where $b_i = 0$ means user i does not join, $b_i = 1$ means she joins platform 1, and $b_i = 2$ stands for joining platform 2. Let us also define

$$J_1 = \{i \in \mathcal{V} \,:\, b_i = 1\} \text{ and } J_2 = \{i \in \mathcal{V} \,:\, b_i = 2\},$$

as the sets of users joining the two platforms.

Similar to the monopoly case in the previous section, the payoff of a platform is a function of leaked information about users and payments to users. So for platform $k \in \{1, 2\}$, we have

$$U^{(k)}(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}) = \sum_{i \in J_k} \mathcal{I}_i(\mathbf{a}^{J_k}) - \sum_{i \in J_k:\ a_i^{J_k}=1} p_i^{J_k}, \tag{7}$$

where $\mathbf{a}^{J_k} \in \{0,1\}^{|J_k|}$ denotes the sharing decision of users belonging to this platform, and $\mathbf{p}^{J_k}$ denotes the vector of prices the platform offers to users in J_k.

The payoff of a user has three parts. First, each user receives a valuable service from the platform it joins. Since we are modeling joining decisions in this section, we will be more explicit about this "joining value" and assume that it depends on who else joins the platform. We therefore write this part of the payoff as $c_i(J_{b_i})$ for user i joining platform b_i, with the convention that $J_0 = \varnothing$, and also normalize $c_i(J) = 0$ for all $J \not\ni i$ and for all $i \in \mathcal{V}$. Second, the user suffers a disutility due to loss of privacy from leaked information as before, and we again denote the value of privacy for user i by v_i. Third, she receives benefits from any payments from the platform in return of the data she shares. Thus the payoff to user i joining platform $k \in \{1, 2\}$ can be written as

$$u_i(J_k, a_i, \mathbf{a}_{-i}^{J_k}, \mathbf{p}^{J_k}) = \begin{cases} p_i^{J_k} - v_i \mathcal{I}_i\left(a_i = 1, \mathbf{a}_{-i}^{J_k}\right) + c_i(J_k), & a_i = 1, \\ \\ -v_i \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_k}\right) + c_i(J_k) & a_i = 0. \end{cases}$$

With our convention, when the user chooses $b_i = 0$, this payoff is equal to zero — there is no data sharing decision, the payment is equal to zero, leaked information is equal to zero, and $c_i(\varnothing) = 0$.

The timing of events is as follows.

1. Users simultaneously decide which platform, if any, to join, i.e., $\mathbf{b} = \{b_i\}_{i \in \mathcal{V}}$, which determines J_1 and J_2.

2. Given J_1 and J_2, the two platforms simultaneously offer price vectors $\mathbf{p}^{J_1}$ and $\mathbf{p}^{J_2}$.

3. Given J_1 and J_2 and $\mathbf{p}^{J_1}$ and $\mathbf{p}^{J_2}$, users simultaneously make their data sharing decisions, $\mathbf{a} : \{0, 1, 2\}^n \times \mathbb{R}^n \to \{0, 1\}$.

The assumption that users join platforms before prices is adopted both for simplicity and to capture the fact that there is currently a limited ability for platforms to attract customers by offering prices for data. We discuss this type of price competition and its implications below. Even though prices are offered at the second stage, users will anticipate these prices and the implications they have for leaked information in making their joining decisions.

4.2 Equilibrium Concept

We first observe that from the second stage on, once the set of users joining a platform is determined, this game is identical to the one we analyzed in the previous section. Hence, the Stackelberg equilibrium from the second stage onwards is defined analogously. Then in the first stage, we define a *joining equilibrium*, as a profile of joining decisions anticipating the equilibrium from the second stage onward. More formally:

Definition 2. Given a joining decision $\mathbf{b}$ and the corresponding sets of users on the two platforms, J_1 and J_2, a pure strategy *Stackelberg equilibrium* from the second stage onwards is given by price vectors $\mathbf{p}^{J_1,\mathrm{E}}$ and $\mathbf{p}^{J_2,\mathrm{E}}$ and action profiles $(\mathbf{a}^{J_1,\mathrm{E}}, \mathbf{a}^{J_2,\mathrm{E}})$ such that $\mathbf{a}^{J_k,\mathrm{E}} \in \mathcal{A}(\mathbf{p}^{J_k,\mathrm{E}})$ and

$$U^{(k)}(J_k, \mathbf{a}^{J_k,\mathrm{E}}, \mathbf{p}^{J_k,\mathrm{E}}) \geq U^{(k)}(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}), \qquad \text{for all } \mathbf{p}^{J_k} \text{ and for all } \mathbf{a}^{J_k} \in \mathcal{A}(\mathbf{p}^{J_k})$$

for $k \in \{1, 2\}$.

Joining decision profile $\mathbf{b}^{\mathrm{E}}$ and the corresponding sets of users on the two platforms, J_1^{E} and J_2^{E}, constitute a pure strategy *joining equilibrium* if no user has a profitable deviation. That is, for all $i \in \mathcal{V}$,

$$\begin{aligned} u_i(J_{b_i}^{\mathrm{E}}, \mathbf{a}^{J_{b_i}^{\mathrm{E}},\mathrm{E}}, \mathbf{p}^{J_{b_i}^{\mathrm{E}},\mathrm{E}}) &\geq u_i(J_k^{\mathrm{E}} \cup \{i\}, \mathbf{a}^{J_k^{\mathrm{E}}\cup\{i\},\mathrm{E}}, \mathbf{p}^{J_k^{\mathrm{E}}\cup\{i\},\mathrm{E}}) \text{ for } k \neq b_i \text{ and} \\ u_i(J_{b_i}^{\mathrm{E}}, \mathbf{a}^{J_{b_i}^{\mathrm{E}},\mathrm{E}}, \mathbf{p}^{J_{b_i}^{\mathrm{E}},\mathrm{E}}) &\geq 0, \quad \text{for all } i \in J_{b_i}^{\mathrm{E}}. \end{aligned}$$

Note that the first condition for the joining equilibrium ensures that each user prefers the platform she joins to the other platform, and given our convention that $J_0 = \varnothing$, it also implies that users not joining either platform prefer this to joining one of the two platforms. The second condition makes sure that a user joining a platform receives nonnegative payoff, since not joining either platform guarantees zero payoff.

4.3 Equilibrium Existence and Characterization

Since our focus is on situations in which users join online platforms and share their data, we impose that joining values are sufficiently large.

Assumption 1. For each $i \in \mathcal{V}$, *we have*

1. *for all J and J' such that $i \in J$ and $J \subset J'$, we have $c_i(J') > c_i(J)$.*
2. $c_i(\{i\}) > \max_{i \in \mathcal{V}} v_i \sigma_i^2$.

This assumption implies that users receive greater services from a platform when there are more users on the platform, which captures the network effects in online services and social media. The fact that this benefit is indexed by i means that users can prefer being on the same platform with different sets of other users. This assumption also imposes that even when there are no other users on a platform, the value of the services provided by the platform exceed the cost of loss of privacy. This second aspect directly yields the next lemma, which simplifies the rest of our analysis. For the rest of this section, we impose Assumption 1 without explicitly stating it.

Lemma 4. *Each user joins one of the two platforms. In other words, $b_i = 1$ or 2 for all $i \in \mathcal{V}$.*

The next theorem is the direct analogue of Theorems 1 and 2 in the previous section and characterizes the Stackelberg equilibrium given joining decisions.

Theorem 4. *Consider a joining profile $\mathbf{b}$ with the corresponding sets of users J_1 and J_2. Then a pure strategy Stackelberg equilibrium exists and satisfies*

$$U^{(k)}(J_k, \mathbf{a}^{J_k,\mathrm{E}}, \mathbf{p}^{J_k,\mathrm{E}}) \geq U^{(k)}(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}), \qquad \textit{for all } \mathbf{p}^{J_k}, \mathbf{a}^{J_k} \in \mathcal{A}(\mathbf{p}^{J_k}), \textit{ and } k = 1, 2.$$

Moreover, for any $i \in J_k$ the equilibrium prices are

$$p_i^{J_k,\mathrm{E}} = v_i \left(\mathcal{I}_i(\mathbf{a}^{J_k,\mathrm{E}}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{J_k,\mathrm{E}}) \right). \tag{8}$$

and

$$u_i(J_k^{\mathrm{E}}, \mathbf{a}^{J_k^{\mathrm{E}},\mathrm{E}}, \mathbf{p}^{J_k^{\mathrm{E}},\mathrm{E}}) = -v_i \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{J_k,\mathrm{E}}) + c_i(J_k). \tag{9}$$

Although a pure strategy Stackelberg equilibrium exists, a pure strategy joining equilibrium may not. This is because of data externalities, which make users sometimes go to the platform where less of their information will be leaked.[10] The next example illustrates this.

Example 5. Suppose that there are three users with

$$\Sigma = \begin{pmatrix} 4 & .05 & .06 \\ .05 & 4 & .05 \\ .06 & .05 & .3 \end{pmatrix},$$

and values of privacy are given by

$$v_1 = 0.01, \qquad v_2 = 0.99, \qquad v_3 = 1.1.$$

Also for simplicity, in this example we take c_i to be a constant function c for all i (meaning that conditional on joining a platform, all users receive the same benefit). In this setting, user 3 has the highest value for privacy, but shares her data if she is on the same platform as user 2 (this is

[10] Note that even though increasing the amount of leaked information about low-value users increases social surplus, it reduces these users' payoff because it enables their platform to pay less for their data. This is the reason low-value users may prefer to have less information leaked about themselves and choose a platform with fewer other low-value users.

because her data are highly correlated with user 2's). However, she does not share her data if she is on the same platform as user 1.

We next list the possible pure strategy joining profiles and show that none of them can be an equilibrium (see Appendix B.3 for details).

1. $J_1 = \{1, 2\}$, $J_2 = \{3\}$: with this joining profile, the resulting Stackelberg equilibrium involves platform 1 buying the data of both users 1 and 2, while platform 2 does not buy user 3's data. In particular, from equation (8), the price of user 1's data can be expressed as $v_1(\mathcal{I}_1(a_1 = 1, a_2 = 1) - \mathcal{I}_1(a_1 = 0, a_2 = 1))$, which yields this user a payoff of $-v_1\mathcal{I}_1(a_1 = 0, a_2 = 1) + c$ in this candidate equilibrium. This implies that user 1 has a profitable deviation, which is to switch to platform 2. To see that this deviation is profitable, note that after this deviation, we have $J_1 = \{2\}$ and $J_2 = \{1, 3\}$, and the resulting Stackelberg equilibrium involves platform 1 buying user 2's data, while platform 2 buys user 1's data but not user 3's data. This gives user 1 a payoff of c, verifying that the deviation is beneficial for her.

2. $J_1 = \{2, 3\}$, $J_2 = \{1\}$: with this joining profile, the resulting Stackelberg equilibrium involves platform 1 buying the data of both users 2 and 3, while platform 2 buys user 1's data. With a similar reasoning, user 2's payoff is $-v_2\mathcal{I}_2(a_2 = 0, a_3 = 1) + c$. But in this case user 2 has a profitable deviation. By switching to platform 2, equation (8) implies that she will be offered a price of $v_2(\mathcal{I}_2(a_2 = 1, a_1 = 1) - \mathcal{I}_2(a_2 = 0, a_1 = 1))$ and receive a payoff of $-v_2\mathcal{I}_1(a_2 = 0, a_1 = 1) + c$, which exceeds her candidate equilibrium payoff.

3. $J_1 = \{1, 3\}$, $J_2 = \{2\}$: in this case, with a similar reasoning, user 3 has a profitable deviation. In the candidate equilibrium, this user receives a payoff of $-v_3\mathcal{I}_3(a_1 = 1, a_3 = 0) + c$, and if she deviates and switches to platform 2, she receives the greater payoff of $-v_3\mathcal{I}_3(a_2 = 1, a_3 = 0) + c$.

4. $J_1 = \{1, 2, 3\}$, $J_2 = \emptyset$: in this case, user 3 again has a profitable deviation and can increase her payoff from $-v_3\mathcal{I}_3(a_1 = 1, a_2 = 1, a_3 = 0) + c$ to c by switching to platform 2. This establishes that there is no pure strategy equilibrium in this game.

Even though a pure strategy equilibrium may fail to exist, we next show that a mixed strategy joining equilibrium always exists.

Definition 3. For any use $i \in \mathcal{V}$, let $\mathcal{B}_i$ be the set of probability measures over $\{1, 2\}$ for user i, $\beta_i \in \mathcal{B}_i$ be a mixed strategy for user i, and $\boldsymbol{\beta} \in \prod_{i \in \mathcal{V}} \mathcal{B}_i$ be a mixed strategy profile. Then $\boldsymbol{\beta}^{\mathrm{E}}$ is a mixed strategy joining equilibrium if

$$u_i(\beta_i^{\mathrm{E}}, \boldsymbol{\beta}_{-i}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) \geq u_i(\beta_i, \boldsymbol{\beta}_{-i}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}), \quad \text{for all } i \in \mathcal{V} \text{ and } \beta_i \in \mathcal{B}_i,$$

where $u_i(\beta_i, \boldsymbol{\beta}_{-i}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) = \mathbb{E}_{b_i \sim \beta_i, \mathbf{b}_{-i} \sim \mathbf{q}_{-i}} \left[u_i(b_i, \mathbf{b}_{-i}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) \right]$.

Theorem 5. *There always exists a mixed strategy joining equilibrium in which all users join each platform with probability 1/2.*

Theorem 5 follows since when all other users are choosing one of the two platforms uniformly at random (each with probability $1/2$), each user is indifferent between the two platforms and can thus randomize with probability $1/2$ herself.

We also note that when the benefit from joining a more crowded platform is sufficiently greater than the benefit from joining a smaller platform, this may restore the existence of a pure strategy equilibrium.[11]

4.4 Inefficiency

The social surplus of strategy profile $(\boldsymbol{\beta}, \mathbf{a}, \mathbf{p})$ is defined analogously to the equilibrium (social) surplus in the previous section as

$$\text{Social surplus}(\boldsymbol{\beta}, \mathbf{a}) = \mathbb{E}_{\mathbf{b}\sim\boldsymbol{\beta}}\left[\sum_{i\in J_1}\big((1-v_i)\mathcal{I}_i(\mathbf{a}^{J_1})) + c_i(J_1)\big) + \sum_{i\in J_2}\big((1-v_i)\mathcal{I}_i(\mathbf{a}^{J_2})) + c_i(J_2)\big)\right],$$

where the sets J_1 and J_2 are defined based on random variable $\mathbf{b} \sim \boldsymbol{\beta}$.

A natural conjecture is that competition might redress some of the inefficiencies of data markets identified so far, either by increasing data prices or by allowing high-value users to go to a platform where less of their information will be leaked. The next example illustrates that this is not necessarily the case and competition may increase or reduce equilibrium surplus.

Example 6. Consider a setting with $\mathcal{V} = \{1, 2\}$, $\sigma_1^2 = \sigma_2^2 = 1$, $\Sigma_{12} > 0$, and $v_1 < 1$, and take c_i to be a constant function c for all i.

- *Competition improves equilibrium surplus:* Suppose that $v_2 > 1$ is sufficiently large that in equilibrium user 2 never shares her data. Under monopoly, equilibrium data sharing decisions are $a_1^{\mathrm{E}} = 1$ and $a_2^{\mathrm{E}} = 0$ with prices given in Theorem 2. The equilibrium surplus is

$$(1 - v_1)\mathcal{I}_1(a_1 = 1, a_2 = 0) + (1 - v_2)\mathcal{I}_2(a_1 = 1, a_2 = 0) + 2c.$$

  With competition, equilibrium joining decisions are $b_1^{\mathrm{E}} = 1$, $b_2^{\mathrm{E}} = 2$ and equilibrium data sharing decisions are $a_1^{J_1,\mathrm{E}} = 1$ and $a_2^{J_2,\mathrm{E}} = 0$ with prices given in equation (8). The equilibrium surplus is therefore

$$(1 - v_1)\mathcal{I}_1(a_1 = 1, a_2 = 0) + 2c,$$

[11] In Appendix B, we also prove that if all users are low-value, then the setup with competing platforms is a potential game and thus a pure strategy equilibrium exists. Moreover, we show that the social surplus of this equilibrium is always less than the social surplus under monopoly, because competition between the platforms leads to an inefficiently fragmented distribution of low-value users as they try to avoid information about them being leaked by other low-value users.

which is strictly greater than the equilibrium surplus under monopoly.

- *Competition reduces equilibrium surplus:* Suppose that $v_2 < 1$. Under monopoly, we have $a_1^{\mathrm{E}} = 1$ and $a_2^{\mathrm{E}} = 1$ with prices given in Theorem 2. The equilibrium surplus is

$$(1 - v_1)\mathcal{I}_1(a_1 = 1, a_2 = 1) + (1 - v_2)\mathcal{I}_2(a_1 = 1, a_2 = 1) + 2c.$$

  With competition, the equilibrium involves $b_1^{\mathrm{E}} = 1$, $b_2^{\mathrm{E}} = 2$, $a_1^{J_1,\mathrm{E}} = 1$ and $a_2^{J_2,\mathrm{E}} = 1$ with prices given in equation (8). The equilibrium surplus is

$$(1 - v_1)\mathcal{I}_1(a_1 = 1, a_2 = 0) + (1 - v_2)\mathcal{I}_2(a_1 = 0, a_2 = 1) + 2c,$$

  which is strictly less than the surplus under monopoly.

The next theorem, which is the analogue of Theorem 3, characterizes the conditions for efficiency and inefficiency in this case and shows that efficiency now requires more stringent conditions.

Theorem 6. *1. Suppose every high-value user is uncorrelated with all other users. Then the equilibrium is efficient if and only if $c_i(\mathcal{V}) - c_i(\{i\}) \geq v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$ for all $i \in \mathcal{V}^{(l)}$.*

2. Suppose at least one high-value user is correlated (has a non-zero correlation coefficient) with a low-value user. Then there exist $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ and $\underline{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(l)}|}$ such that when $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ and $\mathbf{v}^{(l)} \geq \underline{\mathbf{v}}$ the equilibrium is inefficient.

3. Suppose every high-value user is uncorrelated with all low-value users and at least one high-value user is correlated with another high-value user. Let $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ be the subset of high-value users correlated with at least one other high-value user. Then for each $i \in \tilde{\mathcal{V}}^{(h)}$ there exists $\bar{v}_i > 0$ such that if for any $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$, the equilibrium is inefficient.

The results in this theorem are similar to those in Theorem 3 and again establish that the equilibrium is generally inefficient. One important difference is that the condition for efficiency in the first part has become more stringent with the addition of a restriction on direct benefits from joining platforms with different subsets of users. This is because, in the presence of multiple platforms, not all low-value users may end up joining the same platform since they may try to avoid their information being leaked to the platform by other low-value users (recall, as noted in footnote 10, that information leakage about low-value users is socially beneficial but costly for them as it reduces the prices that the platforms pay for their data). This fragmented allocation of users across platforms is costly if there are gains from being on larger platforms.

The second part of the theorem is also slightly different from Theorem 3 and requires that low-value users have some privacy concerns as well. This is to rule out the possibility that low- and high-value users go to separate platforms, and such an allocation may maximize social surplus despite the loss of direct benefits from forming a larger network on a single platform. However, when low-value users care about their privacy, such an allocation cannot be sustained because they will have an incentive to switch to the platform populated by high-value users where there will be no information leaked about them and they will therefore be able to obtain higher prices for their data.[12]

The next proposition generalizes Proposition 4 and provides an upper bound on social surplus, demonstrating that data markets may easily generate negative social surplus, even in the presence of competition. Since our focus is on the implications of data market, we do not include the direct benefits, the c_i's, from online platforms in this social surplus (since these can be enjoyed by users even if there are no data markets), and refer to it as the "data social surplus".

Proposition 6. *Let $\boldsymbol{\beta}^{\mathrm{E}}$ be the uniform mixed joining strategy. Then*

$$\text{Data social surplus}(\boldsymbol{\beta}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}) \leq \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}) - \frac{1}{2} \sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}).$$

Therefore, Proposition 6 implies that if

$$\sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}) \geq 2 \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}),$$

then shutting down the market for data (while continuing to operate the other services on the online platforms) is beneficial.

Note also that compared to Proposition 4, there is a 1/2 in front of the second term in data social surplus. This reflects the fact that in the mixed strategy joining equilibrium, the set of low-value users who will be on the same platform as a high-value user is random (and will typically be less than the entire set of low-values users). Using the submodularity of leaked information, we show that the expected leakage of user i's information from the data sharing of low-value users is greater than 1/2 times the total leaked information about this user when all low-value users are on the same platform and share their data.

The following example provides a more explicit illustration of the results in Proposition 6.

Example 7. We consider the same setting as Example 4 but with two competing platforms. Figure 4 shows the combinations of v_h and ρ for which the data social surplus is negative. As highlighted in Proposition 6, when the value of privacy for high-value users is high and/or when the correlation between high-value and low-value users is large, the data social surplus is negative.

[12] We should add that for such a fragmented allocation of users across platforms to achieve efficiency, though possible, is very difficult in general.

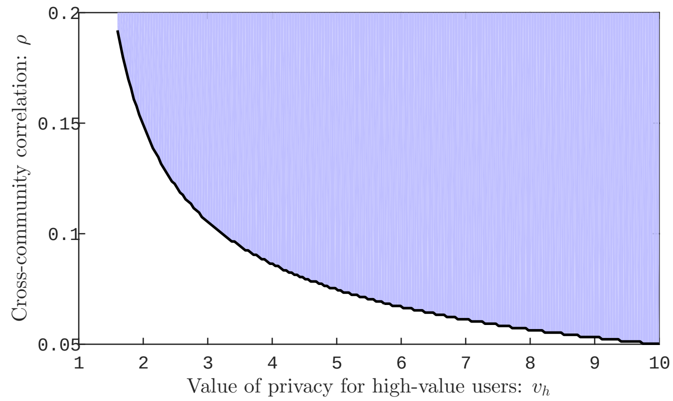


Figure 4: Shaded area shows the pairs of (ρ, v_h) with negative equilibrium surplus in the setting of Example 7.

4.5 Between-Platform Externalities

In this subsection, we generalize our model to allow for between-platform externalities. This captures the fact that some of the loss of privacy may occur for people who are not on the platform, but whose information is revealed to the platform by others' data sharing.

To model this possibility, we now generalize the payoff of users who join platform $k \in \{1,2\}$ to

$$u_i(J_k, a_i, \mathbf{a}_{-i}^{J_k}, \mathbf{p}^{J_k}) = \begin{cases} p_i^{J_k} - v_i \mathcal{I}_i\left(a_i = 1, \mathbf{a}_{-i}^{J_k}\right) - v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right) + c_i(J_k), & a_i = 1, k' \neq k \\ \\ -v_i \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_k}\right) - v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right) + c_i(J_k) & a_i = 0, k' \neq k, \end{cases}$$

where the term $\alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right)$ is leaked information about user i on the platform she has not joined, and the payoff of users who have not joined either of the platforms is

$$-v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_1}\right) - v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_2}\right).$$

In this specification $\alpha \in [0,1]$ captures between-platform externalities. When $\alpha = 0$, we obtain the model studied so far in this section. When $\alpha = 1$, information revealed about an individual who is not on the platform creates the same loss of privacy.

In addition, we assume that platforms also benefit in the same manner from information leaked about individuals who are not their current users, so the payoff of platform $k \in \{1,2\}$ is now

$$U^{(k)}(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}) = \sum_{i \in J_k} \mathcal{I}_i(\mathbf{a}^{J_k}) + \alpha \sum_{i \in J_{k'}} \mathcal{I}_i(\mathbf{a}^{J_k}) - \sum_{i \in J_k:\ a_i^{J_k} = 1} p_i, \qquad k' \neq k, \tag{10}$$

where $\sum_{i \in J_{k'}} \mathcal{I}_i(\mathbf{a}^{J_k})$ is leaked information about users on the other platform.

We next show that our main results generalize to this case.

Theorem 7. *For any joining profile* $\mathbf{b}$ *with the corresponding sets of users* J_1 *and* J_2*, there exists a pure strategy Stackelberg equilibrium with equilibrium prices as in equation (8) and equilibrium payoffs of users joining platform* k *given by*

$$u_i(J_k^{\mathrm{E}}, \mathbf{a}^{J_k^{\mathrm{E}},\mathrm{E}}, \mathbf{p}^{J_k^{\mathrm{E}},\mathrm{E}}) = -v_i\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{J_k,\mathrm{E}}) - v_i\alpha\mathcal{I}_i(\mathbf{a}^{J_{k'},\mathrm{E}}) + c_i(J_k). \tag{11}$$

Moreover, there always exists a mixed strategy joining equilibrium in which all users join each platform with probability 1/2.

Note that the equilibrium price vectors are the same as in equation (8) because the incremental effect of data sharing for a user is the same as before (the information leaked about her on the other platform is unaffected by her data sharing decision). However, because now information about her is leaked by the data shared by users on the other platform, her payoff is less than in equation (9).

The next theorem provides sufficient conditions for the inefficiency of the equilibrium in data markets in this case.

Theorem 8. *1. Suppose that every high-value user is uncorrelated with all other users. Then there exists* $\bar{\alpha}$ *such that for* $\alpha \leq \bar{\alpha}$ *the equilibrium is efficient if and only if* $c_i(\mathcal{V}) - c_i(\{i\}) \geq (1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$ *for all* $i \in \mathcal{V}^{(l)}$.

2. *Suppose that at least one high-value user is correlated (has a non-zero correlation coefficient) with a low-value user. Then there exists* $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ *such that for* $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ *the equilibrium is inefficient.*

3. *Suppose every high-value user is uncorrelated with all low-value users and at least one high-value user is correlated with another high-value user. Let* $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ *be the subset of high-value users correlated with at least one other high-value user. Then for each* $i \in \tilde{\mathcal{V}}^{(h)}$ *there exists* $\bar{v}_i > 0$ *such that if for any* $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$*, the equilibrium is inefficient.*

There are a number of differences from Theorem 6. First, in addition to the conditions in the first part of Theorem 6, we now also require the between-platform spillovers not to be too large — otherwise, the first best may involve splitting low-value users across the two platforms. In addition, the conditions for inefficiency in the second part are slightly weaker because the between-platform externalities make it more likely that the information of high-value users will be leaked inefficiently.

Finally, the next result shows that the data social surplus can again be negative under plausible conditions.

Proposition 7. *Let* $\boldsymbol{\beta}^{\mathrm{E}}$ *be the (uniform) equilibrium mixed joining strategy. Then*

$$\text{Data social surplus}(\boldsymbol{\beta}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}) \leq (1+\alpha)\sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathcal{V}) - \frac{1+\alpha}{2}\sum_{i\in\mathcal{V}^{(h)}}(v_i-1)\mathcal{I}_i(\mathcal{V}^{(l)}).$$

Proposition 7 implies that if

$$\sum_{i\in\mathcal{V}^{(h)}}(v_i-1)\mathcal{I}_i(\mathcal{V}^{(l)})\geq 2\sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathcal{V}),$$

then shutting down the market for data (while continuing to operate the other services provided by online platforms) is once again beneficial.

4.6 Competition over Data Prices

As noted at the beginning of this section, by having users join platforms first, we have removed the ability of online companies to set data prices in order to attract users to their platform. In this subsection, we study this case (with no between-platform spillovers for simplicity). Existence of equilibrium becomes more challenging in this case because of fiercer competition over data and users, but the overall insights are similar.

The exact timing of events is as follows.

1. Platforms simultaneously offer price vectors $\mathbf{p}^1\in\mathbb{R}^n$ and $\mathbf{p}^2\in\mathbb{R}^n$.
2. Users simultaneously decide which platform, if any, to join, i.e., $\mathbf{b}=\{b_i\}_{i\in\mathcal{V}}$ (which determines J_1 and J_2) and whether to share their data.

The payoffs of users and platforms are the same as before. In particular, the payoff of user $i\in\mathcal{V}$ who joins platform $b_i\in\{1,2\}$ is

$$u_i(a_i,b_i,\mathbf{a}_{-i},\mathbf{b}_{-i},\mathbf{p}^1,\mathbf{p}^1)=\begin{cases}p_i^{b_i}-v_i\mathcal{I}_i(\mathbf{a}^{J_{b_i}})+c_i(J_{b_i}), & a_i=1,\\ -v_i\mathcal{I}_i(\mathbf{a}^{J_{b_i}})+c_i(J_{b_i}), & a_i=0,\end{cases}$$

where recall that $\mathbf{a}^{J_k}$ denotes the vector of sharing decisions in the set J_k for $k=1,2$. Assumption 1 ensures that every user joins one of the platforms. The payoff of platform $k\in\{1,2\}$ is

$$U^{(k)}(\mathbf{p}^1,\mathbf{p}^2,\mathbf{a},\mathbf{b})=\sum_{i\,:\,b_i=k}\mathcal{I}_i(\mathbf{a}^{J_k})-\sum_{i\,:\,b_i=k,a_i^{J_k}=1}p_i^1.$$

Definition 4. For given price vectors $\mathbf{p}^1\in\mathbb{R}^n$ and $\mathbf{p}^2\in\mathbb{R}^n$, the joining and sharing profiles $\mathbf{b}$ and $\mathbf{a}$ constitute a user equilibrium if for any $i\in\mathcal{V}$, we have

$$(a_i,b_i)\in\text{argmax}_{a,b}u_i(a,b,\mathbf{a}_{-i},\mathbf{b}_{-i},\mathbf{p}^1,\mathbf{p}^2).$$

Let $\mathcal{A}(\mathbf{p}^1,\mathbf{p}^2)$ denote the set of user equilibria for given price vectors $\mathbf{p}^1$ and $\mathbf{p}^2$. Price vectors $\mathbf{p}^{1,\text{E}}$, $\mathbf{p}^{2,\text{E}}$, joining profile $\mathbf{b}^E$, and sharing profile $\mathbf{a}^{\text{E}}$ constitute a pure strategy equilibrium if $(\mathbf{a}^{\text{E}},\mathbf{b}^{\text{E}})\in$

$\mathcal{A}(\mathbf{p}^{1,\mathrm{E}}, \mathbf{p}^{2,\mathrm{E}})$ and for any $\mathbf{p}$, there exists $(\mathbf{a}, \mathbf{b}) \in \mathcal{A}(\mathbf{p}, \mathbf{p}^{2,\mathrm{E}})$ such that

$$U^{(1)}(\mathbf{p}^{1,\mathrm{E}}, \mathbf{p}^{2,\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{b}^{\mathrm{E}}) \geq U^{(1)}(\mathbf{p}, \mathbf{p}^{2,\mathrm{E}}, \mathbf{a}, \mathbf{b}),$$

and there exists $(\mathbf{a}', \mathbf{b}') \in \mathcal{A}(\mathbf{p}^{1,\mathrm{E}}, \mathbf{p})$ such that

$$U^{(2)}(\mathbf{p}^{1,\mathrm{E}}, \mathbf{p}^{2,\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{b}^{\mathrm{E}}) \geq U^{(2)}(\mathbf{p}^{1,\mathrm{E}}, \mathbf{p}, \mathbf{a}', \mathbf{b}').$$

We next define a mixed strategy equilibrium similarly to before, except that these strategies will define probability distributions over price vectors for the platforms and user actions.[13]

Definition 5. Let $\mathcal{P}$ be the set of probability measures over $\mathbb{R}^n_+$. For any user $i \in \mathcal{V}$, let $\mathcal{A}_i$ be the set of probability measures over $\{0, 1\}$ and $\mathcal{B}_i$ be the set of probability measures over $\{1, 2\}$.

For given price vectors $\mathbf{p}^1 \in \mathbb{R}^n$ and $\mathbf{p}^2 \in \mathbb{R}^n$, the joining and sharing profiles $\boldsymbol{\beta} \in \prod_{i \in \mathcal{V}} \mathcal{B}_i$ and $\boldsymbol{\alpha} \in \prod_{i \in \mathcal{V}} \mathcal{A}_i$ constitute a mixed user equilibrium if for any $i \in \mathcal{V}$, we have

$$u_i(\alpha_i, \beta_i, \boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \mathbf{p}^1, \mathbf{p}^2) \geq u_i(\alpha_i', \beta_i', \boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \mathbf{p}^1, \mathbf{p}^2), \quad \text{for all } \alpha_i' \in \mathcal{A}_i \text{ and } \beta_i' \in \mathcal{B}_i,$$

where $u_i(\alpha_i, \beta_i, \boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \mathbf{p}^1, \mathbf{p}^2) = \mathbb{E}_{a_i \sim \alpha_i, \mathbf{a}_{-i} \sim \boldsymbol{\alpha}_{-i}, b_i \sim \beta_i, \mathbf{b}_{-i} \sim \boldsymbol{\beta}_{-i}} \left[u_i(a_i, b_i, \mathbf{a}_{-i}, \mathbf{b}_{-i}, \mathbf{p}^1, \mathbf{p}^2) \right]$. We let $\mathcal{A}(\boldsymbol{\pi}^1, \boldsymbol{\pi}^2)$ denote the set of mixed strategy user equilibria for given price strategies $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$.

Strategy price profiles $\boldsymbol{\pi}^{k,\mathrm{E}} \in \mathcal{P}$, $k = 1, 2$, joining profile $\boldsymbol{\beta}^{\mathrm{E}}$, and sharing profile $\boldsymbol{\alpha}^{\mathrm{E}}$ constitute a mixed strategy equilibrium if $(\boldsymbol{\alpha}^{\mathrm{E}}, \boldsymbol{\beta}^{\mathrm{E}}) \in \mathcal{A}(\boldsymbol{\pi}^{1,\mathrm{E}}, \boldsymbol{\pi}^{2,\mathrm{E}})$ and for any $\boldsymbol{\pi} \in \mathcal{P}$ there exists $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{A}(\boldsymbol{\pi}, \boldsymbol{\pi}^{2,\mathrm{E}})$ such that

$$\mathbb{E}_{\mathbf{p}^1 \sim \boldsymbol{\pi}^{1,\mathrm{E}}, \mathbf{p}^2 \sim \boldsymbol{\pi}^{2,\mathrm{E}}, \mathbf{a} \sim \boldsymbol{\alpha}^{\mathrm{E}}, \mathbf{b} \sim \boldsymbol{\beta}^{\mathrm{E}}} \left[U^{(1)}(\mathbf{p}^1, \mathbf{p}^2, \mathbf{a}, \mathbf{b}) \right] \geq \mathbb{E}_{\mathbf{p}^1 \sim \boldsymbol{\pi}, \mathbf{p}^2 \sim \boldsymbol{\pi}^{2,\mathrm{E}}, \mathbf{a} \sim \boldsymbol{\alpha}, \mathbf{b} \sim \boldsymbol{\beta}} \left[U^{(1)}(\mathbf{p}^1, \mathbf{p}^2, \mathbf{a}, \mathbf{b}) \right],$$

and there exists $(\boldsymbol{\alpha}', \boldsymbol{\beta}') \in \mathcal{A}(\boldsymbol{\pi}^{1,\mathrm{E}}, \boldsymbol{\pi})$ such that

$$\mathbb{E}_{\mathbf{p}^1 \sim \boldsymbol{\pi}^{1,\mathrm{E}}, \mathbf{p}^2 \sim \boldsymbol{\pi}^{2,\mathrm{E}}, \mathbf{a} \sim \boldsymbol{\alpha}^{\mathrm{E}}, \mathbf{b} \sim \boldsymbol{\beta}^{\mathrm{E}}} \left[U^{(2)}(\mathbf{p}^1, \mathbf{p}^2, \mathbf{a}, \mathbf{b}) \right] \geq \mathbb{E}_{\mathbf{p}^1 \sim \boldsymbol{\pi}^{1,\mathrm{E}}, \mathbf{p}^2 \sim \boldsymbol{\pi}, \mathbf{a} \sim \boldsymbol{\alpha}', \mathbf{b} \sim \boldsymbol{\beta}'} \left[U^{(2)}(\mathbf{p}^1, \mathbf{p}^2, \mathbf{a}, \mathbf{b}) \right].$$

Theorem 9. *There exists a mixed strategy equilibrium strategy.*

We next show that the equilibrium is even more likely to be inefficient when platforms compete using data prices. In particular, in contrast to the settings studied so far, the equilibrium is inefficient not only when high-value users are correlated with other users, but also when there is

[13] Note that in the models analyzed so far, after the platform (or platforms) set data prices, users no longer had the option of switching to another platform, and we focused on the Stackelberg equilibrium where the platform set prices anticipating user choices and selected the most advantageous user equilibrium for itself (when there were multiple user equilibria). This meant that an equilibrium data price ensured a (weakly) greater payoff for the platform than any other price for any other user equilibrium. Because users now make their joining decisions after price offers, we use the standard Nash equilibrium notion and require that for each platform and any other price than its equilibrium price there exists a user equilibrium in which the platform's payoff is no greater than its equilibrium payoff.

correlation only among low-value users. For this theorem, let us define:

$$\delta = \min_{i, T \subset \mathcal{V}} c_i(\mathcal{V}) - c_i(T) \text{ and } \Delta = \max_{i, T \subseteq \mathcal{V}} c_i(\mathcal{V}) - c_i(T).$$

Theorem 10. *1. Suppose every user is uncorrelated with all other users. Then the equilibrium is efficient.*

2. Suppose that every high-value user is uncorrelated with all other users, but at least two low-value users are correlated with each other. Then there exist $\underline{\delta}$, $\bar{\Delta}$, $\tilde{\Delta}$, $\bar{\mathbf{v}}$, and $\tilde{\mathbf{v}}$ such that:

2-1) If $\delta \geq \underline{\delta}$, the equilibrium is efficient.

2-2) If $\Delta \leq \bar{\Delta}$ and $\mathbf{v}^{(l)} \leq \bar{\mathbf{v}}$, the equilibrium is efficient.

2-3) If $\Delta \leq \tilde{\Delta}$ and $\mathbf{v}^{(l)} \geq \tilde{\mathbf{v}}$, the equilibrium is inefficient.

3. Suppose that at least one high-value user is correlated with a low-value user. Then there exist $\tilde{\delta} > \bar{\Delta} > \bar{\delta} > 0$, $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$, and $\underline{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(l)}|}$ such that:

3-1) If $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$, $\mathbf{v}^{(l)} \geq \underline{\mathbf{v}}$, $\Delta \leq \bar{\Delta}$, and $\delta \geq \bar{\delta}$, the equilibrium is inefficient.

3-2) If $\delta \geq \tilde{\delta}$, the equilibrium is efficient.

The first part is straightforward: without correlation there is no data externality, ensuring efficiency.

The second part is new relative to our previous results: now the equilibrium is inefficient even when high-value users are uncorrelated with all other users. This inefficiency is caused by competition using data prices. Since there is no correlation between high-value and low-value users, the first best involves all low-value users sharing their data and all (high-value and low-value) users joining the same platform in order to benefit from the highest joining values. However, we show in part 2.3 that such an allocation is not an equilibrium, because the other platform can attract some of the low-value users who can benefit by having less of their information leaked by other low-value users (even though information leakage about these users is socially beneficial, it is privately costly for them). This leads to a fragmented distribution of users across platforms, leading to inefficiency. Parts 2.1 and 2.2 provide conditions for efficiency in terms of the c function being sufficiently steep or the privacy concerns of low-value users being sufficiently weak.

Part 3.1 of the theorem is similar to our other inefficiency results. In this case, in the first best all users join the same platform (because the c function is sufficiently steep), but only low-value users uncorrelated with high-value users share their data (because $\mathbf{v}^{(h)}$ is sufficiently high). We show, however, that this allocation cannot be an equilibrium because the other platform can deviate and attract a subset of low-value users and induce them to share their data. In part 3.2 the first best is, once again, for all users to join one of platforms. But now because the joining values are even steeper, the other platform can no longer attract a subset of these users, while the threat of all users switching to this other platform supports the first-best allocation (though there also exist inefficient equilibria in this case). Finally, we demonstrate in Appendix B.3 that when high-value users are uncorrelated with low-value users but correlated among themselves, the equilibrium may or may not be efficient.

5 Unknown Valuations

Our analysis has so far assumed that platforms know the value of privacy of different users. In this section, we adopt the more realistic assumption that they do not know the exact valuations of users, but understand that the value of privacy of user i, v_i, has a distribution represented by the cumulative distribution function F_i and density function f_i (with upper support denoted by $v^{\max}$). Users know their own value of privacy. For simplicity, we focus on the case of a monopoly platform and show how the platform can design a mechanism to elicit this information (in the form of users reporting their value of privacy). We then show that all of the main insights from our analysis generalize to this case.

5.1 Second-Best Characterization

We first characterize the "second best" which takes into account that the value of privacy of each user is their private information, and then show that the second best coincides with the first best.

Proposition 8. *Let $\mathbf{v}$ be the reported vector of values of privacy. Then the pricing scheme*

$$p_i(\mathbf{v}) = \left(\mathcal{I}_i(\mathbf{a}(\mathbf{v})) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a}(\mathbf{v}))\right) - \min_{\mathbf{a}\in\{0,1\}^n}\left(\mathcal{I}_i(\mathbf{a}) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a})\right),$$

where $\mathbf{a}(\mathbf{v}) = \text{argmax}_{\mathbf{a}\in\{0,1\}^n}\sum_{i\in\mathcal{V}}(1-v_i)\mathcal{I}_i(\mathbf{a})$ incentivizes users to report their value of privacy truthfully, and thus the second best coincide with the first best.

This mechanism is a variation of Vickery-Clarke-Grove mechanism (Vickrey [1961], Clarke [1971], Groves [1973]). In particular, for any $i \in \mathcal{V}$ the price offered to user i is equal to the surplus of all other users on the platform when user i is present minus by the surplus when user i is absent. The second term in the price $p_i(\mathbf{v})$ can be any function of the values $\mathbf{v}_{-i}$, and the choice specified in Proposition 8 guarantees that the prices are nonnegative.

5.2 Equilibrium Characterization

The next definition generalizes our notion of equilibrium to this incomplete information setup. It is simplified by making use of the revelation principle, which enables us to focus on incentive compatible price and action profiles.

Definition 6. An equilibrium is a pair $(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}})$ of functions of the reported valuations $\mathbf{v} = (v_1, \ldots, v_n)$ such that each user reports its true value and the expected payoff of the platform is maximized. That is,

$$(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}) = \max_{\mathbf{a}:\mathbb{R}^n \to \{0,1\}^n, \mathbf{p}:\mathbb{R}^n \to \mathbb{R}^n} \mathbb{E}_{\mathbf{v}} \left[\sum_{i=1}^{n} \mathcal{I}_i(\mathbf{a}(\mathbf{v})) - \sum_{i:\, a_i(\mathbf{v})=1} p_i(\mathbf{v}) \right]$$

$$p_i(\mathbf{v}) - v_i \mathcal{I}_i(\mathbf{a}(\mathbf{v})) \geq p_i(\mathbf{v}_{-i}, v_i') - v_i \mathcal{I}_i(\mathbf{a}(\mathbf{v}_{-i}, v_i')), \quad \text{for all } v_i', \mathbf{v}, \text{ and } i \in \mathcal{V}.$$

We also impose the following standard assumption on the (reversed) hazard rate and maintain it throughout this section without explicitly mentioning it.

Assumption 2. *For all $i \in \mathcal{V}$, the function $\Phi_i(v) = v + \frac{F_i(v)}{f_i(v)}$ is nondecreasing.*

Here $\Phi_i(v)$ is the well-known "virtual value" in incomplete information models, representing the additional rent that the agent will capture in incentive-compatible mechanisms. In our setting, it will enable users to obtain more of the surplus the platform infers from their data.

A sufficient condition for Assumption 2 to hold is for the reversed hazard rate $f_i(x)/F_i(x)$ to be nonincreasing. This requirement is satisfied for a variety of distributions such as uniform and exponential (see e.g., Burkschat and Torrado [2014]).

Theorem 11. *For any reported vector of values $\mathbf{v}$, the equilibrium is given by*

$$\mathbf{a}^{\mathrm{E}}(\mathbf{v}) = \mathit{argmax}_{\mathbf{a} \in \{0,1\}^n} \sum_{i=1}^{n} (1 - \Phi_i(v_i)) \mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i) \mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0),$$

and

$$p_i^{\mathrm{E}}(v_i) = \int_{v}^{v_{\max}} \left(\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(x, \mathbf{v}_{-i})) - \mathcal{I}_i(\mathbf{a}_{-i}^{\mathrm{E}}(x, \mathbf{v}_{-i}), a_i = 0) \right) dx$$
$$+ v_i \left(\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(v_i, \mathbf{v}_{-i})) - \mathcal{I}_i(\mathbf{a}_{-i}^{\mathrm{E}}(v_i, \mathbf{v}_{-i}), a_i = 0) \right).$$

Moreover, all users report truthfully and thus the expected payoff of the platform is

$$\mathbb{E}_{\mathbf{v}} \left[\max_{\mathbf{a} \in \{0,1\}^n} \sum_{i=1}^{n} (1 - \Phi_i(v_i)) \mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i) \mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0) \right].$$

In the following corollary, we compare the expected payoff of the incentive compatible mechanism in Theorem 11 for the platform with its payoff in the case where these values are known.

Corollary 3. *The expected payoff of the platform with known valuations is (weakly) higher than the case with unknown valuations, and the difference is at least*

$$\mathbb{E}_{\mathbf{v}} \left[\sum_{i=1}^{n} \frac{F_i(v_i)}{f_i(v_i)} \left(\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(\mathbf{v})) - \mathcal{I}_i(\mathbf{a}_{-i}^{\mathrm{E}}(\mathbf{v}), a_i = 0) \right) \right],$$

where

$$\mathbf{a}^{\mathrm{E}}(\mathbf{v}) = \textit{argmax}_{\mathbf{a}\in\{0,1\}^n} \sum_{i=1}^{n} (1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0).$$

This corollary shows that incomplete information about user valuations reduces the payoff of the platform and the extent of this payoff reduction depends on the platform's uncertainty about user values, captured by F_i/f_i.

5.3 Inefficiency

We first establish that the equilibrium is inefficient under fairly plausible conditions in this incomplete information setting as well. The main difference from our analysis so far is that another relevant set is the subset of low-value users with virtual value of privacy less than one, i.e., $\Phi_i(v_i) \leq 1$. For the next theorem, we use the notation $\mathcal{V}_{\Phi}^{(l)} = \{i \in \mathcal{V} : \ \Phi_i(v_i) \leq 1\}$ to denote this set of users.

Theorem 12. *1. Suppose high-value users are uncorrelated with all other users and $\mathcal{V}^{(l)} = \mathcal{V}_{\Phi}^{(l)}$. Then the equilibrium is efficient.*

2. *Suppose some high-value users (those in $\mathcal{V}^{(h)}$) are correlated with users in $\mathcal{V}_{\Phi}^{(l)}$. Then there exists $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ such that for $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ the equilibrium is inefficient.*

3. *Suppose every high-value user is uncorrelated with all users in $\mathcal{V}_{\Phi}^{(l)}$, but users in a nonempty subset $\hat{\mathcal{V}}^{(l)}$ of $\mathcal{V}^{(l)} \setminus \mathcal{V}_{\Phi}^{(l)}$ are correlated with at least one high-value user. Then there exist $\bar{\mathbf{v}}$ and $\tilde{v}$ such that if $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ and $v_i < \tilde{v}$ for some $i \in \hat{\mathcal{V}}^{(l)}$, the equilibrium is inefficient.*

4. *Suppose every high-value user is uncorrelated with all low-value users and at least one high-value user is correlated with another high-value user. Let $\tilde{\mathcal{V}}^{(h)} \subseteq \mathcal{V}^{(h)}$ be the subset of high-value users correlated with at least one other high-value user. Then for each $i \in \tilde{\mathcal{V}}^{(h)}$ there exists $\bar{v}_i > 0$ such that if for any $i \in \tilde{\mathcal{V}}^{(h)}$ $v_i < \bar{v}_i$, the equilibrium is inefficient.*

The inefficiency results in this theorem again have clear parallels to those in Theorem 3, but with some notable differences. First, efficiency now requires all low-value users to also have virtual valuations that are less than one, since otherwise user incentive compatibility constraints prevent the efficient allocation. Second, the conditions for inefficiency are slightly different depending on whether high-value users are correlated with low-value users whose virtual valuations are less than one or greater than one.

We next present the analogue of Proposition 4 in this setting. First, note that for a given vector of user values $\mathbf{v}$ a similar argument to that of Lemma 3 shows that a user i with $\Phi_i(v_i) \leq 1$ will share her data in equilibrium. Using this observation, we can establish the next proposition.

Proposition 9. *Consider a setting with unknown valuations. For a given $\mathbf{v}$, we have*

$$\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) \leq \sum_{i\,:\,v_i\in\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathcal{V}) - \sum_{i\,:\,v_i\in\mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}_{\Phi}^{(l)}).$$

Proposition 9 is similar to Proposition 4 and provides a sufficient condition for equilibrium surplus to be negative. The only difference is that the lower bound on the negative (second) term is evaluated for information leaked by users in $\mathcal{V}_{\Phi}^{(l)}$ (rather than those in $\mathcal{V}^{(l)}$). This is because, with incomplete information, the platform has to compensate users according to their virtual value of privacy and may find it too expensive to purchase the data of low-value users with $\Phi_i(v_i) > 1$. Indeed, because $\Phi_i(v_i) \geq v_i$, we have $\mathcal{V}_{\Phi}^{(l)} \subseteq \mathcal{V}^{(l)}$, and thus, given $\mathbf{v}$, if the equilibrium surplus with unknown valuations is negative ($\sum_{i \in \mathcal{V}^{(h)}}(v_i - 1)\mathcal{I}_i(\mathcal{V}_{\Phi}^{(l)}) > \sum_{i \in \mathcal{V}^{(l)}}(1 - v_i)\mathcal{I}_i(\mathcal{V})$), then the equilibrium surplus with known valuations (as given in Proposition 4) is also negative.

6 Regulation

The inefficiencies documented so far raise the question of whether certain types of government policies or regulations could help data markets function better. We briefly address this question in this section. We first discuss taxes and then turn to a regulation scheme based on "de-correlation" to reduce the informativeness of the data of users about others. For simplicity, we focus on the case of a single platform with complete information.

6.1 Taxation

The next proposition shows that a simple Pigovian tax scheme, using personalized taxes on data transactions, can restore the first best. For this purpose we assume that the government can impose a tax t_i on user i selling her data to the platform.

Proposition 10. *Let* $\mathbf{a}^{\mathrm{W}}$ *denote the first best. Then personalized taxes satisfying*

$$t_i > \sum_{j \in \mathcal{V}^{(l)}} \sigma_j^2 + \sum_{j \in \mathcal{V}^{(h)}} v_j \sigma_j^2 \qquad \textit{for } a_i^{\mathrm{W}} = 0$$
$$t_i = 0 \qquad \textit{for } a_i^{\mathrm{W}} = 1,$$

implements the first-best action profile $\mathbf{a}^{\mathrm{W}}$ *as the unique equilibrium.*

The idea of the proposition is straightforward. We first prove that not taxing users who should be sharing in the first best is sufficient to ensure that they share in the post-tax equilibrium as well regardless of the sharing decisions of the rest of the users. Then imposing prohibitive taxes on the data transactions of users who should not be sharing implements the first best.

Proposition 10 characterizes a set of Pigovian taxes implementing the first best, but these taxes vary across individuals, which presupposes a huge amount of information on the part of the planner/tax authority. A natural question is whether a uniform tax scheme can also improve over the equilibrium allocation. The next proposition provides a simple (and obvious) condition under which this is the case:

Proposition 11. *Suppose equilibrium surplus is negative. Then a uniform and sufficiently high tax on data transactions shuts down the data market and improves the social surplus.*

The next example, however, shows that, beyond this simple case with negative equilibrium surplus, there is no guarantee that uniform taxes on data transactions improve welfare. This is because such taxes may prevent beneficial data trades.

Example 8. Consider a network with three users and suppose that $\sigma_i^2 = 1$ for all $i \in \{1, 2, 3\}$, the correlation between X_1 and the other two random variable is 0.4, and the correlation between X_2 and X_3 is 0. We also set $v_1 = 0.1$, $v_2 = 0.5$, and $v_3 = 8$. The first best is $\mathbf{a}^{\mathrm{W}} = (0, 1, 0)$. If the uniform tax is $t \leq 0.25$, the equilibrium is $\mathbf{a}^{\mathrm{E}} = (1, 1, 0)$ whose surplus is 0.0625. For $0.25 < t < 0.61$, the equilibrium is $\mathbf{a}^{\mathrm{E}} = (1, 0, 0)$, with social surplus -0.15. Finally, for $t \geq 0.61$, the equilibrium is $\mathbf{a}^{\mathrm{E}} = (0, 0, 0)$ with surplus 0. Therefore, no uniform taxation scheme can improve equilibrium surplus in this game.

Intuitively, when the tax is small, user 1 continues to share her data. When the tax rate takes intermediate value, now both users 1 and 2 share, and this leaks considerable information about user 3. With very high taxes, nobody shares their data. In none of these cases can we implement the allocation where only user 2 shares.

6.2 Mediated Data Sharing and De-correlation

In this subsection, we propose a different approach to improving the efficiency of data markets. Our analysis has clarified that a main source of inefficiency in such markets are the data externalities created by the correlation between the information/types of different users. Our approach is based on the observation that it is possible to transform the data of different users in such a manner as to remove the correlation that is at the root of these data externalities. We refer to such a scheme as *de-correlation*.

Suppose that instead of sharing their data with the platform, users share their data with a (trusted third-party) mediator, who can either not share these data with the platform (as instructed) or transform them before revealing them to the platform.[14] Recall that user i's data are represented by $S_i = X_i + Z_i$. The main idea is that the mediator collects all the data from the users and then computes transformed variables for each user removing the correlation with the information of other users and only shares the transformed data of those who are willing to sell their data (but utilizes the data of others for removing the correlation with their information).[15]

Formally, we consider the following de-correlation scheme: $\tilde{\mathbf{S}} = \Sigma^{-1}\mathbf{S}$ where $\mathbf{S} = (S_1, \ldots, S_n)$ is the vector of data of all users. Clearly, $\tilde{\mathbf{S}}$ is jointly normal and has the property that if user i does

[14] Obviously, such a scheme can only work if the mediator is fully reliable and trusted, and this is an important constraint in practice, which we are ignoring in this paper.

[15] In practice, it may be more relevant to remove the correlation between a user's data and the average data of different user types. In that case, we can partition the set of users into K cells and apply this de-correlation procedure to the average data of cells.

not share her data, then the data of other users leak no information about user i's type. This is formally stated in the next lemma.

Lemma 5. *With de-correlation, for any action profile* $\mathbf{a} \in \{0,1\}^n$*, the leaked information about user* i *is*

$$\widetilde{\mathcal{I}}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}\left[\left(X_i - \hat{x}_i\left(\tilde{\mathbf{S}}_{\mathbf{a}}\right)\right)^2\right] = \begin{cases} 0, & a_i = 0, \\ \mathcal{I}_i(a_i, \mathbf{a}_{-i}), & a_i = 1. \end{cases}$$

This lemma clarifies our claim in the Introduction and shows that the de-correlation scheme leaves information leaked about the user sharing her data the same, but removes the leakage about users who are not sharing their data.

We next characterize the equilibrium pricing, denoted by $\tilde{\mathbf{p}}^{\mathrm{E}}$, and sharing profile, denoted by $\tilde{\mathbf{a}}^{\mathrm{E}}$, with this transformation.

Theorem 13. *The equilibrium sharing profile after de-correlation is given by*

$$\tilde{\mathbf{a}}^{\mathrm{E}} = \mathit{argmax}_{\mathbf{a}\in\{0,1\}^n} \sum_{i\in\mathcal{V}} (1 - v_i)\widetilde{\mathcal{I}}_i(\mathbf{a}),$$

with prices $\tilde{p}_i^{\mathrm{E}} = v_i \widetilde{\mathcal{I}}_i(\tilde{\mathbf{a}}^{\mathrm{E}})$ *for any* $i \in \mathcal{V}$ *such that* $\tilde{a}_i^{\mathrm{E}} = 1$.

Theorem 13 follows from the fact that, with de-correlation, there is no information leakage about those who do not share, and therefore they do not contribute to the platform's payoff. Moreover, the price offered to users who share must make them indifferent between sharing and not sharing and thus give them zero payoff (which they can guarantee by not sharing).

We next show that the de-correlation scheme always improves equilibrium surplus and eliminates cases where the social surplus is negative.

Theorem 14. *Let* $(\tilde{\mathbf{a}}^{\mathrm{E}}, \tilde{\mathbf{p}}^{\mathrm{E}})$ *and* $(\mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}})$ *denote the equilibrium with and without the de-correlation scheme, respectively. Then*

$$\text{Social surplus}(\tilde{\mathbf{a}}^{\mathrm{E}}) \geq \max\left\{\text{Social surplus}(\mathbf{a}^{\mathrm{E}}), 0\right\}.$$

That equilibrium surplus increases after de-correlation is a consequence of the fact that in the original equilibrium the contribution of high-value users (who do not share) to social surplus is less than or equal to zero, while after de-correlation their contribution to social surplus is greater than or equal to zero. Moreover, because there are no users with negative contribution to social surplus after de-correlation, equilibrium surplus is always positive. This observation also implies that the de-correlation scheme outperforms policies that shut down data markets — since instead of achieving zero equilibrium surplus by shutting down these markets, e.g., as in Proposition 11, this scheme always guarantees positive social surplus.

Open questions include whether such de-correlation schemes are excessively complex to be implemented in practice and how non-mediated information sharing between platforms and users can be prevented (since otherwise platforms can partially undo the de-correlation implemented by the mediator).

7 Conclusion

Because data generated by economic agents are useful for solving economic, social, or technical problems facing others in society and for designing or inventing new products and services, much of economic analysis in this area argues that the market may produce too little data. This paper develops the perspective that, in the presence of privacy concerns of some agents, the market may generate too much data. Moreover, because the data of a subset of users reveal information about other users, the market price of data tends to be depressed, creating the impression that users do not value their privacy much. The depressed market price of data and excessive data generation are intimately linked.

We exposit these ideas in a simple model in which a platform wishes to estimate the types of a collection of users, and each user has personal data (based on their preferences, past behavior, and contacts) which are correlated both with their type and with the data and types of other users. As a result, when a user decides to share her data with the platform, this enables the platform to improve its estimate of other users' types. We model the market for data by allowing the platform to offer prices (or other services) in exchange of data.

We prove the existence of an equilibrium in the data market and show that there will be too much data shared on the platform and the price of data will be excessively depressed. The result that the platform acquires too much data is a direct consequence of the externalities from the data of others. The root cause of depressed data prices is the submodularity of leaked information: when data sharing by other users already compromises the information of an individual, she has less incentive to protect her data and privacy. We further show that under some simple conditions the social surplus generated by data markets is negative, meaning that shutting down data markets improves (utilitarian) social welfare.

We extend these results to a setting with multiple platforms. Various different types of competition between platforms do not alter the fundamental forces leading to too much data sharing and excessively low prices of data. In fact, competition may make inefficiencies worse. This is in part because more data may be shared in the presence of competition, and also because the desire of some users to avoid excessive data sharing about them may lead to an inefficiently fragmented distribution of users across platforms, even when network externalities would be better exploited by having all users join the same platform. We also extend these results to a setting in which the value of privacy of different users are their private information.

Excessive data sharing may call for policy interventions to correct for the externalities and the excessively low prices of data. Individual-specific (Pigovian) taxes on data transactions can restore the first best. More interestingly, we propose a scheme based on mediated-data sharing that can improve welfare. In particular, in our baseline model, when equilibrium surplus is negative, shutting down data markets, for example with high uniform taxes on all data transactions, would improve welfare. But this prevents the sharing of the data of users with low value of privacy or high benefits from goods and services that depend on the platform accessing their data. We show that if user data are first shared with a mediator which transforms them before revealing them to the platform, the correlation of the data with the information of privacy-conscious users can be eliminated, and this would improve welfare relative to the option of shutting off data markets altogether.

We view our work as part of an emerging literature on data markets and the economics of privacy. Several interesting areas of research are suggested by our results. First, it is important to develop models of the marketplace for data that allow for richer types of competition between different platforms. Second, our modeling of privacy and the use of data by the platform has been reduced-form. Distinguishing the uses of personal data for price discrimination, advertising, and designing of new products and services could lead to additional novel insights. For example, it may enable an investigation of whether applications of personal data for designing personalized services can be unbundled from their use for intrusive marketing, price discrimination, or misleading advertising. Third, we only touched upon the possibility of designing new mechanisms for improving the functioning of data markets while reducing data externalities. Our proposed mechanism can be simplified and made more practical, for example, by aiming to remove the correlation between different user classes, as noted above, or by focusing on only some types of data. Other mediated data sharing arrangements or completely new approaches to this problem could be developed as well, but should take into account the possibility that third parties may not be fully trustworthy either. Finally, our result that market prices, or current user actions for protecting privacy, do not reveal the value of privacy highlights the need for careful empirical analysis documenting and estimating the value of data to platforms and the value that users attach to their privacy in the presence of data externalities.

Appendix A

In this part of the Appendix, we provide some of the proofs omitted from the text. Remaining proofs and details of several of the examples discussed in the text are presented in the online Appendix B.

A.1 Proofs

Proof of Proposition 1

Recall that $\mathbf{a}^{\mathrm{W}}$ denotes the first best. For any $i \in \mathcal{V}$ we have $a_i^{\mathrm{W}} = 1$ if and only if $\text{Social surplus}(\mathbf{a}_{-i}^{\mathrm{W}}, a_i = 1) \geq \text{Social surplus}(\mathbf{a}_{-i}^{\mathrm{W}}, a_i = 0)$. Substituting the expression for the social surplus into this equation yields

$$\sum_{j\in\mathcal{V}}(1-v_j)\left(\mathcal{I}_j(\mathbf{a}^{\mathrm{W}}_{-i},a_i=1)-\mathcal{I}_j(\mathbf{a}^{\mathrm{W}}_{-i},a_i=0)\right)\geq 0. \tag{A-1}$$

Conditional on the data provided by other users, i.e., $k\neq i$ for which $a^{\mathrm{W}}_k=1$, (X_j,X_i) are jointly normal and their covariance matrix is given by

$$\begin{pmatrix}\sigma_j^2-\mathcal{I}_j(\mathbf{a}^{\mathrm{W}}_{-i},a_i=0) & \mathrm{Cov}(X_i,X_j\mid \mathbf{a}^{\mathrm{W}}_{-i},a_i=0)\\ \mathrm{Cov}(X_i,X_j\mid \mathbf{a}^{\mathrm{W}}_{-i},a_i=0) & 1+\sigma_i^2-\mathcal{I}_i(\mathbf{a}^{\mathrm{W}}_{-i},a_i=0)\end{pmatrix}.$$

Therefore, if in addition to users $k\neq i$ for which $a^{\mathrm{W}}_k=1$, user i also shares her data, then the leaked information of user j becomes

$$\mathcal{I}_j(\mathbf{a}^{\mathrm{W}}_{-i},a_i=1)=\mathcal{I}_j(\mathbf{a}^{\mathrm{W}}_{-i},a_i=0)+\frac{\mathrm{Cov}(X_i,X_j\mid \mathbf{a}^{\mathrm{W}}_{-i},a_i=0)^2}{1+\sigma_i^2-\mathcal{I}_i(\mathbf{a}^{\mathrm{W}}_{-i},a_i=0)}. \tag{A-2}$$

Substituting equation (A-2) into equation (A-1) completes the proof. ■

Proof of Lemma 1

Part 1, Monotonicity: In order to show that leaked information is monotonically increasing in the set of users who share, it suffices to establish that for any $i,j\in\mathcal{V}$ and $\mathbf{a}_{-j}\in\{0,1\}^{n-1}$ we have $\mathcal{I}_i(a_j=1,\mathbf{a}_{-j})\geq\mathcal{I}_i(a_j=0,\mathbf{a}_{-j})$. We next consider the two possible cases where $i=j$ and $i\neq j$ and show this inequality.

- $i=j$: conditional on shared data, the joint distribution of (X_i,S_i) is normal with covariance matrix $\begin{pmatrix}\hat\sigma_i^2 & \hat\sigma_i^2\\ \hat\sigma_i^2 & 1+\hat\sigma_i^2\end{pmatrix}$, where $\hat\sigma_i^2=\mathbb{E}[X_i^2\mid\mathbf{a}_{-j}]$. We have $\mathcal{I}_i(a_i=1,\mathbf{a}_{-i})=\sigma_i^2-\left(\hat\sigma_i^2-\frac{\hat\sigma_i^4}{1+\hat\sigma_i^2}\right)\geq\sigma_i^2-\hat\sigma_i^2=\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$, completing the proof of this part.
- $i\neq j$: conditional on shared data, the joint distribution of (X_i,S_j) is normal with covariance matrix $\begin{pmatrix}\hat\sigma_i^2 & \hat\Sigma_{ij}\\ \hat\Sigma_{ij} & 1+\hat\sigma_j^2\end{pmatrix}$, where $\hat\sigma_i^2=\mathbb{E}[X_i^2\mid\mathbf{a}_{-j}]$, $\hat\sigma_j^2=\mathbb{E}[X_j^2\mid\mathbf{a}_{-j}]$, and $\hat\Sigma_{ij}=\mathbb{E}[X_iX_j\mid\mathbf{a}_{-j}]$. We have $\mathcal{I}_i(a_j=1,\mathbf{a}_{-j})=\sigma_i^2-\left(\hat\sigma_i^2-\frac{\hat\Sigma_{ij}^2}{1+\hat\sigma_j^2}\right)\geq\sigma_i^2-\hat\sigma_i^2=\mathcal{I}_i(a_j=0,\mathbf{a}_{-j})$, completing the proof of the monotonicity.

Part 2, Submodularity: We first introduce some additional notation for this proof. For any pair $i,j\in\mathcal{V}$, $\mathbf{a}_{-\{i,j\}}$ is the collection of all users' actions except for user i and user j. To prove this part, it suffices to establish that for any $\mathbf{a}_{-\{i,j\}}\in\{0,1\}^{n-2}$, we have

$$\begin{aligned}&\mathcal{I}_j(\mathbf{a}_{-\{i,j\}},a_j=1,a_i=0)-\mathcal{I}_j(\mathbf{a}_{-\{i,j\}},a_j=0,a_i=0)\\ &\geq\mathcal{I}_j(\mathbf{a}_{-\{i,j\}},a_j=1,a_i=1)-\mathcal{I}_j(\mathbf{a}_{-\{i,j\}},a_j=0,a_i=1).\end{aligned}$$

Conditional on $\mathbf{a}_{-\{i,j\}}$, (X_j, S_j, S_i) has a normal distribution with covariance matrix

$$\begin{pmatrix} \hat{\sigma}_j^2 & \hat{\sigma}_j^2 & \hat{\Sigma}_{ij} \\ \hat{\sigma}_j^2 & 1+\hat{\sigma}_j^2 & \hat{\Sigma}_{ij} \\ \hat{\Sigma}_{ij} & \hat{\Sigma}_{ij} & 1+\hat{\sigma}_i^2 \end{pmatrix},$$

where $\hat{\sigma}_i^2 = \mathbb{E}[X_i^2 \mid \mathbf{a}_{-\{i,j\}}]$, $\hat{\sigma}_j^2 = \mathbb{E}[X_j^2 \mid \mathbf{a}_{-\{i,j\}}]$, and $\hat{\Sigma}_{ij} = \mathbb{E}[X_i X_j \mid \mathbf{a}_{-\{i,j\}}]$. Note that in writing this matrix, we are using the fact that the correlation between X_i and S_j is the same as the correlation between S_i and S_j (this holds because $S_i = X_i + Z_i$ for some independent noise Z_i). Based on this covariance matrix,

$$\begin{aligned} \mathcal{I}_j(\mathbf{a}_{-\{i,j\}}, a_j = 1, a_i = 0) - \mathcal{I}_j(\mathbf{a}_{-\{i,j\}}, a_j = 0, a_i = 0) &= \left(\sigma_j^2 - \left(\hat{\sigma}_j^2 - \frac{\hat{\sigma}_j^4}{1+\hat{\sigma}_j^2}\right)\right) - \left(\sigma_j^2 - \hat{\sigma}_j^2\right) \\ &= \frac{\hat{\sigma}_j^4}{1+\hat{\sigma}_j^2}. \end{aligned} \tag{A-3}$$

We also have

$$\begin{aligned} &\mathcal{I}_j(\mathbf{a}_{-\{i,j\}}, a_j = 1, a_i = 1) - \mathcal{I}_j(\mathbf{a}_{-\{i,j\}}, a_j = 0, a_i = 1) \\ &= \left(\sigma_j^2 - \left(\hat{\sigma}_j^2 - (\hat{\sigma}_j^2, \hat{\Sigma}_{ij}) \begin{pmatrix} 1+\hat{\sigma}_j^2 & \hat{\Sigma}_{ij} \\ \hat{\Sigma}_{ij} & 1+\hat{\sigma}_i^2 \end{pmatrix}^{-1} (\hat{\sigma}_j^2, \hat{\Sigma}_{ij})^T \right)\right) - \left(\sigma_j^2 - \left(\hat{\sigma}_j^2 - \frac{\hat{\Sigma}_{ij}^2}{1+\hat{\sigma}_i^2}\right)\right) \\ &= \frac{\hat{\sigma}_j^4(1+\hat{\sigma}_i^2) + \hat{\Sigma}_{ij}^2(1+\hat{\sigma}_j^2) - 2\hat{\Sigma}_{ij}^2\hat{\sigma}_j^2}{(1+\hat{\sigma}_i^2)(1+\hat{\sigma}_j^2) - \hat{\Sigma}_{ij}^2} - \frac{\hat{\Sigma}_{ij}^2}{1+\hat{\sigma}_i^2}. \end{aligned} \tag{A-4}$$

Comparing (A-3) and (A-4), the submodularity of leaked information becomes equivalent to

$$\hat{\sigma}_j^4(1+\hat{\sigma}_i^2) + \hat{\Sigma}_{ij}^2(1+\hat{\sigma}_j^2) \leq 2\hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2),$$

which holds because

$$\begin{aligned} &\hat{\sigma}_j^4(1+\hat{\sigma}_i^2) + \hat{\Sigma}_{ij}^2(1+\hat{\sigma}_j^2) \leq \hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2) + \hat{\Sigma}_{ij}^2(1+\hat{\sigma}_j^2) \\ &\overset{(a)}{\leq} \hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2) + \hat{\sigma}_i^2\hat{\sigma}_j^2(1+\hat{\sigma}_j^2) \leq \hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2) + \hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2) = 2\hat{\sigma}_j^2(1+\hat{\sigma}_j^2)(1+\hat{\sigma}_i^2), \end{aligned}$$

where (a) follows from $\hat{\Sigma}_{ij}^2 \leq \hat{\sigma}_i^2\hat{\sigma}_j^2$. ■

Proof of Lemma 2

Using Lemma 1, we first establish that the game is supermodular. The rest of the proof follows from Tarski's fixed point theorem. Specifically, for any $i \in \mathcal{V}$, we prove that the game has increasing differences property in $(a_i, \mathbf{a}_{-i})$, i.e., if $\mathbf{a}'_{-i} \geq \mathbf{a}_{-i}$ then we have

$$u_i(a_i = 1, \mathbf{a}'_{-i}) - u_i(a_i = 0, \mathbf{a}'_{-i}) \geq u_i(a_i = 1, \mathbf{a}_{-i}) - u_i(a_i = 0, \mathbf{a}_{-i}).$$

We can write

$$\begin{aligned} &u_i(a_i = 1, \mathbf{a}'_{-i}) - u_i(a_i = 0, \mathbf{a}'_{-i}) = p_i - v_i \left(\mathcal{I}_i(a_i = 1, \mathbf{a}'_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}'_{-i})\right) \\ &\overset{(a)}{\geq} p_i - v_i \left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right) = u_i(a_i = 1, \mathbf{a}_{-i}) - u_i(a_i = 0, \mathbf{a}_{-i}), \end{aligned}$$

where inequality (a) follows from part 2 of Lemma 1. Now consider the mapping $F : \{0,1\}^n \to \{0,1\}^n$ where $F_i(\mathbf{a}) = \text{argmax}_{a \in \{0,1\}} u_i(a, \mathbf{a}_{-i})$. Using supermodularity of the game, this mapping is order preserving and therefore Tarski's theorem establishes that its fixed points form a complete lattice and therefore is non-empty and has greatest and least elements. Finally, note that each fixed point of the mapping F is a user equilibrium and vice versa. Therefore, the set of fixed points of the mapping F is exactly the set of user equilibria denoted by $\mathcal{A}(\mathbf{p})$. ■

Proof of Theorem 1

We prove that the following action profile and price vector constitute an equilibrium:

$$\mathbf{a}^{\text{E}} = \text{argmax}_{\mathbf{a} \in \{0,1\}^n} \sum_{i \in \mathcal{V}} (1 - v_i)\mathcal{I}_i(\mathbf{a}) + v_i \mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0),$$

and $p_i^{\text{E}} = v_i \left(\mathcal{I}_i \left(a_i = 1, \mathbf{a}_{-i}^{\text{E}}\right) - \mathcal{I}_i \left(a_i = 0, \mathbf{a}_{-i}^{\text{E}}\right)\right)$, if $a_i^{\text{E}} = 1$ and $p_i^{\text{E}} = 0$ if $a_i^{\text{E}} = 0$. First note that $\mathbf{a}^{\text{E}} \in \mathcal{A}(\mathbf{p}^{\text{E}})$. This is because the payoff of user i when $a_i^{\text{E}} = 1$ is $p_i^{\text{E}} - v_i \mathcal{I}_i(\mathbf{a}^{\text{E}}) = -v_i \mathcal{I}_i(\mathbf{a}_{-i}^{\text{E}}, a_i = 0)$. If user i deviates and chooses not to share, her payoff would remain unchanged. However, when $a_i^{\text{E}} = 0$, her payoff is $-v_i \mathcal{I}_i(\mathbf{a}_{-i}^{\text{E}}, a_i = 0)$, and deviation to sharing would lead to the lower payoff of $-v_i \mathcal{I}_i(\mathbf{a}_{-i}^{\text{E}}, a_i = 1)$. Therefore, faced with the price vector offer of $\mathbf{p}^{\text{E}}$, the users do not have a profitable deviation from $\mathbf{a}^{\text{E}}$.

We next show that for any $\mathbf{p}$ and $\mathbf{a} \in \mathcal{A}(\mathbf{p})$, we have $U(\mathbf{a}^{\text{E}}, \mathbf{p}^{\text{E}}) \geq U(\mathbf{a}, \mathbf{p})$. Since $\mathbf{a}$ is a user equilibrium for the price vector $\mathbf{p}$, i.e., $\mathbf{a} \in \mathcal{A}(\mathbf{p})$, for all i such that $a_i = 1$, we must have $p_i \geq v_i \left(\mathcal{I}_i \left(a_i = 1, \mathbf{a}_{-i}\right) - \mathcal{I}_i \left(a_i = 0, \mathbf{a}_{-i}\right)\right)$. This is because if $p_i < v_i \left(\mathcal{I}_i \left(a_i = 1, \mathbf{a}_{-i}\right) - \mathcal{I}_i \left(a_i = 0, \mathbf{a}_{-i}\right)\right)$, then user i would have a profitable deviation to not share her data. Thus,

$$\begin{aligned} U(\mathbf{a}, \mathbf{p}) &= \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathbf{a}) - \sum_{i \in \mathcal{V}:\, a_i = 1} p_i \leq \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathbf{a}) - \sum_{i \in \mathcal{V}:\, a_i = 1} v_i \left(\mathcal{I}_i \left(a_i = 1, \mathbf{a}_{-i}\right) - \mathcal{I}_i \left(a_i = 0, \mathbf{a}_{-i}\right)\right) \\ &= \sum_{i \in \mathcal{V}} (1 - v_i)\mathcal{I}_i(\mathbf{a}) + v_i \mathcal{I}_i \left(\mathbf{a}_{-i}, a_i = 0\right) \leq U(\mathbf{a}^{\text{E}}, \mathbf{p}^{\text{E}}). \blacksquare \end{aligned}$$

Proof of Theorem 2

We use the following lemmas in this proof.

Lemma A-1 (Horn and Johnson [1987] Section 0.7)**.** • *The inverse of a matrix in terms of its blocks is*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1}. \end{pmatrix}$$

- *Sherman-Morrison-Woodbury formula for the inverse of rank one perturbation of matrix: Suppose $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u, v \in \mathbb{R}^n$ are column vectors. Then $A + uv^T$ is invertible if and only if $1 + v^T A^{-1}u \neq 0$. If $A + uv^T$ is invertible, then its inverse is*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}.$$

Lemma A-2 (Feller [2008]Chapter 5, Theorem 5)**.** *Suppose $(X_1, \dots, X_n)$ has a normal distribution with covariance matrix Σ. The conditional distribution of X_1 given $X_2, \dots, X_n$ is normal with covariance matrix $\Sigma_{11} - \mathbf{d}^T D^{-1}\mathbf{d}$, where D is the matrix obtained from Σ by removing the first row and the first column and $\mathbf{d} = (\Sigma_{12}, \dots, \Sigma_{1n})^T$.*

We now proceed with the proof of theorem. We first prove the existence of $\mathbf{p^a}$. Let $p_i^{\mathbf{a}} = v_i\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right)$. For any price vector $\mathbf{p}$ such that $\mathbf{a} \in \mathcal{A}(\mathbf{p})$ we have

$$u_i(a_i = 1, \mathbf{a}_{-i}) = p_i - v_i\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) \geq u_i(a_i = 0, \mathbf{a}_{-i}) = -v_i\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}), \quad \text{for all } i \text{ s.t. } a_i = 1.$$

Rearranging this inequality leads to $p_i \geq v_i\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right) = p_i^{\mathbf{a}}$. We next find the price vector $\mathbf{p^a}$ in terms of the matrix Σ. Let $S \subseteq \{1, \dots, n\}$ be the set of users who have shared their data. Leaked information about any user i is only a function of the correlation among users in S and the correlation between user i and the users in S. The relevant covariance matrix for finding leaked information about user i is given by the rows and columns of the matrix Σ corresponding to users in $S \cup \{i\}$. Therefore, without loss of generality, we suppose that $i = 1$ and all users have shared their data and work with the entire matrix Σ. We find the equilibrium price for user 1 (the price offered to other users can be obtained similarly). With $a_1 = \dots, a_n = 1$, $(X_1, S_1, \dots, S_n)$ is normally distributed with covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_1^2 & \Sigma_{12} & \dots & \Sigma_{1n} \\ \sigma_1^2 & 1+\sigma_1^2 & \Sigma_{12} & \dots & \Sigma_{1n} \\ \Sigma_{12} & \Sigma_{12} & 1+\sigma_2^2 & \dots & \Sigma_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{1n} & \Sigma_{1n} & \Sigma_{2n} & \dots & 1+\sigma_n^2 \end{pmatrix}.$$

Therefore, using Lemma A-2, the conditional distribution of X_1 given $s_1, \dots, s_n$ is normal with variance $\sigma_1^2 - (\sigma_1^2, \Sigma_{12}, \dots, \Sigma_{1n})(I + \Sigma)^{-1}(\sigma_1^2, \Sigma_{12}, \dots, \Sigma_{1n})^T$. The best estimator of X_1 given $s_1, \dots, s_n$

is its mean which leads to the following leaked information

$$\mathcal{I}_1(a_1 = 1, \mathbf{a}_{-1}) = (\sigma_1^2, \Sigma_{12}, \dots, \Sigma_{1n})(I + \Sigma)^{-1}(\sigma_1^2, \Sigma_{12}, \dots, \Sigma_{1n})^T. \tag{A-5}$$

If user 1 deviates to $a_1 = 0$, then $(X_1, S_2, \dots, S_n)$ has a normal distribution with covariance

$$\begin{pmatrix} \sigma_1^2 & \Sigma_{12} & \dots & \Sigma_{1n} \\ \Sigma_{12} & \ddots & & \vdots \\ \vdots & & I + D & \\ \Sigma_{1n} & \dots & & \ddots \end{pmatrix},$$

where D is obtained from Σ by removing the first row and column. Therefore, using Lemma A-2, the conditional distribution of X_1 given $s_2, \dots, s_n$ is normal with variance $\sigma_1^2 - (\Sigma_{12}, \dots, \Sigma_{1n})(I + D)^{-1}(\Sigma_{12}, \dots, \Sigma_{1n})^T$ and leaked information of user 1 is

$$\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1}) = (\Sigma_{12}, \dots, \Sigma_{1n})(I + D)^{-1}(\Sigma_{12}, \dots, \Sigma_{1n})^T. \tag{A-6}$$

Using A-5 and A-6, the price offered to user 1 must satisfy

$$\frac{p_1^{\mathbf{a}}}{v_1} = (\sigma_1^2, \mathbf{d}^T)^T \begin{pmatrix} \sigma_1^2 + 1 & \mathbf{d}^T \\ \mathbf{d} & (I + D) \end{pmatrix}^{-1} (\sigma_1^2, \mathbf{d}^T) - \mathbf{d}^T (I + D)^{-1} \mathbf{d},$$

where $\mathbf{d} = (\Sigma_{12}, \dots, \Sigma_{1n})$. We next simplify the right-hand side of the above equation. Using part 1 of Lemma A-1,

$$(\sigma_1^2, \mathbf{d}^T)^T \begin{pmatrix} \sigma_1^2 + 1 & \mathbf{d}^T \\ \mathbf{d} & I + D \end{pmatrix}^{-1} (\sigma_1^2, \mathbf{d}^T) - \mathbf{d}^T (I + D)^{-1} \mathbf{d} = (\sigma_1^2, \mathbf{d}^T)^T M (\sigma_1^2, \mathbf{d}^T) - \mathbf{d}^T (I + D)^{-1} \mathbf{d},$$

for

$$M = \begin{pmatrix} \left((\sigma_1^2 + 1) - \mathbf{d}^T (I + D)^{-1} \mathbf{d}\right)^{-1} & -\frac{1}{\sigma_1^2 + 1} \mathbf{d}^T \left((I + D) - \frac{1}{1 + \sigma_1^2} \mathbf{d}\mathbf{d}^T\right)^{-1} \\ -(I + D)^{-1} \mathbf{d} \left((\sigma_1^2 + 1) - \mathbf{d}^T (I + D)^{-1} \mathbf{d}\right)^{-1} & \left((I + D) - \frac{1}{1 + \sigma_1^2} \mathbf{d}\mathbf{d}^T\right)^{-1} \end{pmatrix}.$$

Using part 2 of Lemma A-1, we can further simplify this equation as follows:

$$
\begin{aligned}
&\sigma_1^2\left((\sigma_1^2+1)-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\right)^{-1}\sigma_1^2-\sigma_1^2\frac{1}{\sigma_1^2+1}\mathbf{d}^T\left((I+D)-\frac{\mathbf{d}\mathbf{d}^T}{\sigma_1^2+1}\right)^{-1}\mathbf{d}\\
&-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\left((\sigma_1^2+1)-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\right)^{-1}\sigma_1^2+\mathbf{d}^T\left((I+D)-\frac{1}{\sigma_1^2+1}\mathbf{d}\mathbf{d}^T\right)^{-1}\mathbf{d}\\
&-\mathbf{d}^T(I+D)^{-1}\mathbf{d}=\sigma_1^2\left((\sigma_1^2+1)-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\right)^{-1}\sigma_1^2\\
&-\left(\sigma_1^2\frac{1}{\sigma_1^1+1}\mathbf{d}^T(I+D)^{-1}\mathbf{d}+\sigma_1^2\frac{1}{\sigma_1^1+1}\frac{\mathbf{d}^T(I+D)^{-1}\frac{\mathbf{d}}{\sigma_1^2+1}\mathbf{d}^T(I+D)^{-1}\mathbf{d}}{1-\mathbf{d}^T(I+D)^{-1}\frac{\mathbf{d}}{\sigma_1^2+1}}\right)\\
&-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\left((\sigma_1^2+1)-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\right)^{-1}\sigma_1^2\\
&+\left(\mathbf{d}^T(I+D)^{-1}\mathbf{d}+\frac{\mathbf{d}^T(I+D)^{-1}\frac{\mathbf{d}}{\sigma_1^2+1}\mathbf{d}^T(I+D)^{-1}\mathbf{d}}{1-\mathbf{d}^T(I+D)^{-1}\frac{\mathbf{d}}{\sigma_1^2+1}}\right)-\mathbf{d}^T(I+D)^{-1}\mathbf{d}\\
&=\frac{\sigma_1^4}{(\sigma_1^2+1)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}-\frac{\sigma_1^2\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}{\sigma_1^2+1}-\frac{\sigma_1^2}{\sigma_1^2+1}\frac{\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})^2}{(\sigma_1^2+1)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}\\
&-\frac{\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})\sigma_1^2}{(\sigma_1^2+1)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}+\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})+\frac{\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})^2}{(\sigma_1^2+1)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})\\
&=\frac{\left(\sigma_1^2-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})\right)^2}{(\sigma_1^2+1)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})},
\end{aligned}
$$

where we used $\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})=\mathbf{d}^T(I+D)^{-1}\mathbf{d}$. This also implies

$$\mathcal{I}_1(a_1=1,\mathbf{a}_{-1})=\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})+\frac{\left(\sigma_1^2-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})\right)^2}{\left(\sigma_1^2+1\right)-\mathcal{I}_1(a_1=0,\mathbf{a}_{-1})}. \blacksquare$$

Proof of Corollary 1

Using Theorem 2, we have $p_i^{(a_i=1,\mathbf{a}_{-i})}=v_i\frac{(\sigma_i^2-\mathcal{I}_i(a_i=0,\mathbf{a}_{-i}))^2}{(\sigma_i^1+1)-\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})}$, which is increasing in σ_i^2 and decreasing in $\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$. Again, using Theorem 2, we have $\mathcal{I}_i(a_i=1,\mathbf{a}_{-i})=\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})+\frac{(\sigma_i^2-\mathcal{I}_i(a_i=0,\mathbf{a}_{-i}))^2}{(\sigma_i^1+1)-\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})}$, which is increasing in both $\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$ and σ_i^2. $\blacksquare$

Proof of Proposition 2

We present the proof in two steps. In the first step we show that $p_i^{\mathbf{a}}$ is decreasing in $|\text{Cov}(X_i,X_j\mid\mathbf{a})|$, denoted by $\tilde{\Sigma}_{ij}$ for any j and in the second step we prove that $p_i^{\mathbf{a}}$ is decreasing in $|\tilde{\Sigma}_{jk}|$ for any j and k.

Step 1: Without loss of generality suppose $i=1$, $j=2$ and the rest of the users share their data.

Note that conditional on $\mathbf{a}_{-\{1,2\}}$, (X_1, S_1, S_2) has a normal distribution. Let

$$\begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\sigma}_1^2 & \tilde{\Sigma}_{12} \\ \tilde{\sigma}_1^2 & \tilde{\sigma}_1^2+1 & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{12} & \tilde{\Sigma}_{12} & \tilde{\sigma}_2^2+1 \end{pmatrix}$$

denote its covariance matrix where $\mathbb{E}[X_1^2 \mid \mathbf{a}_{-\{1,2\}}] = \tilde{\sigma}_1^2$, $\mathbb{E}[X_1X_2 \mid \mathbf{a}_{-\{1,2\}}] = \tilde{\Sigma}_{12}$, and $\mathbb{E}[X_2^2 \mid \mathbf{a}_{-\{1,2\}}] = \tilde{\sigma}_2^2$. Using Theorem 2, the price offered to user 1 is

$$p_1 = \begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\Sigma}_{12} \end{pmatrix} \begin{pmatrix} \tilde{\sigma}_1^2+1 & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{12} & \tilde{\sigma}_2^2+1 \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\sigma}_1^2 \\ \tilde{\rho}_{12} \end{pmatrix} - \left(\frac{\tilde{\Sigma}_{12}^2}{\tilde{\sigma}_2^2+1} \right) = \frac{\left(\tilde{\Sigma}_{12}^2 - \tilde{\sigma}_1^2(1+\tilde{\sigma}_2^2)\right)^2}{(1+\tilde{\sigma}_2^2)\left((1+\tilde{\sigma}_2^2)(1+\tilde{\sigma}_1^2) - \tilde{\Sigma}_{12}^2\right)}.$$

Taking derivative with respect to $\tilde{\Sigma}_{12}^2$ establishes that $p_1^{\mathbf{a}}$ is decreasing in $\tilde{\Sigma}_{12}^2$ and therefore decreasing in $|\tilde{\Sigma}_{12}|$.

Step 2: Without loss of generality suppose $i = 1$, $j = 2$, $k = 3$ and the rest of the users share their data. Again, conditional on $\mathbf{a}_{-\{1,2,3\}}$, (X_1, S_2, S_3) has a normal distribution with covariance matrix

$$\begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{\Sigma}_{12} & \tilde{\Sigma}_{13} \\ \tilde{\Sigma}_{12} & \tilde{\sigma}_2^2+1 & \tilde{\Sigma}_{23} \\ \tilde{\Sigma}_{13} & \tilde{\Sigma}_{23} & \tilde{\sigma}_3^2+1 \end{pmatrix},$$

where $\mathbb{E}[X_1^2 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\sigma}_1^2$, $\mathbb{E}[X_2^2 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\sigma}_2^2$, $\mathbb{E}[X_3^2 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\sigma}_3^2$, $\mathbb{E}[X_1X_2 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\Sigma}_{12}$, $\mathbb{E}[X_1X_3 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\Sigma}_{13}$, and $\mathbb{E}[X_2X_3 \mid \mathbf{a}_{-\{1,2,3\}}] = \tilde{\Sigma}_{23}$. Using Theorem 2 and Corollary 1, the price offered to user 1 is increasing in $\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1})$. Therefore, in order to complete the proof, we need to prove that $\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1})$ is increasing in $|\Sigma_{23}|$. We have

$$\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1}) = \begin{pmatrix} \tilde{\Sigma}_{12} & \tilde{\Sigma}_{13} \end{pmatrix} \begin{pmatrix} \tilde{\sigma}_2^2+1 & \tilde{\Sigma}_{23} \\ \tilde{\Sigma}_{23} & \tilde{\sigma}_3^2+1 \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{13} \end{pmatrix}.$$

Taking the derivative of $\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1})$ with respect to $\tilde{\Sigma}_{23}$ yields

$$\frac{-2}{\left((1+\tilde{\sigma}_2^2)(1+\tilde{\sigma}_3^2) - \tilde{\Sigma}_{23}^2\right)} \left(\tilde{\Sigma}_{13}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{12}(1+\tilde{\sigma}_3^2)\right) \left(\tilde{\Sigma}_{12}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{13}(1+\tilde{\sigma}_2^2)\right).$$

We claim that this derivative is nonnegative for $\tilde{\Sigma}_{23} \geq 0$ and is nonpositive for $\tilde{\Sigma}_{23} \leq 0$, establishing that leaked information is increasing in $|\tilde{\Sigma}_{23}|$.

- Suppose $\tilde{\Sigma}_{23} \geq 0$. The derivative of $\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1})$ with respect to $\tilde{\Sigma}_{23}$ is nonnegative if and only if $\left(\tilde{\Sigma}_{13}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{12}(1+\tilde{\sigma}_3^2)\right) \left(\tilde{\Sigma}_{12}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{13}(1+\tilde{\sigma}_2^2)\right) \leq 0$. This inequality in turn is equivalent to $\begin{pmatrix}\tilde{\Sigma}_{12} & \tilde{\Sigma}_{13}\end{pmatrix} P \begin{pmatrix}\tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{13}\end{pmatrix} \leq 0$, where

$$P = \begin{pmatrix} -\tilde{\Sigma}_{23}(1+\tilde{\sigma}_3^2) & (1+\tilde{\sigma}_2^2)(1+\tilde{\sigma}_3^2) \\ \tilde{\Sigma}_{23}^2 & -\tilde{\Sigma}_{23}(1+\tilde{\sigma}_2^2) \end{pmatrix}. \tag{A-7}$$

The proof of this case completes by noting that the matrix P in this case is negative semi-definite with nonpositive eigenvalues 0 and $-\tilde{\Sigma}_{23}(2+\tilde{\sigma}_2^2+\tilde{\sigma}_3^2)$.

- Suppose $\tilde{\Sigma}_{23} \leq 0$. The derivative of $\mathcal{I}_1(a_1 = 0, \mathbf{a}_{-1})$ with respect to $\tilde{\Sigma}_{23}$ is nonpositive if and only if $\left(\tilde{\Sigma}_{13}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{12}(1+\tilde{\sigma}_3^2)\right)\left(\tilde{\Sigma}_{12}\tilde{\Sigma}_{23} - \tilde{\Sigma}_{13}(1+\tilde{\sigma}_2^2)\right) \geq 0$. This inequality in turn is equivalent to $\begin{pmatrix}\tilde{\Sigma}_{12} & \tilde{\Sigma}_{13}\end{pmatrix} P \begin{pmatrix}\tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{13}\end{pmatrix} \geq 0$. The proof is completed by noting that the matrix P (defined in A-7) in this case is positive semi-definite with nonnegative eigen values 0 and $-\tilde{\Sigma}_{23}(2+\tilde{\sigma}_2^2+\tilde{\sigma}_3^2)$. ■

Proof of Proposition 3

Let $i \in \mathcal{V}$ be such that $a_i' = a_i = 1$. Using Theorem 2, we have

$$p_i^{\mathbf{a}} = v_i\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right) \overset{(a)}{\geq} v_i\left(\mathcal{I}_i(a_i' = 1, \mathbf{a}_{-i}') - \mathcal{I}_i(a_i' = 0, \mathbf{a}_{-i}')\right) = p_i^{\mathbf{a}'},$$

where (a) follows from submodularity of leaked information, i.e., part 2 of Lemma 1. ■

Proof of Lemma 3

Suppose to obtain a contradiction that in equilibrium $a_i^{\mathrm{E}} = 0$ for some $i \in \mathcal{V}$ with $v_i \leq 1$. We prove that there exists a deviation which increases the platform's payoff. In particular, the platform can deviate and offer price $p_i = v_i\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}})\right)$ so that user i shares.

From Theorem 1, the equilibrium action profile $\mathbf{a}^{\mathrm{E}}$ must maximize $\sum_{i\in\mathcal{V}}(1-v_i)\mathcal{I}_i(\mathbf{a}) + v_i\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$. We show that $(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}})$ increases this objective, which yields a contradiction:

$$\begin{aligned}
&\left(\sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}})\right) - \left(\sum_{j\in\mathcal{V}:\ a_j^{\mathrm{E}}=1} p_j^{(a_i=1,\mathbf{a}_{-i}^{\mathrm{E}})} - p_j^{(a_i=0,\mathbf{a}_{-i}^{\mathrm{E}})}\right) \\
&+ \left((1-v_i)\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}}) + v_i\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}})\right) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}}) \\
&\overset{(a)}{\geq} -\left(\sum_{j\in\mathcal{V}:\ a_j^{\mathrm{E}}=1} p_j^{(a_i=1,\mathbf{a}_{-i}^{\mathrm{E}})} - p_j^{(a_i=0,\mathbf{a}_{-i}^{\mathrm{E}})}\right) + \left((1-v_i)\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}}) + v_i\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}})\right) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}}) \\
&\overset{(b)}{\geq} (1-v_i)\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}^{\mathrm{E}}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{E}})\right) \overset{(c)}{\geq} 0,
\end{aligned}$$

where (a) follows from monotonicity of leaked information (i.e., part 1 of Lemma 1), (b) follows from Proposition 3, and (c) follows from the fact that $v_i \leq 1$ and leaked information is monotone. This shows that for any i such that $v_i \leq 1$ we must have $a_i^{\mathrm{E}} = 1$. ■

Proof of Theorem 3

We use the following notation in this proof. For any action profile $\mathbf{a} \in \{0,1\}^n$ and any subset $T \subseteq \{1,\dots,n\}$, we let $\mathbf{a}_T$ denote a vector that include all the entries of a_i for which $i \in T$.
Part 1: For a given action profile $\mathbf{a}$, the social surplus can be written as

$$\text{Social surplus}(\mathbf{a}) = \sum_{i \in \mathcal{V}} (1-v_i)\mathcal{I}_i(\mathbf{a})$$
$$\stackrel{(a)}{=} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}}, \mathbf{a}_{\mathcal{V}^{(h)}} = 0) + \sum_{i \in \mathcal{V}^{(h)}} (1-v_i)\mathcal{I}_i(a_i, \mathbf{a}_{-i} = \mathbf{0}) \stackrel{(b)}{\leq} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathcal{V}^{(l)}),$$

where (a) follows from the fact that the data of high-value users are not correlated with the data of any other user, and (b) follows from the fact that for $i \in \mathcal{V}^{(l)}$, leaked information about user i (weakly) increases in the set of users who share (from part 1 of Lemma 1) and $1 - v_i \geq 0$. Conversely, for $i \in \mathcal{V}^{(h)}$ we have $1 - v_i < 0$. This implies $\mathbf{a}_i^{\mathrm{W}} = 1$ if and only if $i \in \mathcal{V}^{(l)}$.

The payoff of the platform for a given action profile $\mathbf{a}$ (and the corresponding equilibrium prices to sustain it) can be written as

$$U(\mathbf{a}, \mathbf{p}^{\mathbf{a}}) = \sum_{i \in \mathcal{V}} (1-v_i)\mathcal{I}_i(\mathbf{a}) + v_i \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$$
$$\stackrel{(a)}{=} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}}) + v_i \mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)} \setminus \{i\}}) + \sum_{i \in \mathcal{V}^{(h)}} (1-v_i)\mathcal{I}_i(a_i, \mathbf{a}_{-i} = \mathbf{0})$$
$$\stackrel{(b)}{\leq} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}}) + v_i \mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)} \setminus \{i\}}) \stackrel{(c)}{\leq} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) + v_i \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$$

where (a) follows from the fact that the data of high-value users are not correlated with the data of any other user, (b) follows from the fact that $1 - v_i < 0$ for $i \in \mathcal{V}^{(h)}$, and (c) follows from Lemma 3. Therefore, no high-value user shares in equilibrium and we have $\mathbf{a}^{\mathrm{E}} = \mathbf{a}^{\mathrm{W}}$.
Part 2: Let $i \in \mathcal{V}^{(l)}$ and $j \in \mathcal{V}^{(h)}$ be such that $\Sigma_{ij} > 0$. Therefore, there exists $\delta > 0$ such that $\mathcal{I}_j(\mathcal{V}^{(l)}) = \delta > 0$. We next show that for $v_j > 1 + \frac{\sum_{i \in \mathcal{V}^{(l)}} \sigma_i^2}{\delta}$ the surplus of the action profile $\mathbf{a}^{\mathrm{E}}$ is negative, establishing that it does not coincide with the first best. We have

$$\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) = \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) + \sum_{i \in \mathcal{V}^{(h)}} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) \stackrel{(a)}{\leq} \sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\sigma_i^2 + \sum_{i \in \mathcal{V}^{(h)}} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}})$$
$$\stackrel{(b)}{\leq} \left(\sum_{i \in \mathcal{V}^{(l)}} (1-v_i)\sigma_i^2 \right) + (1-v_j)\mathcal{I}_j(\mathcal{V}^{(l)}) \leq \left(\sum_{i \in \mathcal{V}^{(l)}} \sigma_i^2 \right) + (1-v_j)\mathcal{I}_j(\mathcal{V}^{(l)}) \stackrel{(c)}{<} 0$$

where in (a) for low-value users we have upper bounded leaked information with its maximum; in (b) we have removed all the negative terms in the second summation except for the one corresponding to j for which we have replaced the leaked information (of equilibrium action profile) by its minimum (using Lemma 3); and in (c) we have used $v_j > 1 + \frac{\sum_{i \in \mathcal{V}^{(l)}} \sigma_i^2}{\delta}$.

Part 3: Let $i, k \in \mathcal{V}^{(h)}$ be such that $\Sigma_{ik} > 0$. The first best involves all low-values are sharing their data and none of the high-value users doing so. We next show that if the value of privacy for high-value user i is small enough, then at least one high-value user shares in equilibrium. We show this by assuming the contrary and then reaching a contradiction. Suppose that none of high-value users share. We show that if user i shares, the platform's payoff increases. We let $\mathbf{a}'^n$ denote the sharing profile in which all users in $\mathcal{V}^{(l)} \cup \{i\}$ share their data and $\mathbf{a} \in \{0,1\}^n$ denote the sharing profile in which all users in $\mathcal{V}^{(l)}$ share their data. Using this notation, let us write

$$
\begin{aligned}
U(\mathbf{a}', \mathbf{p}^{\mathbf{a}'}) &= (1 - v_i)\mathcal{I}_i(\mathcal{V}^{(l)} \cup \{i\}) + v_i \mathcal{I}_i(\mathcal{V}^{(l)}) + \sum_{k \in \mathcal{V}^{(h)} \setminus \{i\}} \mathcal{I}_k(\mathcal{V}^{(l)} \cup \{i\}) \\
&+ \left(\sum_{j \in \mathcal{V}^{(l)}} (1 - v_j)\mathcal{I}_j(\mathcal{V}^{(l)} \cup \{i\}) + v_j \mathcal{I}_j(\mathcal{V}^{(l)} \cup \{i\} \setminus \{j\}) \right) \\
&\overset{(a)}{=} (1 - v_i)\mathcal{I}_i(\mathcal{V}^{(l)} \cup \{i\}) + v_i \mathcal{I}_i(\mathcal{V}^{(l)}) + \left(\sum_{k \in \mathcal{V}^{(h)} \setminus \{i\}} \mathcal{I}_k(\mathcal{V}^{(l)} \cup \{i\}) \right) + U(\mathbf{a}, \mathbf{p}^{\mathbf{a}}) \overset{(b)}{>} U(\mathbf{a}, \mathbf{p}^{\mathbf{a}}),
\end{aligned}
$$

where (a) follows from the fact that high- and low-value users are uncorrelated and (b) follows by letting

$$
v_i < \frac{\mathcal{I}_i(\mathcal{V}^{(l)} \cup \{i\}) + \sum_{k \in \mathcal{V}^{(h)} \setminus \{i\}} \mathcal{I}_k(\mathcal{V}^{(l)} \cup \{i\})}{\mathcal{I}_i(\mathcal{V}^{(l)} \cup \{i\}) - \mathcal{I}_i(\mathcal{V}^{(l)})} = \frac{\mathcal{I}_i(\{i\}) + \sum_{k \in \mathcal{V}^{(h)} \setminus \{i\}} \mathcal{I}_k(\{i\})}{\mathcal{I}_i(\{i\})}.
$$

Finally, note that using $\Sigma_{ik} > 0$, the right-hand side of the above inequality is strictly larger than 1. The proof is completed by letting $\bar{v}_i = \frac{\mathcal{I}_i(\{i\}) + \sum_{k \in \mathcal{V}^{(h)} \setminus \{i\}} \mathcal{I}_k(\{i\})}{\mathcal{I}_i(\{i\})}$. ■

Proof of Proposition 4

For an equilibrium action profile $\mathbf{a}^{\mathrm{E}}$, social surplus can be written as

$$
\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) = \sum_{i \in \mathcal{V}} (1 - v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) \overset{(a)}{\leq} \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}) + \sum_{i \in \mathcal{V}^{(h)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}^{(l)})
$$

where (a) follows from the fact that for $i \in \mathcal{V}^{(l)}$, leaked information about user i increases in the set of users who share (i.e., part 1 of Lemma 1) and $1 - v_i \geq 0$; and for $i \in \mathcal{V}^{(h)}$ we have $1 - v_i < 0$ and $\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) \geq \mathcal{I}_i(\mathcal{V}^{(l)})$ by using Lemma 3. ■

Proof of Proposition 5

Using Theorem 2, leaked information about a high-value user $i \in \mathcal{V}^{(h)}$ if low-value users share is given by

$$(\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})(2I + M)^{-1}(\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})^T,$$

where low-value users are denoted by $j_1, \ldots, j_k$ and the diagonal entries of M are zero and $M_{r,s}$ is the covariance between two low-value users r and s. We next prove that this leaked information is larger than or equal to $\sum_{l=1}^{k} \frac{\Sigma_{ij_l}^2}{k+1}$. We first show that $(2I + M)^{-1} - ((k+1)I)^{-1} \succeq 0$ (i.e., the matrix $(2I + M)^{-1} - ((k+1)I)^{-1}$ is positive semidefinite). Letting μ_i denote an eigen value of the matrix $(2I + M)^{-1} - ((k+1)I)^{-1}$, it suffices to show that $\mu_i \geq 0$. There exists an eigenvalue, λ_i, of the matrix $2I + M$ for which we have $\mu_i = \frac{1}{\lambda_i} - \frac{1}{k+1}$. We next show that all eigenvalues of the matrix $2I + M$ are (weakly) smaller than $k+1$, which establishes that $\mu_i \geq 0$. Using Gershgorin Circle Theorem, the matrix $(k+1)I - (2I + M)$ is positive semidefinite. This is because for any row of this matrix, the diagonal entry is $k - 1$ which is larger than the summation of the absolute values of the off-diagonal entries. Therefore, for any eigenvalue of the matrix $2I + M$ such as λ_i, we have $\lambda_i \leq k + 1$. We can write

$$\begin{aligned}&(\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})(2I + M)^{-1}(\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})^T \\ &\geq (\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})((k+1)I)^{-1}(\Sigma_{ij_1}, \ldots, \Sigma_{ij_k})^T = \sum_{l=1}^{k} \frac{\Sigma_{ij_l}^2}{k+1}.\end{aligned} \tag{A-8}$$

Using Proposition 4 and Corollary 2, equilibrium surplus is negative if $\sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\mathcal{I}_i(\mathcal{V}^{(l)}) > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V})$. From inequality (A-8) and $\mathcal{I}_i(\mathcal{V}) \leq 1$, this condition holds provided that

$$\sum_{i \in \mathcal{V}^{(h)}} (v_i - 1) \frac{\sum_{j \in \mathcal{V}^{(l)}} \Sigma_{ij}^2}{|\mathcal{V}^{(l)}| + 1} > \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i).$$

This completes the proof. ■

References

A. Acquisti and H. R. Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3): 367–381, 2005.

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

A. R. Admati and P. Pfleiderer. A monopolistic market for information. *Journal of Economic Theory*, 39(2):400–438, 1986.

A. Agarwal, M. Dahleh, and T. Sarkar. A marketplace for data: an algorithmic solution. *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

A. Agrawal, J. Gans, and A. Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.

J. J. Anton and D. A. Yao. The sale of ideas: Strategic disclosure, property rights, and contracting. *The Review of Economic Studies*, 69(3):513–531, 2002.

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

M. Babaioff, R. Kleinberg, and R. Paes Leme. Optimal mechanisms for selling information. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 92–109, 2012.

J. Begenau, M. Farboodi, and L. Veldkamp. Big data in finance and the growth of large firms. *Journal of Monetary Economics*, 97:71–87, 2018.

D. Bergemann and A. Bonatti. Selling cookies. *American Economic Journal: Microeconomics*, 7(3): 259–94, 2015.

D. Bergemann and A. Bonatti. Markets for information: An introduction. *Annual Review of Economics*, 11, 2019.

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

D. Bergemann, A. Bonatti, and T. Gan-a. Markets for information. *mimeo, MIT and Yale*, 2019.

M. Burkschat and N. Torrado. On the reversed hazard rate of sequential order statistics. *Statistics & Probability Letters*, 85:106–113, 2014.

Y. Chen and S. Zheng. Prior-free data acquisition for accurate statistical estimation. *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 659–677, 2019.

Y. Chen, N. Immorlica, B. Lucier, V. Syrgkanis, and J. Ziani. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 27–44, 2018.

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

E. H. Clarke. Multipart pricing of public goods. *Public choice*, 11(1):17–33, 1971.

P. Dasgupta and E. Maskin. The existence of equilibrium in discontinuous economic games, I: Theory. *Review of Economic Studies*, 53(1):1–26, 1986.

P. Eső and B. Szentes. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies*, 74(3):705–731, 2007.

J. A. Fairfield and C. Engel. Privacy as a public good. *Duke Law Journal*, 65(3):385–457, 2015.

W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

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

A. Goldfarb and C. Tucker. Shifts in privacy concerns. *American Economic Review*, 102(3):349–53, 2012.

T. Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, 41(4):617–631, 1973.

R. A. Horn and C. R. Johnson. Matrix analysis. 1985. *Cambridge, Cambridge*, 1987.

J. Horner and A. Skrzypacz. Selling information. *Journal of Political Economy*, 124(6):1515–1562, 2016.

C. I. Jones and C. Tonettil. Nonrivalry and the economics of data. In *2018 Meeting Papers*, number 477. Society for Economic Dynamics, 2018.

K. C. Laudon. Markets and privacy. *Communications of the ACM*, 39(9):92–104, 1996.

M. MacCarthy. New directions in privacy: Disclosure, unfairness and externalities. *I/S: A Journal of Law and Policy for the Information Society*, 6:425–512, 2011.

R. Montes, W. Sand-Zantman, and T. Valletti. The value of personal information in online markets with endogenous privacy. *Management Science*, 65(3):1342–1362, 2018.

F. Pasquale. *The black box society*. Harvard University Press, 2015.

E. A. Posner and E. G. Weyl. *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press, 2018.

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

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

C. R. Taylor. Consumer privacy and the market for customer information. *RAND Journal of Economics*, 35(4):631–650, 2004.

J. Tirole. Digital dystopia. 2019.

H. R. Varian. Economic aspects of personal privacy. In *Internet policy and economics*, pages 101–109, 2009.

L. Veldkamp. Data and the aggregate economy. 2019.

L. Veldkamp, M. Farboodi, R. Mihet, and T. Philippon. Big data and firm dynamics. 2019.

W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

S. D. Warren and L. D. Brandeis. The right to privacy. *Harvard law review*, 4(5):193–220, 1890.

A. F. Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

S. Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books, 2019.

Appendix B: Online Appendix for "Too Much Data: Prices and Inefficiencies in Data Markets"

B.1 Proof of Pure Strategy Existence and Inefficiency When $v_i \leq 1$ for all users

We show that with competition if $v_i < 1$, c_i is equal to constant c, and $1 > \sum_{j=1}^{n} |\Sigma_{ij}|$ for all $i \in \mathcal{V}$, then a pure strategy equilibrium exists. Since all v_i's are below one, in the Stackelberg equilibrium (in the second stage) all users share their data. We next show that the first stage game has a pure strategy equilibrium. In particular, we prove the game is a potential game. Suppose the set S of users join the first platform (and also share their data). The payoff of a user $i \in S$ is

$$u_i = p_i - v_i \mathcal{I}_i(S) + c = v_i \left(\mathcal{I}_i(S) - \mathcal{I}_i(S \setminus \{i\})\right) - v_i \mathcal{I}_i(S) + c = -v_i \mathcal{I}_i(S \setminus \{i\}) + c.$$

Therefore, if in equilibrium set S of users join the first platform and the rest (i.e., the set S^c) join the second platform, then we must have

$$\mathcal{I}_i(S^c) \geq \mathcal{I}_i(S \setminus \{i\}), \text{ for all } i \in S, \text{ and } \mathcal{I}_i(S) \geq \mathcal{I}_i(S^c \setminus \{i\}), \text{ for all } i \in S^c.$$

This means that leaked information of a user on platform 1 (if she does not share) must be smaller than her leaked information on platform 2 (if she joins that platform). That is, the information leakage due to externality on the platform that a user joins is smaller than the other platform.

Before introducing the potential function, we introduce some additional notation. Consider a graph with the set of nodes $\{1, \ldots, n\}$ and edge weights Σ_{ij} (with self-loops of weights σ_i^2). A *walk* is a finite sequence of edges which joins a sequence of vertices. A *closed walk* is a walk in which the first and the last vertices are the same. We define the *weight of a walk* as the product of the weights of its edges. For any $i \in \mathcal{V}$ and a given set $S \subseteq \{1, \ldots, n\}$, S^c, and $n_1, n_2 \in \mathbb{N}$ we let $\mathcal{W}(i; n_1, n_2; S)$ be the sum over the weight of all closed walks that start from node i, visits n_1 nodes in the set $S \setminus \{i\}$ and n_2 nodes in the set $S^c \setminus \{i\}$.

We first provide a reformulation of leaked information in terms of these objects.

Claim 1: For any node $i \in S$ we have

$$\mathcal{I}_i(S \setminus \{i\}) = \sum_{t=1}^{\infty} (-1)^{t+1} \mathcal{W}(i; t, 0; S).$$

Suppose, without loss of generality, $S = \{1, \ldots, k\}$. We also let Σ denote the rows and columns of the covariance matrix corresponding to the users in the set S. Leaked information of user $i \notin S$ as characterized in Theorem 2 is $(\Sigma_{i1}, \ldots, \Sigma_{ik})(I + \Sigma)^{-1}(\Sigma_{i1}, \ldots, \Sigma_{ik})^T$. Given $1 > \sum_{j=1}^{n} |\Sigma_{ij}|$, we have $1 > \sum_{j \in S} |\Sigma_{ij}|$ for all $i \in S$ and Gershgorin Circle Theorem shows that $\rho(\Sigma) < 1$. Therefore, the Taylor expansion of $I+\Sigma$ is convergent and we can rewrite the above expression as

$$\begin{aligned}
&(\Sigma_{i1},\dots,\Sigma_{ik})\left(I-\Sigma+\Sigma^2-\Sigma^3+\dots\right)(\Sigma_{i1},\dots,\Sigma_{ik})^T\\
&=\sum_{j=1}^{k}\Sigma_{ij}^2-\sum_{j_1,j_2\in S}\Sigma_{ij_1}\Sigma_{j_1j_2}\Sigma_{j_2i}+\sum_{j_1,j_2,j_3\in S}\Sigma_{ij_1}\Sigma_{j_1j_2}\Sigma_{j_2j_3}\Sigma_{j_3i}-\dots\\
&=\sum_{t=1}^{\infty}(-1)^{t+1}\mathcal{W}(i;t,0;S).
\end{aligned}$$

This completes the proof of the claim.
Claim 2: For any set S and S^c, we have

$$\sum_{n_1,n_2\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S)=\sum_{n_1,n_2\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S^c).$$

This holds true because by definition of the weight of walks we have $\mathcal{W}(i;n_1,n_2;S)=\mathcal{W}(i;n_2,n_1;S^c)$.
Claim 3: The following is a potential function of the first stage game:

$$\Psi(S)=\sum_{i\in S}\sum_{n_1,n_2:\,n_1\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S)+\sum_{i\in S^s}\sum_{n_1,n_2:\,n_2\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S^c).$$

Moreover, the following set S is the equilibrium of the first stage game:

$$S\in\text{argmax}_T\Psi(T).$$

First note that there are finitely many sets T (i.e., 2^n of them) and therefore the maximizer of $\Psi(\cdot)$ exists. We next prove that the maximizer of $\Psi(\cdot)$ corresponds to a pure strategy equilibrium of the first stage game, i.e.,

$$\mathcal{I}_i(S^c)\geq\mathcal{I}_i(S\setminus\{i\}),\ \text{ for all } i\in S,\ \text{ and } \mathcal{I}_i(S)\geq\mathcal{I}_i(S^c\setminus\{i\}),\ \text{ for all } i\in S^c.$$

Let $i\in S$. Since S is the maximizer of $\Psi(\cdot)$, by moving i from S to S^c we obtain $\Psi(S\setminus\{i\})\leq\Psi(S)$. We further have

$$\begin{aligned}
\Psi(S\setminus\{i\})&\overset{(a)}{=}\Psi(S)-\sum_{n_1,n_2:\,n_1\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S)+\sum_{n_1,n_2:\,n_2\geq 1}(-1)^{n_1+n_2+1}\mathcal{W}(i;n_1,n_2;S^c)\\
&\overset{(b)}{=}\Psi(S)-\sum_{n_1\geq 1}(-1)^{n_1+1}\mathcal{W}(i;n_1,0;S)+\sum_{n_2\geq 1}(-1)^{n_2+1}\mathcal{W}(i;0,n_2;S^c)\\
&\overset{(c)}{=}\Psi(S)-\mathcal{I}_i(S^c)+\mathcal{I}_i(S\setminus\{i\})\leq\Psi(S),
\end{aligned}$$

where (a) follows from the fact that by moving i from S to S^c the terms $\mathcal{W}(j;n_1,n_2;S)$ for all other $j\neq i$ do not change, (b) follows from Claim 2, and (c) follows from Claim 1. Therefore, we have $\mathcal{I}_i(S^c)\geq\mathcal{I}_i(S\setminus\{i\})$. This completes the proof of Claim 2.

Using an identical argument shows that for all $i \in S^c$, we have $\mathcal{I}_i(S) \geq \mathcal{I}_i(S^c \setminus \{i\})$. This completes the proof of existence of a pure strategy equilibrium. ■

We next show that when all values are below 1, the equilibrium is always inefficient. Suppose S and S^c are the equilibrium joining (and sharing) sets on platforms 1 and 2, respectively. The social surplus becomes $\sum_{i \in S}(1 - v_i)\mathcal{I}_i(S) + \sum_{i \in S^c}(1 - v_i)\mathcal{I}_i(S^c) \leq \sum_{i \in \mathcal{V}}(1 - v_i)\mathcal{I}_i(\mathcal{V})$, where the inequality follows from the monotonicity of leaked information (Lemma 1). Note that the right-hand side is the equilibrium surplus when we have a single platform, completing the proof. ■

B.2 Remaining Proofs

Proof of Lemma 4

The payoff of user i from joining platform $k \in \{1, 2\}$ can be lower bounded by

$$c_i(J_k) + p_i^{J_k,\mathbf{E}} - v_i\mathcal{I}_i(\mathbf{a}^{J_k,\mathrm{E}}) \overset{(a)}{=} c_i(J_k) - v_i\mathcal{I}_i(\mathbf{a}_{-i}^{J_k,\mathrm{E}}, a_i = 0) \overset{(b)}{\geq} c_i(\{i\}) - v_i\sigma_i^2,$$

which is positive given Assumption 1. Equality (a) follows from the characterization of equilibrium prices given in Theorem 2, and inequality (b) follows from the monotonicity of the joining value function $c_i(\cdot)$ and the fact that maximum leaked information is σ_i^2. ■

Proof of Theorem 4

After the joining decisions are made, the equilibrium characterization of the sharing profile and the price vector on each platform is identical to that of Theorems 1 and 2. ■

Proof of Theorem 5

The theorem follows from the fact by using a uniform mixed strategy, the two platforms generate the same ex ante expected payoff for users, hence no user has an incentive to deviate from uniform randomization. ■

Proof of Theorem 6

Part 1: When high-value users are not correlated with other users, the first best is to have all users join one of the platforms and to have all low-value users share. This is because: (i) the high-value users do not share because the data of high-value users do not increase leaked information of low-value users and generate negative surplus, (ii) the low value users share because it generates a positive surplus itself and also increases leaked information of other low-value users (and does not leak information about high-value users), and (iii) all users join the same platform because it increases the joining value as well as leaked information of low-value users. We next show that this is an equilibrium if and only if

$$c_i(\mathcal{V}) - c_i(\{i\}) \geq \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) \text{ for all } i \in \mathcal{V}^{(l)}. \tag{B-1}$$

If inequality (B-1) holds, then the first best is an equilibrium: First, note that high-value users do not have an incentive to deviate to the other platform because they would gain $c_i(\{i\})$ instead of $c_i(\mathcal{V})$. Second, consider $i \in \mathcal{V}^{(l)}$ and suppose she deviates and joins the other platform. Her payoff changes from $c_i(\mathcal{V}) - v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$ to $c_i(\{i\})$. This is not a profitable deviation when inequality (B-1) holds.

If inequality (B-1) does not hold, then the first best is not an equilibrium: Given inequality (B-1) does not hold, there exists $i \in \mathcal{V}^{(l)}$ such that $c_i(\mathcal{V}) - c_i(\{i\}) < \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$. We let $v_i = \frac{c_i(\mathcal{V})-c_i(\{i\})}{\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})} + \epsilon$ for $\epsilon = \frac{1}{2}\left(1 - \frac{c_i(\mathcal{V})-c_i(\{i\})}{\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})}\right)$, and show that all users joining the first platform and low-value users sharing on platform 1 is not an equilibrium. With this choice for ϵ, v_i is less than 1 and strictly larger than $\frac{c_i(\mathcal{V})-c_i(\{i\})}{\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})}$. In particular, we show that user i has a profitable deviation by joining platform 2. Using Theorem 4, the payoff of user i on platform 1 is $-v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\mathcal{V})$. If she deviates and joins platform 2, her payoff becomes (again using Theorem 4) $c_i(\{i\})$. This is a profitable deviation because

$$\begin{aligned} &- v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\mathcal{V}) \overset{(a)}{=} -\left(\frac{c_i(\mathcal{V}) - c_i(\{i\})}{\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})} + \epsilon\right)\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\mathcal{V}) \\ &= -\left(c_i(\mathcal{V}) - c_i(\{i\})\right) - \epsilon\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\mathcal{V}) = -\epsilon\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\{i\}) < c_i(\{i\}). \end{aligned}$$

where in (a) we substituted the value of v_i.

Part 2: Let $\mathcal{V}_1^{(l)} = \{i \in \mathcal{V}^{(l)} : \forall j \in \mathcal{V}^{(h)}, \Sigma_{ij} = 0\}$ be the set of low-value users whose correlation with all high-value users are zero and $\mathcal{V}_2^{(l)} = \mathcal{V}^{(l)} \setminus \mathcal{V}_1^{(l)}$ be the set of low-value users with at least one non-zero correlation coefficient to a high-value user. By assumption, there exists at least one non-zero correlation coefficient between high and low-value users, showing $\mathcal{V}_2^{(l)} \neq \emptyset$. We next show that there exist $\bar{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(h)}|}$ and $\underline{\mathbf{v}} \in \mathbb{R}^{|\mathcal{V}^{(l)}|}$ such that for $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$ and $\mathbf{v}^{(l)} \geq \underline{\mathbf{v}}$, the first best is to have all users joining the same platform and low-value users in $\mathcal{V}_1^{(l)}$ share their data. Suppose the contrary, i.e., a subset of users J_1 join the first platform and users in $S_1 \subseteq J_1$ share their data, and the rest of the users $J_2 = \mathcal{V} \setminus J_1$ join the second platform and users in $S_2 \subseteq J_2$ share on platform 2. For large enough $\bar{\mathbf{v}}$ on each platform, leaked information of high-value users must be zero. Therefore, for the set of users sharing on platforms 1 and 2, we have $S_1, S_2 \subseteq \mathcal{V}^{(l)}$. For sufficiently large $\underline{\mathbf{v}}$, the surplus is upper bounded as follows:

$$\begin{aligned} &\sum_{i \in J_1}(1 - v_i)\mathcal{I}_i(S_1) + c_i(J_1) + \sum_{i \in J_2}(1 - v_i)\mathcal{I}_i(S_2) + c_i(J_2) \\ &\overset{(a)}{\leq} \sum_{i \in \mathcal{V}^{(l)}}(1 - v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) + \sum_{i \in J_1} c_i(J_1) + \sum_{i \in J_2} c_i(J_2) \overset{(b)}{\leq} \sum_{i \in \mathcal{V}^{(l)}}(1 - v_i)\mathcal{I}_i(\mathcal{V}_1^{(l)}) + \sum_{i \in \mathcal{V}} c_i(\mathcal{V}), \end{aligned}$$

where (a) follows from the fact that only low-value users have a non-zero leaked information (because the value of high value users is large) and their leaked information is upper bounded by their leaked information when all low-value users share, and (b) follows from the fact if all low-value users are very close to 1, then the inequality in (b) holds because of the monotonicity of c_i, and therefore there exists $\underline{\mathbf{v}}$ close enough to $\mathbf{1}$ for which (b) holds.

Now note that this cannot be an equilibrium because in equilibrium all low-value users share, but in the first best the low-value user who has a non-zero correlation coefficient with one of the high-value users should not be sharing (i.e., those users in $\mathcal{V}_2^{(l)} \neq \emptyset$ should not be sharing).

Part 3: Let users $i, k \in \mathcal{V}^{(h)}$ be such that $\Sigma_{ik} > 0$. The first best is to have all users joining the same platform, which we assume without loss of generality is platform 1, all low-value users share and none of the high-value users share. This is because high-value users have no externality on low-value users and therefore they should never share in the first best. They join the same platform as low-value users to enjoy the higher joining value. Also, all low-value users join the same platform and since they have a zero correlation coefficient with high-value users, they should all share in the first best. An identical argument to that of Part 3 of Theorem 3 shows that for sufficiently small value v_i, user i shares on platform 1, establishing that the first best cannot be equilibrium. ■

Proof of Proposition 6

The proof is very similar to the proof of Proposition 4. In particular, similar to Lemma 3, for any joining profile in the Stackelberg equilibrium users with values below one share. Similar to Proposition 4, we can lower bound leaked information about high-value users such as i with their leaked information when only low-value users shares; and upper bound the leaked information of low-value users such as i with their leaked information when all users share. We next provide the lower bound on the leaked information of a high-value user i:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}}\left[\mathcal{I}_i(\mathbf{a}^{J,\mathrm{E}})\right] &\overset{(a)}{=} \frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathbf{a}^{J_1,\mathrm{E}})\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathbf{a}^{J_2,\mathrm{E}})\right] \overset{(b)}{\geq} \frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathcal{V}^{(l)} \cap J_1)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathcal{V}^{(l)} \cap J_2)\right] \\ &= \mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\frac{1}{2}\mathcal{I}_i(\mathcal{V}^{(l)} \cap J_1) + \frac{1}{2}\mathcal{I}_i\left(\mathcal{V}^{(l)} \setminus \left(\mathcal{V}^{(l)} \cap J_1\right)\right)\right] \overset{(c)}{\geq} \mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\frac{1}{2}\mathcal{I}_i(\mathcal{V}^{(l)})\right] = \frac{1}{2}\mathcal{I}_i(\mathcal{V}^{(l)}).\end{aligned}$$

where (a) follows from the fact that we have uniform mixed joining strategy, (b) follows from the fact that on each platform all low-value users share their information, and (c) follows from the fact that leaked information is submodular. ■

Proof of Theorem 7

Consider platform 1 and suppose the set J_1 of users have joined this platform. Platform 1 can only affect the sharing decision of users that have joined it. An argument identical to that of Theorem

1 shows that

$$\mathbf{a}^{J_1,\mathrm{E}} = \mathrm{argmax}_{\mathbf{a}\in\{0,1\}^{|J_1|}} \sum_{i\in J_1} (1-v_i)\mathcal{I}_i(\mathbf{a}) + v_i \mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$$
$$p_i^{J_1,\mathrm{E}} = v_i(\mathcal{I}_i(\mathbf{a}^{J_1,\mathrm{E}}) - \mathcal{I}_i(a_i=0,\mathbf{a}_{-i}^{J_1,\mathrm{E}})). \quad \text{(B-2)}$$

Therefore, the equilibrium payoff of users $i \in J_1^{\mathrm{E}}$ becomes

$$c_i(J_1^{\mathrm{E}}) + p_i^{J_1,\mathrm{E}} - v_i\mathcal{I}_i(\mathbf{a}^{J_1^{\mathrm{E}},\mathrm{E}}) - \alpha v_i \mathcal{I}_i(\mathbf{a}^{J_2^{\mathrm{E}},\mathrm{E}}) = c_i(J_1^{\mathrm{E}}) - v_i\mathcal{I}_i(a_i=0,\mathbf{a}_{-i}^{J_1^{\mathrm{E}},\mathrm{E}}) - \alpha v_i\mathcal{I}_i(\mathbf{a}^{J_2^{\mathrm{E}},\mathrm{E}}),$$

where the equality follows from B-2. ■

Proof of Theorem 8

Part 1: First note for small enough α the first-best is again to have all users joining the same platform and only low-value users sharing their data. To see this, suppose the set J_1 and J_2 of users have joined platforms 1 and 2, respectively. Also, suppose the set of users S_1 and S_2 have shared on platforms 1 and 2, respectively. We can bound the social surplus as follows:

$$\sum_{i\in J_1}(1-v_i)\mathcal{I}_i(S_1) + \alpha(1-v_i)\mathcal{I}_i(S_2) + c_i(J_1) + \sum_{i\in J_2}(1-v_i)\mathcal{I}_i(S_2) + \alpha(1-v_i)\mathcal{I}_i(S_1) + c_i(J_2)$$
$$\overset{(a)}{\leq} \sum_{i\in J_1\cap\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(S_1\cap\mathcal{V}^{(l)}) + \alpha(1-v_i)\mathcal{I}_i(S_2\cap\mathcal{V}^{(l)}) + \sum_{i\in J_1} c_i(J_1)$$
$$+ \sum_{i\in J_2\cap\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(S_2\cap\mathcal{V}^{(l)}) + \alpha(1-v_i)\mathcal{I}_i(S_1\cap\mathcal{V}^{(l)}) + \sum_{i\in J_2} c_i(J_2) \overset{(b)}{\leq} \sum_{i\in\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) + \sum_{i\in\mathcal{V}} c_i(\mathcal{V}),$$

where (a) follows from the fact that if a high-value user shares, it contributes negatively to the social surplus as it only leaks information about other high-value users and (b) holds for small enough α because in the limit for $\alpha = 0$, this inequality becomes

$$\sum_{i\in J_1\cap\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(S_1\cap\mathcal{V}^{(l)}) + \sum_{i\in J_1} c_i(J_1) + \sum_{i\in J_2\cap\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(S_2\cap\mathcal{V}^{(l)}) + \sum_{i\in J_2} c_i(J_2)$$
$$\leq \sum_{i\in\mathcal{V}^{(l)}} (1-v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) + \sum_{i\in\mathcal{V}} c_i(\mathcal{V}),$$

which holds because both the leaked information and the joining value $c_i(\cdot)$ are monotonically increasing. We next show that this is an equilibrium if and only if we have

$$c_i(\mathcal{V}) - c_i(\{i\}) \geq \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})(1-\alpha), \quad \text{for all } i \in \mathcal{V}^{(l)}. \quad \text{(B-3)}$$

If inequality (B-3) holds, then the first best is an equilibrium: First, note that high-value users do not have an incentive to deviate to the other platform because their payoff becomes $c_i(\{i\})$

instead of $c_i(\mathcal{V})$. Second, consider $i \in \mathcal{V}^{(l)}$ and suppose she deviates and joins the other platform. Her payoff changes from $c_i(\mathcal{V}) - v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$ to $c_i(\{i\}) - \alpha v_i \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$. This is not a profitable deviation if inequality (B-3) holds.

If inequality (B-3) does not hold, then the first best is not an equilibrium: Since inequality (B-3) does not hold, there exists $i \in \mathcal{V}^{(l)}$ such that $c_i(\mathcal{V}) - c_i(\{i\}) < (1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$. We let $v_i = \frac{c_i(\mathcal{V})-c_i(\{i\})}{(1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})} + \epsilon$ for $\epsilon = \frac{1}{2}\left(1 - \frac{c_i(\mathcal{V})-c_i(\{i\})}{(1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})}\right)$, and show that all users joining the first platform and low-value users sharing on platform 1 is not an equilibrium. With this choice for ϵ, v_i is less than 1 and strictly larger than $\frac{c_i(\mathcal{V})-c_i(\{i\})}{(1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})}$. In particular, we show that user i has a profitable deviation by joining platform 2. Using Theorem 7, the payoff of user i on platform 1 is $-v_i\mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\}) + c_i(\mathcal{V})$. If she deviates and joins platform 2, her payoff becomes (again using Theorem 7) $c_i(\{i\}) - \alpha v_i \mathcal{I}_i(\mathcal{V}^{(l)} \setminus \{i\})$. This is a profitable deviation because $v_i > \frac{c_i(\mathcal{V})-c_i(\{i\})}{(1-\alpha)\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\})}$. Therefore, user i has a profitable deviation.

Part 2: In any equilibrium (either pure strategy or mixed strategy), low-value users share on the platform that they join. Now let $\Sigma_{ij} > 0$ for $i \in \mathcal{V}^{(l)}$ and $j \in \mathcal{V}^{(h)}$. In any equilibrium user i shares her data and this leads to a non-zero information leakage for user j larger that $\delta > 0$. Here, unlike the proof of Theorem 6, even if users i and j join different platforms, the data of user i still leaks some information about user j's type. Therefore, for sufficiently large v_j, the surplus becomes negative, showing that any equilibrium is inefficient.

Part 3: First note that since low and high-value users are uncorrelated, in the first-best solution high-value users do not share. We let J_1 and J_2 denote the joining profiles of the first best. We also let S_1 and S_2 denote the sharing profile of the first best. We let $\Sigma_{jk} > 0$ for $j, k \in \mathcal{V}^{(h)}$ and show that for sufficiently small v_j, user j shares her data in equilibrium. Specifically, we establish that if the first best is an equilibrium, then the first platform can deviate to get the sharing profile $S \cup \{j\}$ which is a contradiction. The payoff of platform 1 with sharing profile S is $\sum_{i\in J_1} \mathcal{I}_i(S_1) + \sum_{i \in J_2} \alpha \mathcal{I}_i(S_1) - \sum_{i\in S_1} p_i$, where using Theorem 7 we have $p_i = v_i(\mathcal{I}_i(S) - \mathcal{I}_i(S \setminus \{i\}))$. The payoff of platform 1 with sharing profile $S \cup \{j\}$ improves because we have

$$
\begin{aligned}
&\sum_{i\in J_1} \mathcal{I}_i(S_1 \cup \{j\}) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1 \cup \{j\}) - \sum_{i\in S_1\cup\{j\}} p_i' \\
&= \mathcal{I}_j(\{j\}) + \sum_{i\in J_1\setminus\{j\}} \mathcal{I}_i(S_1 \cup \{j\}) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1 \cup \{j\}) - \sum_{i\in S_1\cup\{j\}} p_i' \\
&\overset{(a)}{\geq} \mathcal{I}_j(\{j\}) + \alpha\mathcal{I}_k(\{j\}) + \sum_{i\in J_1\setminus\{j\}} \mathcal{I}_i(S_1) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1) - \sum_{i\in S_1\cup\{j\}} p_i' \\
&\overset{(b)}{\geq} \mathcal{I}_j(\{j\}) + \alpha\mathcal{I}_k(\{j\}) - v_j\mathcal{I}_j(\{j\}) + \sum_{i\in J_1\setminus\{j\}} \mathcal{I}_i(S_1) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1) - \sum_{i\in S_1} p_i \\
&\overset{(c)}{=} \mathcal{I}_j(\{j\}) + \alpha\mathcal{I}_k(\{j\}) - v_j\mathcal{I}_j(\{j\}) + \sum_{i\in J_1\setminus\{j\}} \mathcal{I}_i(S_1) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1) - \sum_{i\in S_1} p_i \\
&\overset{(d)}{>} \sum_{i\in J_1\setminus\{j\}} \mathcal{I}_i(S_1) + \sum_{i\in J_2} \alpha\mathcal{I}_i(S_1) - \sum_{i\in S_1} p_i,
\end{aligned}
$$

where in (a) we used the fact that S_1 only includes low-value users, and the addition of user j's data has the following contributions: (i) it changes the leaked information of other high-value users (besides j and k) from zero to some positive number, (ii) it changes the leaked information of user j from zero to $\mathcal{I}_j(\{j\})$, and (iii) it changes the leaked information of user k from zero to either $\mathcal{I}_k(\{j\})$ or $\alpha\mathcal{I}_k(\{j\})$ depending on whether $k \in J_1$ or not; in (b) we used the fact that the data of high-value user j has no effect on low-value users' leaked information and hence their offered price, and the price offered to user j to share her data is $v_j\mathcal{I}_j(\{j\})$; (c) follows from the fact that S_1 only included low-value users which leak no information about user j; and (d) follows by letting $v_j < 1 + \frac{\alpha\mathcal{I}_k(\{j\})}{\mathcal{I}_j(\{j\})}$. The proof is completed by letting $\bar{v}_j = 1 + \frac{\alpha\mathcal{I}_k(\{j\})}{\mathcal{I}_j(\{j\})}$. ■

Proof of Proposition 7

We let J_1 and J_2 be the set of users joining platforms 1 and 2, respectively and S_1 and S_2 be the set of users sharing on platforms 1 and 2 respectively. The sum of the platform's utilities can be written as

$$\sum_{i\in J_1}\mathcal{I}_i(S_1)+\alpha\mathcal{I}_i(S_2)-p_i+\sum_{i\in J_2}\mathcal{I}_i(S_2)+\alpha\mathcal{I}_i(S_1)-p_i. \tag{B-4}$$

Also, the summation of users' utilities (excluding the joining value) can be written as

$$\sum_{i\in J_1}p_i-v_i\mathcal{I}_i(S_1)-v_i\alpha\mathcal{I}_i(S_2)+\sum_{i\in J_2}p_i-v_i\mathcal{I}_i(S_2)-v_i\alpha\mathcal{I}_i(S_1)-p_i. \tag{B-5}$$

Taking summation of equations (B-4) and (B-5), the data social surplus becomes

$$\sum_{i\in J_1}(1-v_i)(\mathcal{I}_i(S_1)+\alpha\mathcal{I}_i(S_2))+\sum_{i\in J_2}(1-v_i)(\mathcal{I}_i(S_2)+\alpha\mathcal{I}_i(S_1)).$$

The rest of the proof is identical to that of Proposition 6 with the only difference that for a low-value user $i\in\mathcal{V}$ that joins platform $k\in\{1,2\}$, we use the following bound:

$$\mathcal{I}_i(\mathbf{a}^{J_k,\mathrm{E}})+\alpha\mathcal{I}_i(\mathbf{a}^{J_{k'},\mathrm{E}})\le(1+\alpha)\mathcal{I}_i(\mathcal{V}),$$

and for high-value user i we use the bound

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta}}\left[\mathcal{I}_i(\mathbf{a}^{J_k,\mathrm{E}})+\alpha\mathcal{I}_i(\mathbf{a}^{J_{k'},\mathrm{E}})\right] &\overset{(a)}{\ge}\left(\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathbf{a}^{J_1,\mathrm{E}})\right]+\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathbf{a}^{J_2,\mathrm{E}})\right]\right)(1+\alpha)\\
&\overset{(b)}{\ge}\left(\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathcal{V}^{(l)}\cap J_1)\right]+\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\mathcal{I}_i(\mathcal{V}^{(l)}\cap J_2)\right]\right)(1+\alpha)\\
&=\left(\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\frac{1}{2}\mathcal{I}_i(\mathcal{V}^{(l)}\cap J_1)+\frac{1}{2}\mathcal{I}_i\left(\mathcal{V}^{(l)}\setminus\left(\mathcal{V}^{(l)}\cap J_1\right)\right)\right]\right)(1+\alpha)\\
&\overset{(c)}{\ge}\mathbb{E}_{\boldsymbol{\beta}_{-i}}\left[\frac{1}{2}\mathcal{I}_i(\mathcal{V}^{(l)})\right](1+\alpha)=\frac{1+\alpha}{2}\mathcal{I}_i(\mathcal{V}^{(l)}).
\end{aligned}$$

where (a) follows from the fact that we have uniform mixed joining strategy, (b) follows from the fact that on each platform all low-value users share their information, and (c) follows from the fact that leaked information is submodular. ■

Proof of Theorem 9

For any price vector $\mathbf{p}^1$ and $\mathbf{p}^2$, the second-stage game is a finite game and therefore has a mixed strategy equilibrium. If there are multiple equilibria, we select the one with the highest sum of platform's utilities. We next show that the first-stage game has a mixed strategy equilibirium by using Dasgupta-Maskin theorem stated below.

Theorem B-1 (Dasgupta and Maskin [1986]). *Consider a game with n players where the action space of user i is denoted by a bounded set S_i and her payoff is denoted by u_i. If*

1. *for any i, u_i is continuous except on a subset of $S^*(i)$, where*

$$S^*(i) = \{\mathbf{s} \in \mathbf{S} : \ \exists j \neq i \text{ such that } s_j = f_{ij}^d(s_i)\},$$

   for bijective and continuous functions $f_{ij}^d : S_i \to S_j$ for $d = 1, \ldots, D(i)$.

2. *$\sum_{i=1}^N u_i(\mathbf{s})$ is upper semicontinuous, i.e., for any sequence $\mathbf{s}^k \to \mathbf{s}$, we have*

$$\sum_{i=1}^N u_i(\mathbf{s}) \geq \limsup_{k\to\infty} \sum_{i=1}^N u_i(\mathbf{s}^k).$$

3. *for any i, $u_i(s_i, \mathbf{s}_{-i})$ is weakly lower semicontinuous, i.e., there exists $\lambda \in [0, 1]$ such that*

$$\lambda \liminf_{s_i' \uparrow s_i} u_i(s_i', \mathbf{s}_{-i}) + (1 - \lambda) \liminf_{s_i' \downarrow s_i} u_i(s_i', \mathbf{s}_{-i}) \geq u_i(s_i, \mathbf{s}_{-i}).$$

Then a mixed strategy equilibrium exists.

We next show that the conditions of this theorem are satisfied, establishing a mixed strategy equilibrium exists. First, note that the price each platform offers to any user cannot exceed the highest overall leaked information, i.e., $\sum_{i\in\mathcal{V}} \mathcal{I}_i(\mathcal{V})$. Therefore, without loss of generality, we assume the action space of both platforms is $[0, \sum_{i\in\mathcal{V}} \mathcal{I}_i(\mathcal{V})]^n$.

For two vector of prices $\mathbf{p}^1$ and $\mathbf{p}^2$ and user $i \in \mathcal{V}$ we define functions $f_{12} : \mathbf{p}^1 \to \mathbf{p}^2$ such that

$$[f_{12}(\mathbf{p}^1)]_i = p_i^1 - v_i\mathcal{I}_i(S_1) + c_i(J_1) + v_i\mathcal{I}_i(S_2) - c_i(J_2), \forall S_1, S_2, J_1, J_2 \subseteq \mathcal{V}.$$

Note that there are finitely many such functions and in particular at most $n^{2^n\times 2^n\times 2^n}$ of them (this is because there are n components and for each of them J_1 has 2^n possibilities, $J_2 = \mathcal{V}\backslash J_1$, and each of S_1 and S_2 have 2^n possibilities). Also, note that the functions f_{12} are all linear and hence bijective and continuous. By changing the prices $\mathbf{p}^1$ and $\mathbf{p}^2$, as long as user equilibria of the second-stage game are the same, the payoff functions remain continuous. It becomes discontinuous when a user $i \in \mathcal{V}$ who is sharing on platform 1 changes her decision and starts sharing on platform 2. For this to happen we must have

$$p_i^1 - v_i \mathcal{I}_i(S_1) + c_i(J_1) = p_i^2 - v_i \mathcal{I}_i(S_2) + c_i(J_2),$$

where S_1 is the set of users who are sharing on platform 1, S_2 is the set of users who are sharing on platform 2, and J_1 and J_2 are the sets of users who are joining platforms 1 and 2, respectively. Therefore, for any discontinuity point of $U^1(\mathbf{p}^1, \mathbf{p}^2)$, there exists f_{12} such that $\mathbf{p}^2 = f_{12}(\mathbf{p}^1)$. This establishes that the first condition of Theorem B-1 holds.

The second condition of Theorem B-1 holds because as long as user equilibria of the second-stage game remains the same, payoff functions are continuous in the first stage prices. When user equilibria changes, we select the one with the highest sum of the platforms' utilities. This implies that the sum of platforms' utilities is an upper semicontinuous function in prices.

The third condition of Theorem B-1 holds because by changing $\mathbf{p}^1$, as long as the equilibrium of the second-stage game has not changed, the payoff of platform 1 is continuous. At the point that the equilibrium changes, we have multiplicity of equilibria and we have chosen the one that gives maximum payoff of platforms. Therefore, we have $\liminf_{\mathbf{p}'^1 \downarrow \mathbf{p}^1} U^1(\mathbf{p}'^1, \mathbf{p}^2) = U^1(\mathbf{p}^1, \mathbf{p}^2)$, which by definition is weakly lower semicontinuous with the choice of $\lambda = 0$. ■

Proof of Theorem 10

Part 1: In this case, there is no externality among users, and both the first best and the equilibrium involve all users joining platform 1 (or platform 2) and all low-value users sharing their data. In particular, we show that the following prices with a user equilibrium in which all users join platform 1 and all low-value users share on platform 1 is an equilibrium. For all $i \in \mathcal{V}^{(l)}$ we let

$$p_i^{1,\mathrm{E}} = \begin{cases} v_i \mathcal{I}_i(\{i\}), & c_i(\mathcal{V}) - c_i(\{i\}) \geq (1 - v_i)\mathcal{I}_i(\{i\}), \\ \mathcal{I}_i(\{i\}) - \left(c_i(\mathcal{V}) - c_i(\{i\})\right), & c_i(\mathcal{V}) - c_i(\{i\}) < (1 - v_i)\mathcal{I}_i(\{i\}), \end{cases}$$

and

$$p_i^{2,\mathrm{E}} = \begin{cases} v_i \mathcal{I}_i(\{i\}), & c_i(\mathcal{V}) - c_i(\{i\}) \geq (1 - v_i)\mathcal{I}_i(\{i\}), \\ \mathcal{I}_i(\{i\}), & c_i(\mathcal{V}) - c_i(\{i\}) < (1 - v_i)\mathcal{I}_i(\{i\}). \end{cases}$$

This is a user equilibrium because the payoff of a user on platform 1 is $c_i(\mathcal{V})$ that is larger than her payoff on platform 2 which is $c_i(\{i\})$. We next show that platform 1 does not have a profitable deviation. For any user i for which $c_i(\mathcal{V}) - c_i(\{i\}) \geq (1 - v_i)\mathcal{I}_i(\{i\})$ platform 1 cannot increase its payoff by reducing its price offer because the user would then stop sharing her data. For any user with $c_i(\mathcal{V}) - c_i(\{i\}) < (1 - v_i)\mathcal{I}_i(\{i\})$, a lower price would make the user join platform 2. This establishes that the first platform does not have a profitable deviation. We next show that platform 2 does not have a profitable deviation. The maximum price that platform 2 can offer to user i

without making negative profits is $\mathcal{I}_i(\{i\})$ (this is because there exists no externality). Such a price offer is not sufficient to attract users from platform 1. In particular, if $c_i(\mathcal{V})-c_i(\{i\}) < (1-v_i)\mathcal{I}_i(\{i\})$ then the price $\mathcal{I}_i(\{i\})$ offered to user i by the second platform would make her indifferent between the two platforms and if $c_i(\mathcal{V}) - c_i(\{i\}) \geq (1 - v_i)\mathcal{I}_i(\{i\})$ then the price $\mathcal{I}_i(\{i\})$ offered to user i by the second platform would give her a lower payoff. Therefore, the second platform does not have a profitable deviation. This completes the proof of the first part.

Part 2: Note that the first best is to have all users join the same platform and low-value users share on it, which we assume is platform 1. Since there is no externality from high-value users, without loss of generality, we show the proof when they are removed from the market as they will join (and not share) the platform that has a higher joining value for them.

Part 2-1: We show that if $\delta \geq \underline{\delta}$ for

$$\underline{\delta} = \max_{i\in\mathcal{V}} \left(\sum_{j\in\mathcal{V}} \mathcal{I}_j(\mathcal{V}) \right) + v_i(\mathcal{I}_i(\mathcal{V} \setminus \{i\}) - \mathcal{I}_i(\{i\})), \tag{B-6}$$

the first best is an equilibrium, supported by the following prices:

$$\begin{aligned} p_i^{1,\mathrm{E}} &= v_i(\mathcal{I}_i(\mathcal{V}) - \mathcal{I}_i(\mathcal{V} \setminus \{i\})), \quad i \in \mathcal{V} \\ p_i^{2,\mathrm{E}} &= c_i(\mathcal{V}) - c_i(\{i\}) - v_i\mathcal{I}_i(\mathcal{V} \setminus \{i\}) + v_i\mathcal{I}_i(\{i\}), \quad i \in \mathcal{V}. \end{aligned} \tag{B-7}$$

Note that all (low-value) users joining and sharing on platform 1 is a user equilibrium. This is because the payoff of each user i is she deviates and shares on the other platform is equal to her current payoff. Platform 1 does not have a profitable deviation. This is because, using Theorem 2, platform 1 is paying the minimum prices to get the data of all users. We next show that platform 2 does not have a profitable deviation, establishing the prices specified above and all users sharing on platform 1 is an equilibrium. The highest price that platform 2 can offer to get one of users, e.g., user i, share on it is bounded by the total information leakage $\sum_{j\in\mathcal{V}} \mathcal{I}_j(\mathcal{V})$. We show that given the condition on δ all users joining and sharing on platform 1 is always a user equilibrium which gives the second platform zero payoff. Consider user $i \in \mathcal{V}$, given all other users are sharing on platform 1 it is best response for this user to share on platform 1 because

$$\begin{aligned} p_i^{1,\mathrm{E}} - v_i\mathcal{I}_i(\mathcal{V}) + c_i(\mathcal{V}) &\overset{(a)}{=} -v_i\mathcal{I}_i(\mathcal{V} \setminus \{i\}) + c_i(\mathcal{V}) \overset{(b)}{\geq} -v_i\mathcal{I}_i(\mathcal{V} \setminus \{i\}) + \delta + c_i(\{i\}) \\ &\overset{(c)}{\geq} \sum_{j\in\mathcal{V}} \mathcal{I}_j(\mathcal{V}) - v_i\mathcal{I}_i(\{i\}) + c_i(\{i\}) \geq \tilde{p}_i^2 - v_i\mathcal{I}_i(\{i\}) + c_i(\{i\}). \end{aligned}$$

where (a) follows from plugging in the prices given in (B-7), (b) follows from the definition of δ, and (c) follows from (B-6).

Part 2-2: We show that given $\Delta \leq \bar{\Delta}$ and $v_i \leq \bar{v}$, where

$$\bar{\Delta} = \max\Bigg\{ \max_{i,S:\ i\notin S}(1-v_i)\mathcal{I}_i(\mathcal{V}) + v_i(2\mathcal{I}_i(\{i\}) - \mathcal{I}_i(S^c)), \tag{B-8}$$

$$\max_{S\subseteq\mathcal{V}} \frac{1}{2|S|}\sum_{i\in S}\mathcal{I}_i(\mathcal{V}) - (1-v_i)\mathcal{I}_i(S) - v_i\mathcal{I}_i(S^c\cup\{i\})\Bigg\}. \tag{B-9}$$

and

$$\bar{v} = \min\left\{\frac{1}{2}, \min_{i,S:\ i\in S, \mathcal{I}_i(\mathcal{V})-\mathcal{I}_i(S)\neq 0} \frac{\mathcal{I}_i(\mathcal{V}) - \mathcal{I}_i(S)}{\mathcal{I}_i(\mathcal{V}) - \mathcal{I}_i(S) + \mathcal{I}_i(S^c\cup\{i\}) - \mathcal{I}_i(\{i\})}\right\}. \tag{B-10}$$

the first best is an equilibrium.

Before we proceed with the proof of this case, note that both $\bar{\Delta}$ and $\bar{v}$ are non-zero. The latter is by hypothesis. Consider the former, (B-8). The first term on the right-hand side of this expression is non-zero if $v_i \leq \frac{1}{2}$. The second term on the right-hand side of this expression is non-zero, because $v_i \leq 1$ and leaked information is monotonically increasing in the set of users who share. If $\mathcal{I}_i(\mathcal{V}) = \mathcal{I}_i(S)$ for some S, then user i is uncorrelated with users and the complement of S, S^c. If $\mathcal{I}_i(\mathcal{V}) = \mathcal{I}_i(S^c)$ for some S, then user i is uncorrelated with users in S. Therefore, for the right-hand side of the above expression to be zero all users must be uncorrelated with all other users. But this contradicts the assumption that at least the data of two low-value users are correlated.

We next show that the following prices form an equilibrium:

$$p_i^{1,\mathrm{E}} = \mathcal{I}_i(\mathcal{V}) - (c_i(\mathcal{V}) - c_i(\{i\})), \quad i \in \mathcal{V}, \tag{B-11}$$

$$p_i^{2,\mathrm{E}} = (1-v_i)\mathcal{I}_i(\mathcal{V}) + \mathcal{I}_i(\{i\}), \quad i \in \mathcal{V}. \tag{B-12}$$

Note that with these prices all users sharing on platform 1 is a user equilibrium. This is because if a user deviates and shares on the second platform, she receives the same payoff. We next show that the second platform does not have a profitable deviation. Suppose the second platform deviates to get users in a set S share on it by offering prices $\tilde{p}_i^2$. We show that the payoff of platform 2 becomes strictly negative. Consider one of the user equilibria after this deviation and suppose that the set $J_1 \subseteq S^c$ of users join platform 1 and a subset of J_1, S_1, shares on platform 1. Users in S must prefer to share on platform 2, which leads to

$$\begin{aligned}\tilde{p}_i^2 - v_i\mathcal{I}_i(S) + c_i(S) &\geq p_1^i - v_i\mathcal{I}_i(S_1\cup\{i\}) + c_i(J_1\cup\{i\}) \\ &\geq \mathcal{I}_i(\mathcal{V}) - (c_i(\mathcal{V}) - c_i(\{i\})) - v_i\mathcal{I}_i(S^c\cup\{i\}) + c_i(\{i\}).\end{aligned}$$

Hence

$$\begin{aligned}\tilde{p}_i^2 &\geq v_i\mathcal{I}_i(S) - c_i(S) + \mathcal{I}_i(\mathcal{V}) - (c_i(\mathcal{V}) - c_i(\{i\})) - v_i\mathcal{I}_i(S^c\cup\{i\}) + c_i(\{i\}) \\ &\overset{(a)}{\geq} -2\Delta + \mathcal{I}_i(\mathcal{V}) + v_i\mathcal{I}_i(S) - v_i\mathcal{I}_i(S^c\cup\{i\}).\end{aligned} \tag{B-13}$$

Therefore, the payoff of platform 2 becomes

$$\sum_{i\in S}\mathcal{I}_i(S)-\tilde{p}_i^2 \overset{(a)}{\leq} \sum_{i\in S}\mathcal{I}_i(S)+2\Delta-\mathcal{I}_i(\mathcal{V})-v_i\mathcal{I}_i(S)+v_i\mathcal{I}_i(S^c\cup\{i\})$$
$$=2\Delta|S|+\sum_{i\in S}(1-v_i)\mathcal{I}_i(S)+v_i\mathcal{I}_i(S^c\cup\{i\})-\mathcal{I}_i(\mathcal{V}) \overset{(b)}{<} 0,$$

where (a) follows by using (B-13) and (b) follows from the choice of $\bar{\Delta}$ given in (B-8).

We next show that platform 1 does not have a profitable deviation. Suppose that platform 1 deviates to get users in the set S to share on it with prices $\tilde{p}_i^1$. We first claim that if $\Delta \leq \bar{\Delta}$ (where $\bar{\Delta}$ is given in (B-8)), in one of the user equilibria all users prefer to share on platform 2. In particular, for a user $i \in S^c$, her payoff if she shares on platform 2 is higher than her payoff if she joins platform 2 and does not share because

$$p_i^2-v_i\mathcal{I}_i(S^c)+c_i(S^c) \overset{(a)}{=} (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c)+c_i(S^c) \overset{(b)}{\geq} -v_i\mathcal{I}_i(S^c\setminus\{i\})+c_i(S^c),$$

where (a) follows by using the prices given in (B-11) and (b) follows the submodularity of leaked information and in particular $v_i\mathcal{I}_i(\{i\}) \geq v_i(\mathcal{I}_i(S^c)-\mathcal{I}_i(S^c\setminus\{i\}))$. Also, for a user $i \in S^c$, her payoff if she shares on platform 2 is higher than her payoff if she joins platform 1 because (she does not share on platform 1 as the price offered to her is 0)

$$p_i^2-v_i\mathcal{I}_i(S^c)+c_i(S^c) \overset{(a)}{=} (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c)+c_i(S^c)$$
$$\overset{(b)}{\geq} (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c)+c_i(S\cup\{i\})-\Delta$$
$$\overset{(c)}{\geq} -v_i\mathcal{I}_i(\{i\})+c_i(S\cup\{i\}) \geq -v_i\mathcal{I}_i(S\cup\{i\})+c_i(S\cup\{i\})$$

where (a) follows by using the prices given in (B-11), (b) follows from the definition of Δ, (c) follows from the choice of $\bar{\Delta}$ given in (B-8). To have users in the set S share on platform 1 the new prices must satisfy

$$\tilde{p}_i^1-v_i\mathcal{I}_i(S)+c_i(S) \geq p_i^2-v_i\mathcal{I}_i(S^c\cup\{i\})+c_i(S^c\cup\{i\})$$
$$\overset{(a)}{=} (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c\cup\{i\})+c_i(S^c\cup\{i\}),$$

where (a) follows from substituting for prices from (B-11). Therefore,

$$\tilde{p}_i^1 \geq (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c\cup\{i\})+v_i\mathcal{I}_i(S)+c_i(S^c\cup\{i\})-c_i(S)$$
$$\geq (1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(\{i\})-v_i\mathcal{I}_i(S^c\cup\{i\})+v_i\mathcal{I}_i(S)-\Delta. \tag{B-14}$$

We next show that this user equilibrium generates no higher payoff for platform 1 (compared to its equilibrium payoff of $\Delta|\mathcal{V}|$). In particular, the payoff of platform 1 can be written as

$$\sum_{i\in S}\mathcal{I}_i(S)-\tilde{p}_i^1 \overset{(a)}{\leq} \sum_{i\in S}(1-v_i)\mathcal{I}_i(S)-(1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(S^c\cup\{i\})-v_i\mathcal{I}_i(\{i\})+\Delta \overset{(b)}{\leq} \Delta|S|\leq\Delta|\mathcal{V}|$$

where in (a) we used the inequality (B-14), and (b) follows from the choice of $\bar{v}$ given in (B-10). In particular, using $v_i\leq\bar{v}$, for any i and S such that $\mathcal{I}_i(\mathcal{V})-\mathcal{I}_i(S)\neq 0$, we have

$$(1-v_i)\mathcal{I}_i(S)-(1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(S^c\cup\{i\})-v_i\mathcal{I}_i(\{i\})\leq 0.$$

For any S and $i\in S$ for which $\mathcal{I}_i(\mathcal{V})-\mathcal{I}_i(S)=0$, user i's data must be independent of the data of users in S^c. Therefore, we have $\mathcal{I}_i(S^c\cup\{i\})=\mathcal{I}_i(\{i\})$. This leads to

$$(1-v_i)\mathcal{I}_i(S)-(1-v_i)\mathcal{I}_i(\mathcal{V})+v_i\mathcal{I}_i(S^c\cup\{i\})-v_i\mathcal{I}_i(\{i\})= 0.$$

This completes the proof of this case.

Part 2-3: We show that if $\Delta\leq\tilde{\Delta}$ where

$$\tilde{\Delta}=\min\left\{\min_{i\in\mathcal{V}}\frac{1}{2}(1-v_i)\mathcal{I}_i(\{i\}),\max_{i\in\mathcal{V}}\frac{\sum_{i\neq j}\mathcal{I}_i(\{i,j\})-\mathcal{I}_i(\{i\})}{2(3|\mathcal{V}|-2)}\right\}, \tag{B-15}$$

then there exist $\tilde{\mathbf{v}}$ such that for $\mathbf{v}^{(l)}\geq\tilde{\mathbf{v}}$ the equilibrium is inefficient.

Before, we proceed with the proof, note that the second argument of maximum is non-zero since there exits at least two low-value users whose data are correlated.

We show that there exists no prices for both platforms to sustain all (low-value) users share on platform 1 as an equilibrium. We suppose the contrary and then reach a contradiction. In particular, we let $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$ be the equilibrium prices offered by platform 1 and 2. Since all users sharing on platform 1 is a user equilibrium we must have $x_i-v_i\mathcal{I}_i(\mathcal{V})+c_i(\mathcal{V})\geq y_i-v_i\mathcal{I}_i(\{i\})+c_i(\{i\})$. Also, note that we must have $x_i-v_i\mathcal{I}_i(\mathcal{V})\leq c_i(\mathcal{V})-c_i(\{i\})+y_i-v_i\mathcal{I}_i(\{i\})$. This is because, otherwise platform 1 can deviate by decreasing its prices and increase its payoff. Therefore, we have

$$x_i-v_i\mathcal{I}_i(\mathcal{V})\in\left[-(c_i(\mathcal{V})-c_i(\{i\}))+y_i-\mathcal{I}_i(\{i\}),(c_i(\mathcal{V})-c_i(\{i\}))+y_i-\mathcal{I}_i(\{i\})\right],\quad i\in\mathcal{V}. \tag{B-16}$$

Moreover, we also have

$$x_i-v_i\mathcal{I}_i(\mathcal{V})+c_i(\mathcal{V})\geq(1-v_i)\mathcal{I}_i(\{i\})+c_i(\{i\}),\quad i\in\mathcal{V}. \tag{B-17}$$

This is because if this inequality does not hold for some $i\in\mathcal{V}$, i.e., $\epsilon=((1-v_i)\mathcal{I}_i(\{i\})+c_i(\{i\}))-(x_i-v_i\mathcal{I}_i(\mathcal{V})+c_i(\mathcal{V}))>0$, then platform 2 will have a profitable deviation by letting $y_i'=\mathcal{I}_i(\{i\})-\frac{\epsilon}{2}$ for all $i\in\mathcal{V}$. This is because in any user equilibria after this deviation at least one user shares on this platform, guaranteeing a positive profit.

We now consider the deviation of platform 1. For any $j \in \mathcal{V}$, the following prices guarantee that the only user equilibria is to have all users i in $\mathcal{V} \setminus \{j\}$ to share on platform 1 and user j share on platform 2:

$$x_i' = c_i(\mathcal{V}) - c_i(\{i\}) + y_i - v_i \mathcal{I}_i(\{i,j\}) + v_i \mathcal{I}_i(\mathcal{V} \setminus \{j\}) + \Delta, \quad i \in \mathcal{V} \setminus \{j\},$$

This is a user equilibrium because for any user $i \in \mathcal{V} \setminus \{j\}$ the price offered to her with the maximum leaked information on platform 1 and minimum joining value is larger than her payoff with the price offered on platform 2 with minimum leaked information and maximum joining value, i.e., we have

$$x_i' - v_i \mathcal{I}_i(\mathcal{V} \setminus \{j\}) + c_i(\{i\}) > y_i - v_i \mathcal{I}_i(\{i,j\}) + c_i(\mathcal{V}).$$

Also, user j shares on platform 2 because we have

$$\begin{aligned} y_j - v_j \mathcal{I}_j(\{j\}) + c_j(\{j\}) &\overset{(a)}{\geq} x_j - v_j \mathcal{I}_j(\mathcal{V}) - c_j(\mathcal{V}) + c_j(\{j\}) + c_j(\{j\}) \\ &\overset{(b)}{\geq} (1 - v_j)\mathcal{I}_j(\{j\}) + c_j(\{j\}) - c_j(\mathcal{V}) - c_j(\mathcal{V}) + c_j(\{j\}) + c_j(\{j\}) \\ &\overset{(c)}{\geq} (1 - v_j)\mathcal{I}_j(\{j\}) - 2\Delta + c_j(\{j\}) \overset{(d)}{\geq} c_j(\{j\}) \end{aligned}$$

where (a) follows from (B-16), (b) follows from (B-17), (c) follows from the definition of Δ, and (d) follows from condition (B-15).

This should not be a profitable deviation for platform 1, which leads to

$$\begin{aligned} &\sum_{i \neq j} \mathcal{I}_i(\mathcal{V} \setminus \{j\}) - x_i' = \sum_{i \neq j} \mathcal{I}_i(\mathcal{V} \setminus \{j\}) - c_i(\mathcal{V}) + c_i(\{i\}) - y_i + v_i \mathcal{I}_i(\{i,j\}) - v_i \mathcal{I}_i(\mathcal{V} \setminus \{j\}) \\ &\overset{(a)}{=} \sum_{i \neq j} \mathcal{I}_i(\mathcal{V} \setminus \{j\}) - c_i(\mathcal{V}) + c_i(\{i\}) - x_i - \Delta + v_i \mathcal{I}_i(\mathcal{V}) - v_i \mathcal{I}_i(\{i\}) - c_i(\mathcal{V}) + c_i(\{i\}) + v_i \mathcal{I}_i(\{i,j\}) - v_i \mathcal{I}_i(\mathcal{V} \setminus \{j\}) \\ &\overset{(b)}{\leq} \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathcal{V}) - x_i, \end{aligned}$$

where in (a) we used (B-16), and in (b) we used the fact that in the only user equilibrium after this deviation the payoff of platform 1 cannot increase. Rearranging the previous inequality and using the definition of Δ, for any $j \in \mathcal{V}$ we obtain

$$x_j \leq \mathcal{I}_j(\mathcal{V}) + \sum_{i \neq j} (1 - v_i) \left(\mathcal{I}_i(\mathcal{V}) - \mathcal{I}_i(\mathcal{V} \setminus \{j\}) \right) + v_i \left(\mathcal{I}_i(\{i\}) - \mathcal{I}_i(\{i,j\}) \right) + 3\Delta, \quad j \in \mathcal{V}. \tag{B-18}$$

We now consider the deviation of platform 2. In particular, we show that for any $j \in \mathcal{V}$, with price $y_j' = x_j - v_j \mathcal{I}_j(\mathcal{V}) + c_j(\mathcal{V}) + v_j \mathcal{I}_j(\{j\}) - c_j(\{j\}) + \Delta$ and zero for all other users the only user equilibrium is to have user j share on platform 2 and all other users share on platform 1. First note that all other users will still share on platform 1 as their information leakage has weakly decreased (since j is not sharing on platform 1) and they receive the same payment. now consider user j. She shares her data on platform 2, because with the choice of the price y'_j we have $y'_j - v_j\mathcal{I}_j(\{j\}) + c_j(\{j\}) > x_j - v_j\mathcal{I}_j(\mathcal{V}) + c_j(\mathcal{V})$. This cannot be a profitable deviation for platform 2 leading to $\mathcal{I}_j(\{j\}) - y'_j = \mathcal{I}_j(\{j\}) - x_j + v_j\mathcal{I}_j(\mathcal{V}) - v_j\mathcal{I}_i(\{j\}) \leq 0$. Therefore, we have

$$x_j \geq (1 - v_j)\mathcal{I}_j(\{j\}) + v_j\mathcal{I}_j(\mathcal{V}) - \Delta. \tag{B-19}$$

We next show that for sufficiently large $\mathbf{v}^{(l)}$ given the choice of $\tilde{\Delta}$, the inequalities (B-19), and (B-18) cannot simultaneously hold. In particular, consider $j \in \mathcal{V}$ who is with at least one other low-value user. For boundary low-values, i.e., $v_i = 1$, we must have

$$\Delta(3|\mathcal{V}| - 2) + \mathcal{I}_j(\mathcal{V}) + \sum_{i \neq j} (\mathcal{I}_i(\{i\}) - \mathcal{I}_i(\{i, j\})) \geq \mathcal{I}_j(\mathcal{V}),$$

which does not hold provided that $\Delta < \frac{1}{3|\mathcal{V}|-2} \sum_{i \neq j} \mathcal{I}_i(\{i, j\}) - \mathcal{I}_i(\{i\})$. Since $\Delta \leq \tilde{\Delta}$ for sufficiently large values $\mathbf{v}^{(l)}$, there exists no prices for which the first best can be sustained as an equilibrium, establishing inefficiency in this case.

Part 3-1: We first show that there exists $\bar{\mathbf{v}}$ and $\underline{\mathbf{v}}$ such that if $\mathbf{v}^{(h)} \geq \bar{\mathbf{v}}$, $\mathbf{v}^{(l)} \geq \underline{\mathbf{v}}$, and

$$\delta > \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)(\mathcal{I}_i(\mathcal{V}^{(l)}) - \mathcal{I}_i(\mathcal{V}_1^{(l)})), \tag{B-20}$$

the first best is to have all users joining the same platform, say platform 1, and only low-value users who are not correlated with high-value users share their data. Let $\mathcal{V}_1^{(l)} = \{i \in \mathcal{V}^{(l)} : \forall j \in \mathcal{V}^{(h)}, \Sigma_{ij} = 0\}$, be those low-value users uncorrelated with all high-value users and $\mathcal{V}_2^{(l)} = \mathcal{V} \setminus \mathcal{V}_2^{(l)}$ to denote the rest of the low-value users (i.e., low-value users correlated with at least one high-value user). The first best is to have all users join one of the platforms, say platform 1, because we can upper bound the social surplus when set J_k ($\neq \emptyset$ and $\neq \mathcal{V}$) joins platform $k \in \{1, 2\}$ and set S_k share on it as

$$\begin{aligned}
&\sum_{i \in J_1} (1 - v_i)\mathcal{I}_i(S_1) + c_i(J_1) + \sum_{i \in J_2} (1 - v_i)\mathcal{I}_i(S_2) + c_i(J_2) \\
&\overset{(a)}{\leq} -|\mathcal{V}|\delta + \sum_{i \in \mathcal{V}} c_i(\mathcal{V}) + \sum_{i \in J_1} (1 - v_i)\mathcal{I}_i(S_1) + \sum_{i \in J_2} (1 - v_i)\mathcal{I}_i(S_2) \\
&\overset{(b)}{\leq} -|\mathcal{V}|\delta + \sum_{i \in \mathcal{V}} c_i(\mathcal{V}) + \sum_{i \in J_1 \cap \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(S_1 \cap \mathcal{V}^{(l)}) + \sum_{i \in J_2 \cap \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(S_2 \cap \mathcal{V}^{(l)}) \\
&\leq -|\mathcal{V}|\delta + \sum_{i \in \mathcal{V}} c_i(\mathcal{V}) + \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) \overset{(c)}{\leq} \sum_{i \in \mathcal{V}} c_i(\mathcal{V}) + \sum_{i \in \mathcal{V}_1^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}_1^{(l)}),
\end{aligned}$$

where (a) follows from the definition of δ, (b) follows from the fact that for sufficiently large $\bar{v}$ only low-value users share and the information leakage of high-value users is zero, and (c) follows from condition (B-20). We next consider two possible cases and show that in both of them platform 2 has a profitable deviation.

Case 1: There exists no non-zero correlation between users in $\mathcal{V}_1^{(l)}$ and users in $\mathcal{V}_2^{(l)}$. We show that in this case platform 2 can induce all users in $\mathcal{V}_2^{(l)}$ to join and share their data. In particular, the following prices form a profitable deviation for platform 2:

$$\tilde{p}_i^2 = \Delta + v_i \mathcal{I}_i(\mathcal{V}_2^{(l)}), \quad i \in \mathcal{V}_2^{(l)},$$

and zero price to all other users. We first show that with these prices in any user equilibrium all users in $\mathcal{V}_2^{(l)}$ will join and share on platform 2. This is because the payoff of a user $i \in \mathcal{V}_2^{(l)}$ after deviating to share on platform 2 is

$$\tilde{p}_i^2 - v_i \mathcal{I}_i(S_2 \cup \{i\}) + c_i(J_2 \cup \{i\}) \overset{(a)}{\geq} \tilde{p}_i^2 - v_i \mathcal{I}_i(\mathcal{V}_2^{(l)}) + c_i(J_2 \cup \{i\}) \overset{(b)}{=} \Delta + c_i(J_2 \cup \{i\}) \overset{(c)}{\geq} c_i(J_1 \cup \{i\})$$

where (a) follows from the fact that the price offered by platform 2 to users outside of $\mathcal{V}_2^{(l)}$ is zero and hence they never share on platform 2, (b) follows from the choice of price offered by platform 2, and (c) follows from the definition of Δ. We next show that if

$$\sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i) \mathcal{I}_i(\mathcal{V}_2^{(l)}) > \Delta |\mathcal{V}_2^{(l)}|, \tag{B-21}$$

then the payoff of platform 2 after this deviation becomes strictly positive. Note that we can lower bound the payoff of platform 2 by

$$\sum_{i \in \mathcal{V}_2^{(l)}} \mathcal{I}_i(\mathcal{V}_2^{(l)}) - \tilde{p}_i^2 = \sum_{i \in \mathcal{V}_2^{(l)}} \mathcal{I}_i(\mathcal{V}_2^{(l)}) - \Delta - v_i \mathcal{I}_i(\mathcal{V}_2^{(l)}) = \sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i) \mathcal{I}_i(\mathcal{V}_2^{(l)}) - \Delta |\mathcal{V}_2^{(l)}| > 0.$$

Finally, we show that there exist $\bar{\delta}$ and $\bar{\Delta}$ such that for $\delta \geq \bar{\delta}$ and $\Delta \leq \bar{\Delta}$ both (B-20) and (B-21) hold. First note that since in this case users in $\mathcal{V}_1^{(l)}$ and those in $\mathcal{V}_2^{(l)}$ are uncorrelated, condition (B-20) becomes $\delta > \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i) \mathcal{I}_i(\mathcal{V}_2^{(l)})$. Therefore, letting

$$\Delta < \bar{\Delta} = \frac{1}{|\mathcal{V}_2^{(l)}|} \sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i) \mathcal{I}_i(\mathcal{V}_2^{(l)}), \text{ and } \delta > \bar{\delta} = \frac{1}{|\mathcal{V}^{(l)}|} \sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i) \mathcal{I}_i(\mathcal{V}_2^{(l)}),$$

completes the proof in this case.

Case 2: There exists $j \in \mathcal{V}_2^{(l)}$ who is correlated with at least one other user in $\mathcal{V}_1^{(l)}$. We show that platform 2 has a profitable deviation to take user j join and share on it. In particular, the following prices constitute a profitable deviation for platform 2:

$$\tilde{p}_j^2 = c_j(\mathcal{V}) - c_j(\{j\}) + v_j \mathcal{I}_j(\{j\}) - v_j \mathcal{I}_j(\mathcal{V}_1^{(l)}),$$

and zero price to all other user. First note that in any user equilibrium user j shares on platform 2. This is because $\tilde{p}_j^2 - v_j \mathcal{I}_j(\{j\}) + c_j(J_2 \cup \{j\}) = c_j(\mathcal{V}) - c_j(\{j\}) - v_j \mathcal{I}_j(\mathcal{V}_1^{(l)}) + \epsilon + c_j(J_2 \cup \{j\}) \geq c_j(\mathcal{V}) - v_j \mathcal{I}_j(\mathcal{V}_1^{(l)})$. This deviation is profitable for platform 2 provided that

$$(1 - v_j)\mathcal{I}_j(\{j\}) + v_j \mathcal{I}_j(\mathcal{V}_1^{(l)}) > \Delta. \tag{B-22}$$

Note that in this case $\mathcal{I}_j(\mathcal{V}_1^{(l)}) > 0$. Letting $\bar{\Delta} = \mathcal{I}_j(\mathcal{V}_1^{(l)}) \geq \Delta$ and $\bar{\delta} = \frac{1}{|\mathcal{V}^{(l)}|} \sum_{i \in \mathcal{V}_2^{(l)}} (1 - v_i)\mathcal{I}_i(\mathcal{V}_2^{(l)}) < \delta$, guarantees that the first best is not an equilibrium. The proof is completed by observing that for $\mathbf{v}^{(l)}$ sufficiently close to 1, we have $\bar{\Delta} > \bar{\delta}$.

Part 3-2: We prove that for

$$\delta > \tilde{\delta} = \max_{i \in \mathcal{V}} v_i \mathcal{I}_i(\mathcal{V}) + \sum_{j \in \mathcal{V}} \mathcal{I}_j(\mathcal{V}), \tag{B-23}$$

the first best is an equilibrium.

We first show that when δ is large, the first best involves all users joining the same platform, say platform 1. This is proved by showing that if users split and join different platforms social surplus decreases. Let $J_1 \neq \emptyset, \mathcal{V}$ be the set of users who join platform 1 and J_2 be the set of users who join platform 2. Also, let S_1 and S_2 be the sets of users who share on platforms 1 and 2, respectively. We can upper bound the surplus as

$$\begin{aligned}
&\left(\sum_{i \in J_1} c_i(J_1) + (1 - v_i)\mathcal{I}_i(S_1) \right) + \left(\sum_{i \in J_2} c_i(J_2) + (1 - v_i)\mathcal{I}_i(S_2) \right) \\
&\overset{(a)}{\leq} \left(\sum_{i \in \mathcal{V}} c_i(\mathcal{V}) - \delta \right) + \sum_{i \in J_1} (1 - v_i)\mathcal{I}_i(S_1) + \sum_{i \in J_2} (1 - v_i)\mathcal{I}_i(S_2) \\
&\overset{(b)}{\leq} -\delta|\mathcal{V}| + \sum_{i \in \mathcal{V}} c_i(\mathcal{V}) + \sum_{i \in \mathcal{V}} \mathcal{I}_i(\mathcal{V}) \overset{(c)}{<} \sum_{i \in \mathcal{V}} c_i(\mathcal{V}),
\end{aligned}$$

where (a) follows from the definition of δ and $J_1, J_2 \neq \emptyset$, (b) follows from replacing $(1 - v_i)\mathcal{I}_i(S_k)$ for $k = 1, 2$ by its upper bound $\mathcal{I}_i(\mathcal{V})$, and (c) follows from the condition given in (B-23). We next show that the first best, $\mathbf{a}^{\mathrm{W}}$, can be supported as an equilibrium. In particular, the prices $p_i^1 = v_i(\mathcal{I}_i(\mathbf{a}^{\mathrm{W}}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}^{\mathrm{W}}))$ for all $i \in \mathcal{V}$ and $p_i^2 = 0$ for all $i \in \mathcal{V}$ makes $\mathbf{a}^{\mathrm{W}}$ an equilibrium. We next verify that at these prices $\mathbf{a}^{\mathrm{W}}$ is indeed a user equilibrium and than that none of the platforms have a profitable deviation. It is a user equilibrium because the payoff of any user such as user i on platform 1 is larger than or equal to $c_i(\mathcal{V}) - v_i \mathcal{I}_i(\mathcal{V})$. If user i deviates and joins platform 2, then her payoff is smaller than or equal to $c_i(\{i\}) + \sum_{j \in \mathcal{V}} \mathcal{I}_j(\mathcal{V})$ (because the highest price offered to users on platform 2 is the total leaked information). User i does not have a profitable deviation because $c_i(\mathcal{V}) - v_i \mathcal{I}_i(\mathcal{V}) > c_i(\{i\}) + \delta - v_i \mathcal{I}_i(\mathcal{V}) \geq c_i(\{i\}) + \sum_{j \in \mathcal{V}} \mathcal{I}_j(\mathcal{V})$. We next show that platforms do not have a profitable deviation. Next suppose that platform 1 deviates and offers price vector $\tilde{\mathbf{p}}^1$. Since $c_i(\mathcal{V}) - v_i \mathcal{I}_i(\mathcal{V}) > c_i(\{i\}) + \delta - v_i \mathcal{I}_i(\mathcal{V}) \geq c_i(\{i\}) + \sum_{j \in \mathcal{V}} \mathcal{I}_j(\mathcal{V})$, all users joining platform 2 is a user equilibrium of the price vectors $(\tilde{\mathbf{p}}^1, \mathbf{p}^2)$. This user equilibrium gives platform 1 zero payoff and therefore deviating and offering price vector $\tilde{\mathbf{p}}^1$ is not profitable for platform 1. The same argument also establishes that platform 2 does not have a profitable deviation. ■

Proof of Proposition 8

We prove that with the price given by $p_i(\mathbf{v}) = \left(\mathcal{I}_i(\mathbf{a}(\mathbf{v})) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a}(\mathbf{v}))\right) + h_i(\mathbf{v}_{-i})$, for any function $h_i(\cdot)$ which depends only on $\mathbf{v}_{-i}$ where $\mathbf{a}(\mathbf{v})$ is defined as

$$\mathbf{a}(\mathbf{v}) = \text{argmax}_{\mathbf{a}\in\{0,1\}^n} \sum_{i\in\mathcal{V}}(1-v_i)\mathcal{I}_i(\mathbf{a}),$$

the users report their value of privacy truthfully. Consider user $i \in \mathcal{V}$ and suppose she reports v' instead of her true value of privacy v_i. Her payoff becomes

$$p_i(v', \mathbf{v}_{-i}) - v_i\mathcal{I}_i(\mathbf{a}(v', \mathbf{v}_{-i})) = \left(\mathcal{I}_i(\mathbf{a}(v', \mathbf{v}_{-i})) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a}(v', \mathbf{v}_{-i}))\right) + h_i(\mathbf{v}_{-i}) - v_i\mathcal{I}_i(\mathbf{a}(v', \mathbf{v}_{-i}))$$

$$= \left(\sum_{j\in\mathcal{V}}(1-v_j)\mathcal{I}_j(\mathbf{a}(v', \mathbf{v}_{-i}))\right) + h_i(\mathbf{v}_{-i}) \overset{(a)}{\leq} \left(\sum_{j\in\mathcal{V}}(1-v_j)\mathcal{I}_j(\mathbf{a}(v_i, \mathbf{v}_{-i}))\right) + h_i(\mathbf{v}_{-i}) \overset{(b)}{=} p_i(\mathbf{v}) - v_i\mathcal{I}_i(\mathbf{a}(\mathbf{v})),$$

where inequality (a) follows from the definition of $\mathbf{a}(\mathbf{v})$ as the maximizer of the surplus and equality (b) follows from the definition of $\mathbf{a}(\mathbf{v})$ and $\mathbf{p}(\mathbf{v})$, and $p_i(\mathbf{v}) - v_i\mathcal{I}_i(\mathbf{a}(\mathbf{v}))$ is equal to the payoff of user i if she reports truthfully.

Finally, note that with the choice of $h_i(\mathbf{v}_{-i}) = -\min_{\mathbf{a}\in\{0,1\}^n}\left(\sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a}) + \mathcal{I}_i(\mathbf{a})\right)$ the prices become nonnegative because

$$p_i(\mathbf{v}) = \left(\mathcal{I}_i(\mathbf{a}(\mathbf{v})) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a}(\mathbf{v}))\right) - \min_{\mathbf{a}\in\{0,1\}^n}\left(\mathcal{I}_i(\mathbf{a}) + \sum_{j\neq i}(1-v_j)\mathcal{I}_j(\mathbf{a})\right) \geq 0.$$

Proof of Theorem 11

Using the revelation principle, we can focus on direct mechanisms and then find the optimal direct incentive compatible mechanism. For the given prices, users in the set $\mathbf{a}(\mathbf{v})$ share their data. Letting $A_i(v, \mathbf{v}_{-i}) = \mathcal{I}_i(\mathbf{a}(v, \mathbf{v}_{-i})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}(v, \mathbf{v}_{-i}))$, the incentive compatibility constraint can be written as

$$p_i(v, \mathbf{v}_{-i}) - vA_i(v, \mathbf{v}_{-i}) \geq p_i(v', \mathbf{v}_{-i}) - vA_i(v', \mathbf{v}_{-i}). \quad \text{(B-24)}$$

Writing the first order condition for inequality (B-24) yields

$$p_i'(v, \mathbf{v}_{-i}) = vA_i'(v, \mathbf{v}_{-i}), \quad \forall v.$$

Taking integral of both sides yields

$$p_i(v, \mathbf{v}_{-i}) = -\int_v^{v_{\max}} x A_i'(x, \mathbf{v}_{-i}) dx \overset{(a)}{=} \int_v^{v_{\max}} A_i(x, \mathbf{v}_{-i}) dx + v A_i(v, \mathbf{v}_{-i}), \tag{B-25}$$

where we used integration by part in (a). The expected payment to user i is

$$\begin{aligned}
\mathbb{E}_{v \sim F_i}\left[p_i(v, \mathbf{v}_{-i})\right] &= \int p_i(v, \mathbf{v}_{-i}) f_i(v) dv \overset{(a)}{=} -\int_0^{v_{\max}} \int_v^{v_{\max}} x A_i'(v, \mathbf{v}_{-i}) dx f_i(v) dv \\
&\overset{(b)}{=} -\int_0^{v_{\max}} \int_0^{x} x A_i'(x, \mathbf{v}_{-i}) f_i(v) dv dx = -\int_0^{v_{\max}} F_i(x) x A_i'(x, \mathbf{v}_{-i}) dx \\
&\overset{(c)}{=} \left(\int_0^{v_{\max}} (f_i(x)x + F_i(x)) A_i(x, \mathbf{v}_{-i}) dx\right) - F_i(x) x A_i(x, \mathbf{v}_{-i})|_0^{v_{\max}} \\
&= \int_0^{v_{\max}} \left(x + \frac{F_i(x)}{f_i(x)}\right) f_i(x) A_i(x, \mathbf{v}_{-i}) dx.
\end{aligned}$$

where (a) follows from equation (B-25), (b) follows from changing the order of integrals, and (c) follows from integration by part. Therefore, the expected payment to user i becomes $\mathbb{E}_{\mathbf{v}}\left[p_i(\mathbf{v})\right] = \mathbb{E}_{\mathbf{v}}\left[\Phi_i(v_i) A_i(\mathbf{v})\right]$. Therefore, the expected payoff of the platform is

$$\mathbb{E}_{\mathbf{v}}\left[\sum_{i=1}^{n} \mathcal{I}_i(\mathbf{a}(\mathbf{v})) - \Phi_i(v_i)(\mathcal{I}_i(\mathbf{a}(\mathbf{v})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}(\mathbf{v})))\right].$$

The equilibrium sharing profile maximizes

$$\begin{aligned}
&\sum_{i=1}^{n} \mathcal{I}_i(\mathbf{a}(\mathbf{v})) - \Phi_i(v_i)\left(\mathcal{I}_i(\mathbf{a}(\mathbf{v})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}(\mathbf{v}))\right) \\
&= \sum_{i=1}^{n} (1 - \Phi_i(v_i)) \mathcal{I}_i(\mathbf{a}(\mathbf{v})) + \Phi_i(v_i) \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}(\mathbf{v})),
\end{aligned}$$

for any reported vector $\mathbf{v}$. Finally, note that maximizing this expression yield the following property: if $\mathbf{a}(v, \mathbf{v}_{-i}) = 1$, then for all $v' \leq v$ we have $\mathbf{a}(v', \mathbf{v}_{-i}) = 1$. This follows from the fact that by Assumption 2 we have $\Phi_i(v') \leq \Phi_i(v)$ for all $v' \leq v$. Since the leaked information is monotone, we obtain that $A_i(v, \mathbf{v}_i)$ is decreasing in v because as we increase v, $\Phi_i(v)$ increases, which means $a_i(v, \mathbf{v}_{-i})$ decreases, which in turn means that $A_i(v, \mathbf{v}_i) = \mathcal{I}_i(\mathbf{a}(v, \mathbf{v}_{-i})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}(v, \mathbf{v}_{-i}))$ decreases. This monotonicity property together with the payment identity (B-25) guarantees that the incentive compatibility constraint holds as we show next. Using the payment identity (B-25), the incentive compatibility constraint $p_i(v, \mathbf{v}_{-i}) - v A_i(v, \mathbf{v}_{-i}) \geq p_i(v', \mathbf{v}_{-i}) - v A_i(v', \mathbf{v}_{-i})$, is equivalent to

$$\left(\int_v^{v_{\max}} A_i(x, \mathbf{v}_{-i}) dx\right) + v A_i(v, \mathbf{v}_{-i}) - v A_i(v, \mathbf{v}_{-i}) \geq \left(\int_{v'}^{v_{\max}} A_i(x, \mathbf{v}_{-i}) dx\right) + v' A_i(v', \mathbf{v}_{-i}) - v A_i(v', \mathbf{v}_{-i}).$$

After canceling out the term $vA_i(v, \mathbf{v}_{-i})$ and rearranging the other terms, this inequality becomes equivalent to

$$\int_v^{v'} A_i(x, \mathbf{v}_{-i})dx \geq (v' - v)A_i(v', \mathbf{v}_{-i}).$$

To show this inequality we consider the following two possible cases:

- $v' \geq v$: we have

$$\int_v^{v'} A_i(x, \mathbf{v}_{-i})dx \overset{(a)}{\geq} \int_v^{v'} A_i(v', \mathbf{v}_{-i})dx = (v' - v)A_i(v', \mathbf{v}_{-i}),$$

  where (a) follows from the fact that $A_i(x, \mathbf{v}_{-i})$ is decreasing in x and hence for all $x \in [v, v']$ we have $A_i(x, \mathbf{v}_{-i}) \geq A_i(v', \mathbf{v}_{-i})$.

- $v' < v$: we have

$$\int_v^{v'} A_i(x, \mathbf{v}_{-i})dx = \int_{v'}^{v} -A_i(x, \mathbf{v}_{-i})dx \overset{(a)}{\geq} \int_{v'}^{v} -A_i(v', \mathbf{v}_{-i})dx = (v' - v)A_i(v', \mathbf{v}_{-i}),$$

  where (a) follows from the fact that $A_i(x, \mathbf{v}_{-i})$ is decreasing in x and hence for all $x \in [v', v]$ we have $-A_i(x, \mathbf{v}_{-i}) \geq -A_i(v', \mathbf{v}_{-i})$.

Proof of Corollary 3

Using Theorems 2 and 11, the difference between the expected payoff with and without knowing the value of privacy is

$$\begin{aligned}
&\mathbb{E}_{\mathbf{v}}\left[\max_{\mathbf{a}\in\{0,1\}^n} \sum_{i=1}^{n}(1 - v_i)\mathcal{I}_i(\mathbf{a}) + v_i\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0) - \max_{\mathbf{a}\in\{0,1\}^n} \sum_{i=1}^{n}(1 - \Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0)\right] \\
&\geq \mathbb{E}_{\mathbf{v}}\left[\sum_{i=1}^{n}(1 - v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(\mathbf{v})) + v_i\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}_{-i}(\mathbf{v}), a_i = 0) - \sum_{i=1}^{n}(1 - \Phi_i(v_i))\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(\mathbf{v})) + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}_{-i}(\mathbf{v}), a_i = 0)\right] \\
&\geq \mathbb{E}_{\mathbf{v}}\left[\sum_{i=1}^{n}\frac{F_i(v_i)}{f_i(v_i)}\left(\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}(\mathbf{v})) - \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}_{-i}(\mathbf{v}), a_i = 0)\right)\right]
\end{aligned}$$

Proof of Theorem 12

We first show a similar result to Lemma 3 that we use in this proof.

Lemma B-1. *All users for which we have $\Phi_i(v_i) \leq 1$ share their data in equilibrium.*

We establish this result by assuming the contrary and reaching a contradiction. Suppose, to obtain a contradiction, that in equilibrium $a_i^{\mathrm{E}} = 0$ for some $i \in \mathcal{V}$ where $\Phi_i(v_i) \leq 1$. We prove that there exists a deviation which increases the platform's payoff. In particular, the platform can deviate so that user i shares.

Using Theorem 11 the equilibrium action profile $\mathbf{a}^{\mathrm{E}}$ must maximize $\sum_{i\in\mathcal{V}}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$. We show that $(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i})$ increases this objective, which is a contradiction.

$$\begin{aligned}
&\sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_j(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_j(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) \\
&- \left(\sum_{j\in\mathcal{V}\setminus\{i\}} \Phi_j(v_j)(\mathcal{I}_j(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_j(a_i=1,a_j=0,\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}))\right) \\
&- \left(\sum_{j\in\mathcal{V}\setminus\{i\}} -\Phi_j(v_j)(\mathcal{I}_j(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_j(a_i=0,a_j=0,\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}))\right) \\
&+ \big((1-\Phi_i(v_i))\mathcal{I}_i(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i})\big) - \mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) \\
&\overset{(a)}{\geq} - \left(\sum_{j\in\mathcal{V}\setminus\{i\}} \Phi_j(v_j)(\mathcal{I}_j(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_j(a_i=1,a_j=0,\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}))\right) \\
&- \left(\sum_{j\in\mathcal{V}\setminus\{i\}} -\Phi_j(v_j)(\mathcal{I}_j(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_j(a_i=0,a_j=0,\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}))\right) \\
&+ \big((1-\Phi_i(v_i))\mathcal{I}_i(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i})\big) - \mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) \\
&\overset{(b)}{\geq} \big((1-\Phi_i(v_i))\mathcal{I}_i(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i})\big) - \mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}) \\
&= (1-\Phi_i(v_i))\big(\mathcal{I}_i(a_i=1,\mathbf{a}^{\mathrm{E}}_{-i}) - \mathcal{I}_i(a_i=0,\mathbf{a}^{\mathrm{E}}_{-i})\big) \overset{(c)}{\geq} 0,
\end{aligned}$$

where (a) follows from monotonicity of leaked information (part 1 of Lemma 1), (b) follows from the submodularity of information (part 2 of Lemma 1), and (c) follows from the fact that $\Phi_i(v_i)\leq 1$ and the leaked information is monotone. This is a contradiction because the equilibrium action profile must be the maximizer of $\sum_{i\in\mathcal{V}}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(a_i=0,\mathbf{a}_{-i})$. ■

We now proceed with the proof of theorem.

Part 1: Recall that for any action profile $\mathbf{a}$ and set $T\subseteq\mathcal{V}$, $\mathbf{a}_T$ denotes $(a_i:\ i\in T)$. Since, high-value users are uncorrelated with all other users, the first best is to have all low-value users share. This is because for any sharing profile $\mathbf{a}\in\{0,1\}^n$ we have

$$\begin{aligned}
\text{Social surplus}(\mathbf{a}) &= \sum_{i=1}^{n}(1-v_i)\mathcal{I}_i(\mathbf{a}) \overset{(a)}{=} \sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}}) + \sum_{i\in\mathcal{V}^{(h)}}(1-v_i)\mathcal{I}_i(\mathbf{a}_{\{i\}}) \\
&\overset{(b)}{\leq} \sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}}) \overset{(c)}{\leq} \sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathcal{V}^{(l)}) = \text{Social surplus}(\mathbf{1}_{\mathcal{V}^{(l)}}),
\end{aligned}$$

where (a) holds because high-value users' data are uncorrelated from all other users, (b) holds because for high-value users $(1-v_i)\mathcal{I}_i(\mathbf{a}_{\{i\}})\leq 0$, and (c) holds because the leaked information is monotone. We next show that in equilibrium only users in $\mathcal{V}^{(l)}_{\Phi}=\mathcal{V}^{(l)}$ share their data. Using Theorem 11, in the equilibrium the platform maximizes $\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a})+\Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i},a_i=$

0). We can write

$$\begin{aligned}
&\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a})+\Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i},a_i=0)\\
&\overset{(a)}{=}\sum_{i\in\mathcal{V}^{(l)}}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}})+\Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{\mathcal{V}^{(l)}\setminus\{i\}})+\sum_{i\in\mathcal{V}^{(h)}}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}_{\{i\}})\\
&\overset{(b)}{\le}\sum_{i\in\mathcal{V}^{(l)}}(1-\Phi_i(v_i))\mathcal{I}_i(\mathcal{V}^{(l)})+\Phi_i(v_i)\mathcal{I}_i(\mathcal{V}^{(l)}\setminus\{i\}),
\end{aligned}$$

where (a) holds because high-value users' data are uncorrelated from all other users and (b) holds because for high-value users $(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}_{\{i\}})\le 0$ and for users in $\mathcal{V}_\Phi^{(l)}$ we have replaced their leaked information by its maximum. This shows that in both equilibrium and first best all users in $\mathcal{V}^{(l)}=\mathcal{V}_\Phi^{(l)}$ share their data and no other user shares her data.

Part 2: We let $i\in\mathcal{V}_\Phi^{(l)}$ and $j\in\mathcal{V}^{(h)}$ be such that $\Sigma_{ij}>0$. Therefore, there exists $\delta>0$ such that $\mathcal{I}_j(\mathcal{V}_\Phi^{(l)})=\delta>0$. We next show that for $v_j>1+\frac{\sum_{i\in\mathcal{V}^{(l)}}\sigma_i^2}{\delta}$ the surplus of the action profile $\mathbf{a}^{\mathrm{E}}$ is negative, establishing that it does not coincide with the first best. We have

$$\begin{aligned}
\text{Social surplus}(\mathbf{a}^{\mathrm{E}})&=\sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}})+\sum_{i\in\mathcal{V}^{(h)}}(1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}})\overset{(a)}{\le}\sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\sigma_i^2+\sum_{i\in\mathcal{V}^{(h)}}(1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}})\\
&\overset{(b)}{\le}\left(\sum_{i\in\mathcal{V}^{(l)}}(1-v_i)\sigma_i^2\right)+(1-v_j)\mathcal{I}_j(\mathcal{V}_\Phi^{(l)})\le\left(\sum_{i\in\mathcal{V}^{(l)}}\sigma_i^2\right)+(1-v_j)\mathcal{I}_j(\mathcal{V}_\Phi^{(l)})\overset{(c)}{<}0
\end{aligned}$$

where in (a) for users in $\mathcal{V}_\Phi^{(l)}$ we have upper bounded the leaked information with its maximum, in (b) we have removed all the negative terms in the second summation except for the one corresponding to j for which we have replaced leaked information (of equilibrium action profile) by its minimum (using Lemma B-1), and in (c) we have used $v_j>1+\frac{\sum_{i\in\mathcal{V}^{(l)}}\sigma_i^2}{\delta}$.

Part 3: First note that for sufficiently large $\bar{\mathbf{v}}$, all users who are correlated with at least one of the high-value users do not share in the first best, and let this first best profile denoted by $\mathbf{a}$. Since user i is correlated with some of the high-value users, in the first best user $i\in\mathcal{V}^{(l)}\setminus\mathcal{V}_\Phi^{(l)}$ does not share and we also have

$$\sum_{j\in\mathcal{V}^{(h)}}\mathcal{I}_j(\mathbf{a})=0. \tag{B-26}$$

We next show that for sufficiently small $\bar{v}_i$ the action profile $\mathbf{a}$ cannot be equilibrium. Using Theorem 11, in the equilibrium the platform maximizes $\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a})+\Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i},a_i=0)$. We show that the action profile $\mathbf{a}'=(\mathbf{a}_{-i},a_i=1)$ leads to a higher $\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a})+$

$\Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0)$. We can write

$$\left(\sum_{j=1}^{n}(1-\Phi_j(v_j))\mathcal{I}_j(\mathbf{a}') + \Phi_j(v_j)\mathcal{I}_j(\mathbf{a}'_{-j}, a_j = 0)\right) - \left(\sum_{j=1}^{n}(1-\Phi_j(v_j))\mathcal{I}_j(\mathbf{a}) + \Phi_j(v_j)\mathcal{I}_j(\mathbf{a}_{-j}, a_j = 0)\right)$$
$$= \sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_j(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i})$$
$$- \left(\sum_{j\in\mathcal{V}\setminus\{i\}} \Phi_j(v_j)(\mathcal{I}_j(\mathbf{a}') - \mathcal{I}_j(\mathbf{a}'_{-j}, a_j = 0)) - \Phi_j(v_j)(\mathcal{I}_j(\mathbf{a}) - \mathcal{I}_j(\mathbf{a}_{-j}, a_j = 0))\right)$$
$$+ ((1-\Phi_i(v_i))\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$$
$$\overset{(a)}{\geq} \sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_j(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i})$$
$$+ \big((1-\Phi_i(v_i))\mathcal{I}_i(a_i = 1, \mathbf{a}^{\mathrm{E}}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i = 0, \mathbf{a}^{\mathrm{E}}_{-i})\big) - \mathcal{I}_i(a_i = 0, \mathbf{a}^{\mathrm{E}}_{-i})$$
$$\overset{(b)}{\geq} \left(\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})\right) + (1-\Phi_i(v_i))\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right)$$
$$\overset{(c)}{\geq} \left(\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})\right) + (1-\Phi_i(v_i))\mathcal{I}_i(\{i\}) \overset{(d)}{>} 0,$$

where (a) follows from submodularity of leaked information, (b) follows from equation B-26, (c) follows from submodularity of leaked information and the fact that $1-\Phi_i(v_i) < 0$, and (d) follows by letting $\Phi_i(v_i) \leq 1 + \frac{\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})}{\mathcal{I}_i(\{i\})}$. Letting $\tilde{v}$ be the minimum of $\Phi_i^{-1}\left(1 + \frac{\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})}{\mathcal{I}_i(\{i\})}\right)$ over all $i \in \hat{\mathcal{V}}^{(l)}$, completes the proof of this part.

Part 4: Let $i, k \in \mathcal{V}^{(h)}$ be such that $\Sigma_{ik} > 0$. First, note that in the first-best solution none of the high-value users share. We next show that if the value of privacy for high-value user i is small enough, then at least one high-value user shares in equilibrium. This establishes that any equilibrium is not efficient. We show this by assuming the contrary and then reaching a contradiction. Suppose that none of high-value users share. We show that if user i shares, the objective $\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0)$ increases. We let $\mathbf{a} \in \{0,1\}^n$ denote the equilibrium sharing profile and $\mathbf{a}'^n$

be the same actions profile except that user i is also sharing in $\mathbf{a}'$. Using this notation, let us write

$$\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}') + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}'_{-i}, a_i = 0) - \left(\sum_{i=1}^{n}(1-\Phi_i(v_i))\mathcal{I}_i(\mathbf{a}) + \Phi_i(v_i)\mathcal{I}_i(\mathbf{a}_{-i}, a_i = 0)\right)$$
$$= \sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_j(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i})$$
$$- \left(\sum_{j\in\mathcal{V}\setminus\{i\}} \Phi_j(v_j)(\mathcal{I}_j(\mathbf{a}') - \mathcal{I}_j(\mathbf{a}'_{-j}, a_j = 0)) - \Phi_j(v_j)(\mathcal{I}_j(\mathbf{a}) - \mathcal{I}_j(\mathbf{a}_{-j}, a_j = 0))\right)$$
$$+ ((1-\Phi_i(v_i))\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})$$
$$\overset{(a)}{\geq} \sum_{j\in\mathcal{V}\setminus\{i\}} \mathcal{I}_j(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_j(a_i = 0, \mathbf{a}_{-i})$$
$$+ \big((1-\Phi_i(v_i))\mathcal{I}_i(a_i = 1, \mathbf{a}^{\mathrm{E}}_{-i}) + \Phi_i(v_i)\mathcal{I}_i(a_i = 0, \mathbf{a}^{\mathrm{E}}_{-i})\big) - \mathcal{I}_i(a_i = 0, \mathbf{a}^{\mathrm{E}}_{-i})$$
$$\overset{(b)}{\geq} \left(\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})\right) + (1-\Phi_i(v_i))\left(\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i})\right)$$
$$\overset{(c)}{\geq} \left(\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})\right) + (1-\Phi_i(v_i))\mathcal{I}_i(\{i\}) \overset{(d)}{>} 0,$$

where (a) follows from the submodularity of leaked information, (b) follows from the fact that in the first best action profile $\mathbf{a}$ the leaked information of high-value users is zero, (c) follows from submodularity of leaked information and the fact that $1 - \Phi_i(v_i) < 0$, and (d) follows by letting $\Phi_i(v_i) < 1 + \frac{\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})}{\mathcal{I}_i(\{i\})}$. The proof is completed by letting $\bar{v}_i = \Phi^{-1}\left(1 + \frac{\sum_{j\in\mathcal{V}^{(h)}} \mathcal{I}_j(\{i\})}{\mathcal{I}_i(\{i\})}\right)$. ■

Proof of Proposition 10

We first show that for any user $i \in \mathcal{V}$ for which $a_i^{\mathrm{W}} = 0$, we have $a_i^{\mathrm{E}} = 0$. Suppose the contrary, i.e., $a_i^{\mathrm{W}} = 0$ and $a_i^{\mathrm{E}} = 1$. From the choice of taxation policy, since $a_i^{\mathrm{W}} = 0$, we have a non-zero tax on user i's data. We next compare the platform's payoff with and without user i's data and show that the sharing profile $\mathbf{a}^E$ in which $a_i^{\mathrm{E}} = 1$ cannot be equilibrium. Using the price characterization of Theorem 2,

$$U(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 1, \mathbf{p}^{\mathbf{a}^{\mathrm{E}}}) - U(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 0, \mathbf{p}^{a_i=0,\mathbf{a}^{\mathrm{E}}_{-i}})$$
$$= \left(\sum_{j=1}^{n}(1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 1) + v_j\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 1)\right) - t_i$$
$$- \left(\sum_{j=1}^{n}(1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 0) + v_j\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 0)\right) \overset{(a)}{\leq} -t_i + \sum_{i\,:v_j\leq 1} \sigma_j^2 + \sum_{j\,:v_j>1} v_j\sigma_j^2 \overset{(b)}{<} 0,$$

where (a) follows from the fact that for $v_j \leq 1$, we have $(1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 1) + v_j \mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 1) \leq \mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 1) \leq \sigma_j^2$ and $(1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 0) + v_j \mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 0) \geq 0$. Also, for $v_j \geq 1$ we have $(1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 1) - (1-v_j)\mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-i}, a_i = 0) \leq 0$ and $v_j \mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 1) - v_j \mathcal{I}_j(\mathbf{a}^{\mathrm{E}}_{-\{i,j\}}, a_j = 0, a_i = 0) \leq v_j \sigma_j^2$. The inequality (b) follows because of the choice of taxation t_i. On the other hand, if $a_i^{\mathrm{E}} = 1$, then we must have $U(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 1, \mathbf{p}^{\mathbf{a}^{\mathrm{E}}}) \geq U(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 0, \mathbf{p}^{a_i = 0, \mathbf{a}^{\mathrm{E}}_{-i}})$. This is a contradiction, proving the claim.

Therefore, we must have $\mathbf{a}^{\mathrm{E}} \leq \mathbf{a}^{\mathrm{W}}$. We next prove that the equilibrium is $\mathbf{a}^{\mathrm{W}}$, i.e., this inequality cannot be strict. Suppose the contrary, i.e., $\mathbf{a}^{\mathrm{E}} < \mathbf{a}^{\mathrm{W}}$ is equilibrium. Using Theorem 2 we have

$$\sum_{i=1}^{n} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) + v_i \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 0) \geq \sum_{i=1}^{n} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{W}}) + v_i \mathcal{I}_i(\mathbf{a}^{\mathrm{W}}_{-i}, a_i^{\mathrm{W}} = 0). \tag{B-27}$$

On the other hand, using Lemma 1 and $\mathbf{a}^{\mathrm{E}} < \mathbf{a}^{\mathrm{W}}$, we have

$$\sum_{i=1}^{n} v_i \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}_{-i}, a_i^{\mathrm{E}} = 0) < \sum_{i=1}^{n} v_i \mathcal{I}_i(\mathbf{a}^{\mathrm{W}}_{-i}, a_i^{\mathrm{W}} = 0). \tag{B-28}$$

Putting (B-27) and (B-28) together, we obtain

$$\sum_{i=1}^{n} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) > \sum_{i=1}^{n} (1-v_i)\mathcal{I}_i(\mathbf{a}^{\mathrm{W}}),$$

which contradicts the fact that $\mathbf{a}^{\mathrm{W}}$ is the first-best solution. ■

Proof of Lemma 5

Without loss of generality, suppose $a_2 = \cdots = a_n = 1$. We have $\mathbb{E}[\mathbf{X}\tilde{\mathbf{S}}^T] = \mathbb{E}[\mathbf{X}\mathbf{S}^T]\Sigma^{-1} = I$. We also have $\mathbb{E}[\tilde{\mathbf{S}}\tilde{\mathbf{S}}^T] = \Sigma^{-1}\mathbb{E}[\mathbf{S}\mathbf{S}^T]\Sigma^{-1} = \Sigma^{-1}(I+\Sigma)\Sigma^{-1}$. We first find the leaked information of user 1 if she does not share. Since, the correlation between user 1's type and the shared data $\tilde{S}_2, \ldots, \tilde{S}_n$ is zero, this leaked information is zero.

We next find the leaked information of user 1 if she shares her information. Note that $\tilde{\mathbf{S}}$ and X_1 are jointly normal. Using the characterization of Theorem 2, this leaked information is equal to

$$\begin{aligned}
(1,0,\ldots,0)\left(\Sigma^{-1}(I+\Sigma)\Sigma^{-1}\right)^{-1}(1,0,\ldots,0)^T &= (1,0,\ldots,0)\Sigma(I+\Sigma)^{-1}\Sigma(1,0,\ldots,0)^T \\
&= (\sigma_1^2, \Sigma_{12}, \ldots, \Sigma_{1n})(I+\Sigma)^{-1}(\sigma_1^2, \Sigma_{12}, \ldots, \Sigma_{1n})^T \\
&= \mathcal{I}_1(a_1 = 1, \mathbf{a}_{-1}),
\end{aligned}$$

where the last equality follows from Theorem 2. This completes the proof of Lemma. ■

Proof of Theorem 13

First note that the minimum price offered to user i to share her information with action profile $\mathbf{a}_{-i}$ must make her indifferent between her payoff if she shares which is given by $p_i - v_i \mathcal{I}_i(a_i = 1, \mathbf{a}_{-1})$ (where we used Lemma 5) and her payoff if she does not share which is zero. Therefore, the minimum price offered to users i to share is $v_i \mathcal{I}_i(a_i = 1, \mathbf{a}_{-i})$. Since the leaked information of users who do not share is zero, for a given action profile $\mathbf{a} \in \{0,1\}^n$, the payoff of the platform with minimum prices becomes $\sum_{i:\ a_i=1} \mathcal{I}_i(\mathbf{a}) - \sum_{i:\ a_i=1} p_i = \sum_{i:\ a_i=1} (1 - v_i) \mathcal{I}_i(\mathbf{a})$. Choosing the action profile that maximizes this payoff, completes the proof.

Proof of Theorem 14

Suppose $\mathbf{a}^{\mathrm{E}}$ is the equilibrium action profile before de-correlation. Using Lemma 3, all low-value users share in $\mathbf{a}^{\mathrm{E}}$. We have

$$\begin{aligned}
\text{Social surplus}(\mathbf{a}^{\mathrm{E}}) &= \sum_{i \in \mathcal{V}} (1 - v_i) \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) = \sum_{i:\ a_i^{\mathrm{E}}=1} (1 - v_i) \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) + \sum_{i:\ a_i^{\mathrm{E}}=0} (1 - v_i) \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) \\
&\overset{(a)}{\leq} \sum_{i:\ a_i^{\mathrm{E}}=1} (1 - v_i) \mathcal{I}_i(\mathbf{a}^{\mathrm{E}}) \overset{(b)}{\leq} \text{Social surplus}(\tilde{\mathbf{a}}^{\mathrm{E}}),
\end{aligned}$$

where (a) follows from the fact that for all i such that $a_i^{\mathrm{E}} = 0$ we must have $1 - v_i < 0$ (because all low-value users share in equilibrium), and (b) follows because using Theorem 13 the action profile $\tilde{\mathbf{a}}^{\mathrm{E}}$ is the maximizer of $\sum_{i:\ a_i=1} (1 - v_i) \mathcal{I}_i(\mathbf{a})$. Finally, note that the equilibrium social surplus after de-correlation cannot be negative because it is equal to $\sum_{i \in \mathcal{V}} (1 - v_i) \tilde{\mathcal{I}}_i(\mathbf{a}^{\mathrm{E}}) \geq \sum_{i \in \mathcal{V}} (1 - v_i) \tilde{\mathcal{I}}_i(\mathbf{a} = \mathbf{0}) = 0$.

B.3 Examples

Details of Example 1 and Section 3.4

We first present a lemma that enables us to find all possible prices for which a given action profile is a user equilibrium in Example 1.

Lemma B-2. *Consider a game with $\mathcal{V} = \{1,2\}$, values of privacy (v_1, v_2), and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The price vectors (p_1, p_2) that can sustain each of the four possible user equilibria are as follows.*

1. *$a_1 = a_2 = 1$ is a user equilibrium if $p_1 \geq v_1 \frac{(2-\rho^2)^2}{2(4-\rho^2)}$ and $p_2 \geq v_2 \frac{(2-\rho^2)^2}{2(4-\rho^2)}$,*
2. *$a_1 = 1$ and $a_2 = 0$ is a user equilibrium if $p_1 \geq \frac{v_1}{2}$ and $p_2 \leq v_2 \frac{(2-\rho^2)^2}{2(4-\rho^2)}$.*
3. *$a_1 = 0$ and $a_2 = 1$ is a user equilibrium if $p_2 \geq \frac{v_2}{2}$ and $p_1 \leq v_1 \frac{(2-\rho^2)^2}{2(4-\rho^2)}$.*
4. *$a_1 = 0$ and $a_2 = 0$ is a user equilibrium if $p_1 \leq \frac{v_1}{2}$ and $p_2 \leq \frac{v_2}{2}$.*

1. With $a_1 = 1$ and $a_2 = 1$, the payoff of user 1 is $u_1(a_1 = 1, a_2 = 1, \mathbf{p}) = p_1 - v_1\left(1 - \frac{2-\rho^2}{4-\rho^2}\right)$. If user 1 deviates and plays $a_1 = 0$, her payoff becomes $u_1(a_1 = 0, a_2 = 1, \mathbf{p}) = -v_1\left(1 - \frac{2-\rho^2}{2}\right)$. Therefore, as long as we have $p_1 \geq v_1\left(\frac{2-\rho^2}{2} - \frac{2-\rho^2}{4-\rho^2}\right) = v_1\frac{(2-\rho^2)^2}{2(4-\rho^2)}$, user 1 does not have an incentive to deviate. Because of symmetry, similarly, user 2 does not have an incentive to deviate either.

2. With $a_1 = 1$ and $a_2 = 0$, the payoff of user 1 is $u_1(a_1 = 1, a_2 = 0, \mathbf{p}) = p_1 - v_1\left(1 - \frac{1}{2}\right)$, and the payoff of user 2 is $u_2(a_1 = 1, a_2 = 0, \mathbf{p}) = -v_2\left(1 - \frac{2-\rho^2}{2}\right)$. If user 1 deviates and plays $a_1 = 0$, her payoff becomes $u_1(a_1 = 0, a_2 = 0, \mathbf{p}) = 0$. Therefore, as long as we have $p_1 \geq \frac{v_1}{2}$, user 1 does not have an incentive to deviate. If user 2 deviates and plays $a_2 = 1$, her payoff becomes $u_1(a_1 = 1, a_2 = 1, \mathbf{p}) = p_2 - v_1\left(1 - \frac{2-\rho^2}{4-\rho^2}\right)$. Therefore, as long as $p_2 \leq v_2\left(\frac{2-\rho^2}{2} - \frac{2-\rho^2}{4-\rho^2}\right) = v_2\frac{(2-\rho^2)^2}{2(4-\rho^2)}$, user 2 does not have an incentive to deviate.

3. The proof of this case is the same as the previous case and can be obtained by swapping users 1 and 2.

4. With $a_1 = 0$ and $a_2 = 0$, the payoff of user 1 is $u_1(a_1 = 0, a_2 = 0, \mathbf{p}) = 0$. If user 1 deviates and plays $a_1 = 1$, her payoff becomes $u_1(a_1 = 1, a_2 = 0, \mathbf{p}) = p_1 - v_1\left(1 - \frac{1}{2}\right)$. Therefore, as long as we have $p_1 \leq v_1\frac{1}{2}$, user 1 does not have an incentive to deviate. Similarly, user 2 does not have an incentive to deviate. ■

Using this lemma we have the following characterization which is depicted in Example 1.

Proposition B-1. *Consider a game with* $\mathcal{V} = \{1, 2\}$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *and* $v_1 = v_2 = v$.

- *The equilibrium prices to sustain* $a_1 = a_2 = 1$ *are* $p_1 = p_2 = v\frac{(2-\rho^2)^2}{2(4-\rho^2)}$. *The corresponding payoffs are*
$$u_1(\mathbf{a}, \mathbf{p}) = u_2(\mathbf{a}, \mathbf{p}) = \frac{-v\rho^2}{2}, \quad U(\mathbf{a}, \mathbf{p}) = \frac{4}{4-\rho^2} - v\frac{(2-\rho^2)^2}{4-\rho^2}.$$
- *The equilibrium prices to sustain* $a_1 = 1$ *and* $a_2 = 0$ *are* $p_1 = \frac{v}{2}$, $p_2 = 0$. *The corresponding payoffs are*
$$u_1(\mathbf{a}, \mathbf{p}) = 0, u_2(\mathbf{a}, \mathbf{p}) = \frac{-v\rho^2}{2}, \quad U(\mathbf{a}, \mathbf{p}) = \frac{1+\rho^2}{2} - \frac{v}{2}.$$
- *The equilibrium prices to sustain* $a_1 = a_2 = 0$ *are* $p_1, p_2 \leq \frac{v}{2}$. *The corresponding payoffs are all equal to* 0.

The equilibrium price vector for action profile $(a_1 = 1, a_2 = 1)$ is $p_1 = p_2 = v\frac{(2-\rho^2)^2}{2(4-\rho^2)}$ with the corresponding platform's payoff equal to

$$U\left((a_1 = 1, a_2 = 1), \left(p_1 = \frac{(2-\rho^2)^2}{2(4-\rho^2)}, p_2 = \frac{(2-\rho^2)^2}{2(4-\rho^2)}\right)\right) = -2v\frac{(2-\rho^2)^2}{2(4-\rho^2)} + 2\left(1 - \frac{2-\rho^2}{4-\rho^2}\right)$$
$$= \frac{4}{4-\rho^2} - v\frac{(2-\rho^2)^2}{4-\rho^2}. \qquad \text{(B-29)}$$

We also have

$$u_i\left((a_1=a_2=1),\left(p_1=p_2=v\frac{(2-\rho^2)^2}{2(4-\rho^2)}\right)\right)=v\frac{(2-\rho^2)^2}{2(4-\rho^2)}-v\frac{2}{4-\rho^2}=-v\frac{\rho^2}{2},\quad i=1,2.$$

The equilibrium price vector for actions $(a_1=1,a_2=0)$ is $p_1=\frac{v}{2}$ and $p_2\le\frac{v}{2}$ with the corresponding payoff equal to

$$U\left((a_1=1,a_2=0),\left(p_1=\frac{v}{2},p_2=0\right)\right)=\frac{-v}{2}+\left(1-\frac{2-\rho^2}{2}\right)+\left(1-\frac{1}{2}\right)=\frac{-v}{2}+\frac{1+\rho^2}{2}. \tag{B-30}$$

Similarly, the equilibrium price vector for actions $(a_1=0,a_2=1)$ is $p_2=\frac{v}{2}$ and $p_1\le\frac{v}{2}$ with the corresponding payoff equal to

$$U\left((a_1=0,a_2=1),\left(p_1=0,p_2=\frac{v}{2}\right)\right)=\frac{-v}{2}+\frac{1+\rho^2}{2}.$$

Finally, the equilibrium price vector for actions $(a_1=0,a_2=0)$ is $p_1=p_2=0$ with the corresponding payoff equal to $U((a_1=0,a_2=0),(p_1=0,p_2=0))=0$. This completes the proof.

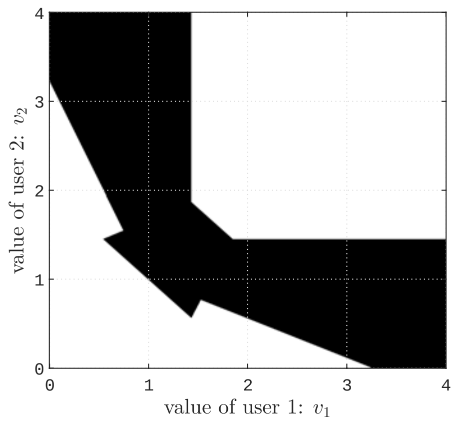
Lemma B-2 leads to the following characterization of equilibrium which is depicted in Section 3.4.

Proposition B-2. *Consider a game with $\mathcal{V}=\{1,2\}$, $\Sigma=\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}$ and $v_1=v_2=v$. We have the following cases:*

1. *If $v\ge\frac{4}{(2-\rho^2)^2}$, then the equilibrium sharing profile is $a_1=a_2=0$ with equilibrium prices $p_1,p_2\le\frac{v}{2}$. The corresponding social surplus in this case is 0.*
2. *If $v\le\frac{4}{(2-\rho^2)^2}$, then the equilibrium sharing profile is $a_1=a_2=1$ with equilibrium prices $p_1=p_2=v\frac{(2-\rho^2)^2}{2(4-\rho^2)}$. The corresponding social surplus in this case is $\frac{4}{4-\rho^2}(1-v)$.*

Using the characterization of user equilibria, and in particular comparing (B-29) and (B-30), we conclude that if $v\le\frac{4}{(2-\rho^2)^2}$, then the equilibrium decisions are given by $p_1=p_2=\frac{(2-\rho^2)^2}{2(4-\rho^2)}$, and $a_1=a_2=1$, with the corresponding platform's payoff $\frac{4}{4-\rho^2}-v\frac{(2-\rho^2)^2}{4-\rho^2}$. Otherwise, if $v>\frac{4}{(2-\rho^2)^2}$, then the equilibrium decisions are given by $p_1,p_2\le\frac{v}{2}$ and $a_1=a_2=0$ with the corresponding platform's payoff 0.

We end the discussion of this example by showing what happens if the users have different values for privacy. Figure B-1 shows the pair of (v_1,v_2) for which the equilibrium surplus is negative and therefore shutting down the data market improves the social surplus. As we can see in two cases the equilibrium surplus becomes negative: (i) one of the users has a relatively low value for privacy, denoted by i, and the other one denoted by j has a very high value for privacy (i.e., much larger than 1). In this case, the platform purchases the data of user i and because of data externality this leaks information about user j, leading to a negative surplus (if the value of



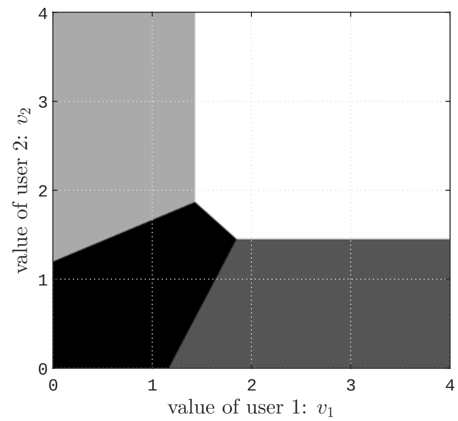


Figure B-1: In Panel (a), the black area shows the pair of (v_1, v_2) for which the equilibrium surplus is negative in the setting of the example given in Section 3.4 when $\rho = 2/3$. In panel (b), the black area shows the pair of (v_1, v_2) for which the equilibrium action profile is $(a_1^{\mathrm{E}}, a_2^{\mathrm{E}}) = (1, 1)$, the gray area on the southeast shows the pair of (v_1, v_2) for which $(a_1^{\mathrm{E}}, a_2^{\mathrm{E}}) = (0, 1)$, and the gray area on the northwest shows the pair of (v_1, v_2) for which $(a_1^{\mathrm{E}}, a_2^{\mathrm{E}}) = (1, 0)$.

user i is less than 1, the platform purchases her information even if there were no externality, but if the value of user i is larger but close to 1, the platform buys user i's data because it also leaks about user j), (ii) both users have an intermediate value for privacy (larger than 1, but not too large). In this case, users would have not shared if there were no externalities. However, since the externality of data exists, both users share in equilibrium which leads to a negative surplus.

Details of Example 5

We first find a closed-form expression for the leaked information with 3 users. For any $i \in \{1, 2, 3\}$, we have

$$\mathcal{I}_i(\{i\}) = \frac{\sigma_i^4}{1 + \sigma_i^2}.$$

For any $i \neq j$, we have

$$\mathcal{I}_i(\{j\}) = \frac{\Sigma_{ij}^2}{1 + \sigma_j^2}.$$

For any $i \neq j$, we have

$$\mathcal{I}_i(\{i, j\}) = \frac{\Sigma_{ij}^2 - \sigma_i^2 \Sigma_{ij}^2 + \sigma_i^4 (1 + \sigma_j^2)}{1 - \Sigma_{ij}^2 + \sigma_j^2 + \sigma_i^2 (1 + \sigma_j^2)}.$$

Substituting the specific numbers given in the example in the expressions above lead to the following inequalities

$$\mathcal{I}_1(\{2\}) > 0 \tag{B-31}$$

$$\mathcal{I}_2(\{1\}) < \mathcal{I}_2(\{3\}) \tag{B-32}$$

$$\mathcal{I}_3(\{2\}) < \mathcal{I}_3(\{1\}) \tag{B-33}$$

$$(1 - v_3)\mathcal{I}_3(\{2,3\}) + v_3\mathcal{I}_3(\{2\}) + (1 - v_2)\mathcal{I}_2(\{2,3\}) + v_2\mathcal{I}_2(\{3\}) \geq \mathcal{I}_3(\{2\}) + (1 - v_2)\mathcal{I}_2(\{2\}) \tag{B-34}$$

$$(1 - v_1)\mathcal{I}_1(\{1\}) + \mathcal{I}_3(\{1\}) \geq (1 - v_1)\mathcal{I}_1(\{1,3\}) + v_1\mathcal{I}_1(\{3\}) + (1 - v_3)\mathcal{I}_3(\{1,3\}) + v_3\mathcal{I}_3(\{1\}) \tag{B-35}$$

We are now ready to show that first-stage game does not have a pure strategy equilibrium. We have the following four possible candidates for a pure-strategy equilibrium (and the symmetric cases of these):

1. $J_1 = \{1,2\}$: Since $v_1 < 1$ and $v_2 < 1$, in the equilibrium of the second-stage game both users 1 and 2 share and user 3 does not share on platform 2. Therefore, the payoff of users become

$$u_1 = -v_1\mathcal{I}_1(\{2\}) + c, \quad u_2 = -v_2\mathcal{I}_2(\{1\}) + c, \quad u_3 = c.$$

2. $J_1 = \{2,3\}$: Since $v_1 < 1$ user 1 share on Platform 2. Now on platform 1 since $v_2 < 1$, user 2 share her data in the second-stage game. The question is whether user 3 shares on platform 1. The answer is positive because (B-34) shows that platform 1's payoff when both users 2 and 3 share is larger than when only user 2 shares. Therefore, the payoff of users in this case becomes

$$u_1 = c, \quad u_2 = -v_2\mathcal{I}_2(\{3\}) + c, \quad u_3 = -v_3\mathcal{I}_3(\{2\}) + c.$$

3. $J_1 = \{1,3\}$: Since $v_2 < 1$ user 2 share on Platform 2. Now on platform 1 since $v_1 < 1$, user 1 share her data in the second-stage game. The question is whether user 3 shares on platform 1. The answer is negative because (B-35) shows that platform 1's payoff when both users 1 and 3 share is smaller than when only user 1 shares. Therefore, the payoff of users in this case becomes

$$u_1 = c, \quad u_2 = c, \quad u_3 = -v_3\mathcal{I}_3(\{1\}) + c.$$

4. $J_1 = \{1,2,3\}$: Since both v_1 and v_2 are below 1, users 1 and 2 share in equilibrium. This means that the payoff of user 3 becomes less than c. User 3 has an incentive to deviate and join the second platform with payoff c. Therefore, this case can never be the equilibrium of the first-stage game.

We next prove that in each one of the cases, there exists a user with a profitable deviations. We have shown this for case 4 already. For case 1, because of (B-31), user 1 has a profitable deviation from joining the second platform (together with user 3). For case 2, because of (B-32), user 2 has a profitable deviation from joining the second platform (together with user 1). For case 3, because of (B-33), user 3 has a profitable deviation from joining the second platform (together with user 2).

Therefore, none of the above candidates can be a pure strategy equilibrium of the first stage game, completing the proof of the claim in Example 5.

Details of Examples 4 and 7

We let $1, \ldots, n$ denote the users with value below 1 in community 1 and $n+1, \ldots, n+m$ denote the users in community 2 whose value of privacy are above 1. Using Lemma 3, in equilibrium all users of community 1 will share their data. We also denote the cross-community correlations by ρ and with-in community correlations by r. We next find the leaked information of users if k users of community 2 share their data.

Let $\mathcal{I}(0, k_1, k_2)$ denote the leaked information of a user in community 1 when she does not share her data, k_1 users in community 1 share their data, and k_2 users in community 2 share their data. Note that this leaked information only depends on k_1 and k_2 and is given below

$$\mathcal{I}(0, k_1, k_2) = \frac{k_1 r^2(2 + (k_2 - 1)r) + 2k_2\rho^2 - k_2(k_1 + 1)r\rho^2}{4 + (k_2 - 1)(k_1 - 1)r^2 + 2(k_2 + k_1 - 2)r - k_2 k_1 \rho^2}.$$

Now, using this expression for the leaked information and Theorem 2, letting $\mathcal{I}(1, k_1, k_2)$ denote the leaked information of user 1 when she also shares her data we obtain

$$\mathcal{I}(1, k_1, k_2) = \mathcal{I}(0, k_1, k_2) + \frac{(1 - \mathcal{I}(0, k_1, k_2))^2}{2 - \mathcal{I}(0, k_1, k_2)}.$$

Using these notations, the equilibrium of the setting of Example 4 will be determined by the number of users in community 2 who share their data which is the solution of the following optimization

$$\max_{1 \le x \le m} n\left(v_l \mathcal{I}(0, n-1, x) + (1 - v_l)\mathcal{I}(1, n-1, x)\right)$$
$$+ x\left(v_h \mathcal{I}(0, x-1, n) + (1 - v_h)\mathcal{I}(1, x-1, n)\right) + (m - x)\mathcal{I}(0, x, n),$$

where $v_l = .9$, $n = 10$, $m = 10$, and v_h is as specified in Example 4. Solving this optimization problem numerically, and then considering the surplus of equilibrium which is equal to

$$n(1 - v_l)\mathcal{I}(1, n-1, x^*) + x^*(1 - v_h)\mathcal{I}(1, x^* - 1, n) + (m - x^*)(1 - v_h)\mathcal{I}(0, x^*, n),$$

where x^* is the number of users from community 2 who share in equilibrium leads to the results of Example 4. Similarly, for the setting of Example 7, we first uniformly at random select the joining decisions and then solve the corresponding optimization of the platforms to find the user equilibrium for each of them.

Examples of Section 4.6

We provide two examples of a setting with only high-value users and show that the equilibrium could be efficient or inefficient. We also provide an example that illustrates the results of part 3 of Theorem 10.

Example B-1 (Inefficiency with High-values). Consider a setting with two users and suppose the joining value $c(\cdot)$ is a function of the cardinality of the joining set. Also, suppose we have two high-value users 1 and 2. There exists $\bar{v}$ such that if $v_1, v_2 \leq \bar{v}$, then provided that

$$c(2) - c(1) \leq \min \left\{ \mathcal{I}_2(\{1\}), \frac{\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{1\})}{2} \right\},$$

the first best is not an equilibrium.

This example also establishes that there exists a setting with only high-value users in which for values of privacy close to 1, the equilibrium is inefficient.

Example B-2 (Efficiency with High-values). Consider a setting with $n \geq 3$ users and suppose the joining value $c(\cdot)$ is a function of the cardinality of the joining set. Suppose in addition that two high-value users 1 and 2 are correlated and the rest of the high-value users are uncorrelated with all others and have large values of privacy. In this setting, the first best is an equilibrium.

In contrast to Example B-1, this example establishes that there exists a setting with only high-value users for which even for values of privacy close to 1, the equilibrium is efficient.

We next provide an example illustrating the results of part 3 of Theorem 10.

Example B-3 (Efficiency when High-Value and Low-Value Users are Correlated). Consider a setting with $n = 2$ users and suppose the joining value $c_i(\cdot)$ is a function of the cardinality of the joining set, i.e., $c_i(S) = \delta|S|$ for $i = 1, 2$. Also, suppose $v_1 \leq 1$ and $v_2 > \max\{1 + \frac{\mathcal{I}_1(\{1\})}{\mathcal{I}_2(\{1\})}, \frac{\delta}{\mathcal{I}_2(\{1\})}\}$ and the data of these two users are correlated. For $\delta \geq \frac{1}{2}(1 - v_1)\mathcal{I}_1(\{1\})$, the first best is to have both users joining (and not sharing) one of the platforms, which without loss of generality we assume is platform 1. Depending on the value of δ, we have the following cases:

- If $\delta < (1 - v_1)\mathcal{I}_1(\{1\})$, the first best is not an equilibrium.
- If $\delta \geq (1 - v_1)\mathcal{I}_1(\{1\})$, the first best is an equilibrium.

This example shows that, as claimed in part 3 of Theorem 10, for intermediate values of δ the equilibrium is inefficient while for large values of δ the equilibrium becomes efficient.

We next show the proof of the statements of Examples B-1, B-2, and B-3.

Proof of the statement of Example B-1: Note that the first best is to have both users join one of the platforms, say platform 1, and none of them shares.

We show that the following prices form a profitable deviation for platform 2:

$$\tilde{p}_1^2 = \max\{v_1\mathcal{I}_1(\{1\}) + c(2) - c(1), v_1\mathcal{I}_1(\{1,2\}) + c(1) - c(2), v_1(\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{2\}))\} + \epsilon, \tag{B-36}$$

$$\tilde{p}_2^2 = \max\{v_2\mathcal{I}_2(\{1,2\}) + c(1) - c(2), v_2(\mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\}))\} + \epsilon. \tag{B-37}$$

for some $\epsilon > 0$ that we choose later.

Claim 1: With these prices the only user equilibrium is to have both users join and share on platform 2.

We next list all candidate user equilibria and show both users joining and sharing on platform 2 is the only user equilibria.

1. No user joins platform 2: this is not a user equilibrium because user 1 has a profitable deviation to join and share on platform 2. In particular, her payoff from this deviation is larger than her her payoff if she stays on platform 1, i.e.,

$$\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1\}) + c(1) \overset{(a)}{\geq} v_1\mathcal{I}_1(\{1\}) + c(2) - c(1) + \epsilon - v_1\mathcal{I}_1(\{1\}) + c(1) > c(2),$$

where (a) follows from (B-36).

2. User 1 joins platform 2: first note that the price given in (B-36) guarantees that user 1 shares on platform 2 if she joins because $\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1\}) \geq c(2) - c(1) > 0$. We next show that user 2 has a profitable deviation to join and share on platform 2. This is because her payoff if she joins and shares on platform 2 becomes larger than her payoff if she stays on platform 1, i.e.,

$$\tilde{p}_2^2 - v_2\mathcal{I}_2(\{1,2\}) + c(2) \overset{(a)}{\geq} v_2\mathcal{I}_2(\{1,2\}) + c(1) - c(2) + \epsilon - v_2\mathcal{I}_2(\{1,2\}) + c(2) = c(1) + \epsilon,$$

where (a) follows from (B-37). User 2's payoff if she joins and shares on platform 2 becomes larger than her payoff if she joins platform 2 and not shares, i.e.,

$$\begin{aligned}\tilde{p}_2^2 - v_2\mathcal{I}_2(\{1,2\}) + c(2) &\overset{(a)}{\geq} v_2(\mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\})) + \epsilon - v_2\mathcal{I}_2(\{1,2\}) + c(2) \\ &= -v_2\mathcal{I}_2(\{1\}) + c(2) + \epsilon.\end{aligned}$$

where (a) follows from (B-37).

3. User 2 joins platform 2: There are two possible cases depending on whether user 2 shares on platform 2 or not.

   - User 2 does not share on platform 2: user 1 has a profitable deviation from joining and sharing on platform 2 because her payoff after this deviation is larger than her payoff if she stays on platform 1:

$$\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1\}) + c(2) \overset{(a)}{\geq} v_1\mathcal{I}_1(\{1\}) + c(2) - c(1) + \epsilon - v_1\mathcal{I}_1(\{1\}) + c(2)$$
$$= \epsilon + 2c(2) - c(1) > c(1),$$

where (a) follows from (B-36).

- User 2 does not share on platform 2: we show that user 1 has a profitable deviation to join and share on platform 2. In particular, the payoff of user 1 if she joins and shares on platform 2 is larger than her payoff if she stays on platform 1 because

$$\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1,2\}) + c(2) \overset{(a)}{\geq} v_1\mathcal{I}_1(\{1,2\}) + c(1) - c(2) + \epsilon - v_1\mathcal{I}_1(\{1,2\}) + c(2)$$
$$= \epsilon + c(1) > c(1),$$

where (a) follows from (B-36). Payoff of user 1 if she joins and shares on platform 2 is also higher than her payoff if she joins platform 2 and does not share because

$$\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1,2\}) + c(2) \overset{(a)}{\geq} v_1(\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{2\})) + \epsilon - v_1\mathcal{I}_1(\{1,2\}) + c(2)$$
$$= -v_1\mathcal{I}_1(\{2\}) + \epsilon + c(2) > -v_1\mathcal{I}_1(\{2\}) + c(2).$$

4. Both users join platform 2: we show both users sharing on platform 2 is an equilibrium. In particular, for user 1 her payoff if she joins and shares on platform 2 is larger than her payoff if she joins and not shares on platform 2 and larger than her payoff if she stays on platform 1 because $\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1,2\}) + c(2) \geq \max\{-v_1\mathcal{I}_1(\{2\}) + c(2), c(1)\}$, where this inequality follows from (B-36). Also, for user 2 her payoff if she joins and shares on platform 2 is larger than her payoff if she joins and not shares on platform 2 and larger than her payoff if she stays on platform 1 because $\tilde{p}_2^2 - v_2\mathcal{I}_2(\{1,2\}) + c(2) \geq \max\{-v_2\mathcal{I}_2(\{1\}) + c(2), c(1)\}$, where this inequality follows from (B-37). This completes the proof Claim 1.

Claim 2: The prices specified above is a profitable deviation for platform 2 for sufficiently small v_1, v_2, and $c(2) - c(1)$. We prove that the payoff of platform 2 is strictly positive for $v_1 = v_2 = 1$. This establishes for sufficiently small v_1, v_2 the payoff of platform 2 is positive.

We let $c(2) - c(1) = \Delta$. Note that from submodularity of leaked information, we can simplify the price $\tilde{p}_1^2$ as

$$\tilde{p}_1^2 = \max\{\mathcal{I}_1(\{1\}) + c(2) - c(1), \mathcal{I}_1(\{1,2\}) + c(1) - c(2), (\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{2\}))\} + \epsilon$$
$$= \max\{\mathcal{I}_1(\{1\}) + c(2) - c(1), \mathcal{I}_1(\{1,2\}) + c(1) - c(2)\} + \epsilon = \max\{\mathcal{I}_1(\{1\}) + \Delta, \mathcal{I}_1(\{1,2\}) - \Delta\} + \epsilon$$

We can also simplify the price $\tilde{p}_2^2$ as

$$\begin{aligned}\tilde{p}_2^2 &= \max\{\mathcal{I}_2(\{1,2\}) - c(2) + c(1), (\mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\}))\} + \epsilon \\ &= \max\{\mathcal{I}_2(\{1,2\}) - \Delta, \mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\})\} + \epsilon.\end{aligned}$$

We next show that for

$$\Delta \leq \min\{\mathcal{I}_2(\{1\}), \frac{\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{1\})}{2}\}, \tag{B-38}$$

the payoff of platform 2 is strictly positive:

$$\begin{aligned}&\mathcal{I}_1(\{1,2\}) + \mathcal{I}_2(\{1,2\}) - \tilde{p}_1^2 - \tilde{p}_2^2 = \mathcal{I}_1(\{1,2\}) + \mathcal{I}_2(\{1,2\}) \\ &- \max\{\mathcal{I}_1(\{1\}) + \Delta, \mathcal{I}_1(\{1,2\}) - \Delta\} - \epsilon - \max\{\mathcal{I}_2(\{1,2\}) - \Delta, \mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\})\} - \epsilon \\ &\overset{(a)}{=} \mathcal{I}_1(\{1,2\}) + \mathcal{I}_2(\{1,2\}) - (\mathcal{I}_1(\{1,2\}) - \Delta) - \epsilon - \max\{\mathcal{I}_2(\{1,2\}) - \Delta, \mathcal{I}_2(\{1,2\}) - \mathcal{I}_2(\{1\})\} - \epsilon \\ &\overset{(b)}{=} \mathcal{I}_1(\{1,2\}) + \mathcal{I}_2(\{1,2\}) - (\mathcal{I}_1(\{1,2\}) - \Delta) - (\mathcal{I}_2(\{1,2\}) - \Delta) - 2\epsilon = 2\Delta - 2\epsilon \overset{(c)}{>} 0,\end{aligned}$$

where (a) follows form (B-38) and in particular inequality $\Delta \leq \frac{\mathcal{I}_1(\{1,2\}) - \mathcal{I}_1(\{1\})}{2}$, (b) again follows from (B-38) and in particular inequality $\Delta \leq \mathcal{I}_2(\{1\})$, and (c) follows by choosing $\epsilon < \Delta$. ■

Proof of the statement of Example B-2: Note that the first best is to have all users join one of the platforms, e.g., platform 1, and none of them share.

We show that both platforms offering zero prices to all users is an equilibrium. Suppose, the contrary, i.e., one of the platforms have a profitable deviation. We next show this cannot happen. Consider platform 1 and suppose it has a profitable deviation. Note that the only way for this platform to have a positive payoff is to offer prices $\tilde{p}_1^2$ and $\tilde{p}_2^2$ to get both users 1 and 2 join and share on it (since the value of all other users is very high). One user equilibrium is to have users 1 and 2 share on platform 2 and all other users join platform 1. For this to be a user equilibrium the payoff of a user $i \in \{1,2\}$ on platform 2 must be higher than her payoff on platform 1, i.e., $\tilde{p}_i^2 - v_1\mathcal{I}_i(\{1,2\}) + c(2) \geq c(n-1)$, for $i = 1,2$. Since $n \geq 3$, this inequality implies that $\tilde{p}_i^2 \geq v_1\mathcal{I}_i(\{1,2\})$ for $i = 1,2$. Therefore, the payoff of platform 2 is upper bounded by $\sum_{i=1,2}\mathcal{I}_i(\{1,2\})(1-v_i)$, which is strictly negative given that $v_1, v_2 > 1$.

Similarly, platform 1 cannot have a profitable deviation. Because the only candidate for a profitable deviations is to have both users 1 and 2 share on this platform. In one of the user equilibria of this form, all the other users join platform 2. Therefore, an identical argument to the previous case shows that platform 1 does not have a profitable deviation. ■

Proof of the statement of Example B-3: The first best is to have both users join platform 1 and none of them share because $v_2 > \max\left\{1 + \frac{\mathcal{I}_1(\{1\})}{\mathcal{I}_2(\{1\})}, \frac{\delta}{\mathcal{I}_2(\{1\})}\right\}$ guarantees that the social surplus of this allocation is larger than the allocation in which both users join platform 1 and user 1 shares and the allocation in which the two users join different platforms and user 1 shares, i.e.,

$$4\delta \geq \max\{4\delta + (1-v_1)\mathcal{I}_1(\{1\}) + (1-v_2)\mathcal{I}_2(\{1\}), 2\delta + (1-v_1)\mathcal{I}_1(\{1\})\}.$$

- If $\delta < (1-v_1)\mathcal{I}_1(\{1\})$, the first best is not equilibrium. In particular, we show that platform 2 has a profitable deviation by offering a price to get user 1 join and share on it. In particular, offering price $\tilde{p}_1^2 = \mathcal{I}_1(\{1\}) - \epsilon$ for $\epsilon < (1-v_1)\mathcal{I}_1(\{1\}) - \delta$ is a profitable deviation for platform 2. This is because, with this price the only user equilibrium is to have user 1 join and share on platform 2 and user 2 join and not share on platform 1. In this case, user 1 deviates and shares on platform 2 because it generates a higher payoff than joining platform 1, i.e., $\mathcal{I}_1(\{1\}) - \epsilon - v_1\mathcal{I}_1(\{1\}) + \delta > 2\delta$. Moreover, since user 1 shares on platform 1, user 2 prefers to join platform 1 but not share her data because $2\delta - v_2\mathcal{I}_2(\{1\}) < \delta$. Finally note that the payoff of platform 2 after this deviation increases from 0 to $\epsilon > 0$.

- If $\delta \geq (1-v_1)\mathcal{I}_1(\{1\})$, then the first best is an equilibrium. We show that the first best can be supported as an equilibrium with zero prices. First, note that, zero prices, both users joining platform 1 and none of them sharing is a user equilibria. We next show that neither platform has a profitable deviation. Consider platform 2. The only candidate for a profitable deviation is to have user 1 sharing on it (in which case user 2 joins platform 1 because $\delta < v_2\mathcal{I}_2(\{1\})$). This is not a profitable deviation for platform 2 because to incentivize user 1 to share on platform 1, the price offered to her, $\tilde{p}_1^2$, must satisfy $\tilde{p}_1^2 - v_1\mathcal{I}_1(\{1\}) + \delta \geq 2\delta$. This price generates a nonpositive payoff for platform 2 because $\mathcal{I}_1(\{1\}) - \tilde{p}_1^2 \leq (1-v_1)\mathcal{I}_1(\{1\}) - \delta \leq 0$. Similarly, the only possible profitable deviation for platform 1 is to induce user 1 to share her data, but this cannot generate a strictly positive payoff for platform and one. ■