Social Data Trading is Privacy Welfare Damaging

Ranjan Pal, *Member, IEEE,* Yixuan Wang, *Student Member, IEEE,* Junhui Li, *Student Member, IEEE,*
Mingyan Liu, *Fellow, IEEE,* Jon Crowcroft, *Fellow, IEEE,*
Yong Li, *Senior Member, IEEE,* Sasu Tarkoma, *Senior Member, IEEE*

***Abstract*—This paper argues that private but correlated data of individuals (such as those on a social networking platform) are under-priced in general, and that market competition for such data (such as that traded in a duopoly setting) does not alleviate this issue, resulting in diminishing economic utilitarian welfare. This is in stark contrast to the commonly held intuition that increased amount of end-user data in a market improves its efficiency or social welfare. The main reason behind this inefficiency lies in negative externality, in the form of privacy loss to one user as a result of the disclosure of another user's data, caused by the statistical correlation between their data, which is very common among community users. This externality is not sufficiently internalized at market equilibrium as reflected by the under-pricing of the data, thereby leading to damaged social welfare. Our mathematical model (a) provides regulatory insights on user information management for community social platforms, and (b) paves the way for a future general theory of community data trading in n-platform oligopoly markets. To the best of our knowledge, this is the first work of its kind in relation to assessing via theory the economically inefficient nature of community data trading ventures.**

***Index Terms*—community data, duopoly trade, social welfare**

I. Introduction

Data of billions of online individuals are currently gathered, processed, and analyzed for personalized advertising or other online service[1]. This trend is on the rise, both fueling and fueled by an increase in online apps, IoT technologies, and advanced artificial intelligence (AI) and machine learning (ML) methodologies. It is a widely accepted notion in economics (see [1] [2] [3] [4] [5] [6] [7] [8] [9]) that sharing individual information with the demand side of an information market is beneficial to targeted customization, demand side profit, and the growth of data-'hungry' AI/ML controlled businesses. It has also been argued by economists [10] [11] that because of these benefits that individual data brings to a market, a competitive market mechanism might generate too little data sharing from the supply side.

In this paper, we first show through a counter-example (Section II) that this popular economic intuition *does not hold* in general. Specifically, when two individuals' data are highly correlated, then the first selling his/her data would effectively render the second's data worthless to the buyer, regardless of the second user's own privacy valuation. This creates a type of "race to the bottom" pricing the buyer is able to extract. Depending on how the sellers value their data, this could be extremely damaging to the total welfare. The intuition behind this example lies in the negative externalities created by trading statistically correlated end-user data when users have heterogeneous privacy valuations of their data. This example also adds another facet to an emerging intuition that privacy is detrimental to information market efficiency [12] [5] [9]: we show that even when privacy preservation is not required, information markets can nonetheless be inefficient in a community setting where user data tends to be highly correlated in a statistical fashion.

We then present a mathematical model (Section IV) for the trading of such private but correlated data in a duopoly setting. The intention is to examine whether market competition can alleviate the inefficiency caused by the negative data externality. *Our conclusion is a negative one: at market equilibrium the externality cannot be sufficiently internalized and as a result, end-user privacy is jeopardized in such a community setting.* We analyze this market model (Section V) under a variety of market conditions. We first consider trading environments where negative data externalities are only among end-users of the same platform. We then extend this to environments where the data externalities exist across the two platforms. Numerical results are presented in VI.

As a special case particularly relevant to social community platforms, we also investigate platform duopoly markets (Section VII) where prices are pre-set (i.e., posted prices) so as to attract a larger client base. These markets extend those considered in Section V where platform pricing occurs post user lock-in. *As a primary result common to the analyses in Sections V and VII, we formally prove that community data trading markets decreases societal welfare.*

We discuss related work in Section III, and conclude the paper in Section IX.

II. Intuition

In this section, we provide an example-driven intuition that leads us to formally investigate the invalidity (or otherwise) of the popular economic notion that end-user information promotes their efficient trading markets.

Our intuition is based on the widely popular *Cambridge Analytica* scandal, where the Cambridge Analytica company created an app for mapping personality traits called *This is your digital life*; launched it on Facebook; and acquired private information of millions (approximately 50 million) of

R. Pal. M. Liu, J. Li, and Y. Wang are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA. E-mail: {palr, mingyan, opheelia, joywyx}@umich.edu

Y. Li is with the Department of Electronic Engineering, Tsinghua University, China, E-mail: liyong07@tsinghua.edu.cn

J. Crowcroft is with the Computer Laboratory, University of Cambridge, UK, and the Alan Turing Institute, UK, E-mail: jac22@cam.ac.uk

S. Tarkoma is with the Department of Computer Science, University of Helsinki, Finland, E-mail: sasu.tarkoma@helsinki.fi

[1]Facebook itself has approximately 2.5 billion monthly active individual users.

individuals from data shared by 270,000 Facebook users who voluntarily downloaded this app. The app accessed users' news feed, timeline, posts, and messages, and revealed information about other Facebook users. Cambridge Analytica then deployed such mass information for designing personalized political messages and advertising in the Brexit referendum and the 2016 US presidential election.

This scandal highlighted two important facets: (i) private information (e.g., behavior, habits, preferences) of users part of an online social community such as Facebook are correlated and results in knowing such information about other users[2] whose data is not leaked, and (ii) once it is openly publicized that valuable user information has been breached to satisfy external objectives, users are often miffed resulting in a huge social uproar as did happen in the case of the Cambridge Analytica scandal. These observations motivated us to develop a skepticism regarding the popular economic notion that more data implies increased information market efficiency. It could also be that trading in return for incentives for such community settings might not go down well[3] with the user privacy preferences, consequently hampering societal welfare (see Section IV for a definition).

To state our intuition in a relatively more formal manner, consider a community platform with two users, $i = 1, 2$. Each user owns its own personal data, which we represent with a random variable X_i (from the viewpoint of the platform). The personal data of the two users are related, which we statistically capture by assuming that random variables that represent such data are jointly normally distributed with mean zero and a correlation coefficient ρ. The community platform can acquire or buy the data of a user for "optimizing" their commercial interests through better estimating the user's preferences or actions. For example, an objective for the platform can be to minimize the mean square error of its estimates of user types, or maximize the amount of leaked information about them. Suppose that the intrinsic valuation of the platform for the users' leaked information is 1[4], while the value that the first user attaches to its privacy, again in terms of leaked information about it, is $\frac{1}{2}$ 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 its data (because its valuation of privacy, $\frac{1}{2}$, 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 $\rho \simeq 1$, and the fact that this is common knowledge, a high privacy-sensitive but perfectly rational user gains nothing in not selling his/her data - moreover, his/her data is automatically inferred through user 1. User 2 is better off selling her data for a benefit, even if that means compromising with his/her privacy - that would likely be anyway compromised via inference from user 1's data. Using a game-theoretic argument, if the buyer knows of this mindset (that both users would prefer to sell their data given common knowledge about correlations), then he would offer to buy the data at the cheapest possible price (approximately 0 given $\rho \simeq 1$), as it still generates positive profit for the sellers. Therefore in this simple example, the community platform buyer 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, despite data externalities changing the distribution of economic surplus between the platform and users. However, it can be arbitrarily inefficient when v is sufficiently high, simply by selling its data, the first user is creating a negative externality on the second user.

[2]Habits and preferences of a highly educated gay from a particular locality is informative about others with the same profile residing in the same area.

[3]This is a high chance in scenarios of social uproar post publicly known data breaches, if not in cases where data breaches go un-noticed.

[4]It is usual in practice to associate this parametric value with a "function" that characterizes the importance of information in the platform's commercial prospects. This parameter then is multiplied with the quality of user data (see Section IV for details) to arrive at a scalar value characterizing the monetary valuation of user data to the platform.

III. Related Work

In this section, we briefly review related literature pertaining to the existence and design of information trading markets. We identified two strands of research in this context: one rooted in the economics literature, and the other rooted in the technological literature on privacy-aware mechanism design.

The vision and benefits for information trading had their roots in arguments made in the 1970s by Chicago school economists, Posner [1][2] and Stigler [3], in favor of having increased social welfare. In later years, their arguments were upvoted by information economists such as Laudon [4] and Acquisiti [5] Varian [13], Odlyzko [6], Schwarz [8], and Samuelson [7][14][15]. The primary thesis of these scholars being that the lack of use of personal client data will lead to opportunity costs and market inefficiencies (sub-optimal states of economic social welfare) since it conceals potentially relevant information from other economic agents that eventually hamper the profitability of these agents. In contrast to the Chicago-school views, a number of economists including Hirshleifer [16][17], Burke [18], Wagman [19], Daughety & Reinganum [20], and Spence [21] are of the opinion that the costs to the demand side of the market to acquire quality client information may outweigh its social benefit, thereby decreasing social welfare.

According to Varian [13], Odlyzko [6], and Acquisiti [5], consumer data obtained (with or without consent) can have negative effects on society simply because post transaction the consumers have little knowledge or control over how and by whom their personal data will later be used. The firm (e.g., ad-networks) may sell the consumer's data to third parties (e.g., advertisers), which may lead to spam and adverse price discrimination, among other concerns, and subsequently lead to consensual consumers opting out of trade in future. Regulation here can curb the adverse effects of these negative externalities arising from trading and significantly contribute to welfare efficient and complete markets (where supply equals demand) [22][23]. Examples of practical ways to implement regulations suggested in existing literature include legislative property rights on consumer personal data shared between

the supply and demand side [4], and technical metrics (e.g., differential privacy) being adopted by demand side data intermediaries such as ad-networks to keep a check on the degree of privacy breach [24]. There have also been recent efforts that characterize cross-user negative externalities such as ours, with respect to trade of user information with a platform. Notable among them are the works by the authors in [25][26][27][28]. However, none of these works (unlike ours) are explicit in their mathematical characterization of data externalities. They simply assume that social surplus (welfare) negatively depends on the number of users on a platform (as more users imply greater chances of increased negative externality due to data breaches). From a non-privacy perspective, De Corni et al. [29] state that targeted advertising in markets driven by the presence of private non-perturbed consumer information can lead to higher equilibrium prices. This result is in line with Levin and Milgrom [30], Bergemann and Bonatti [31], and Cowan [32] who design markets improving match quality by disclosing consumer information to downstream firms. The common takeaway message from these papers is that such markets might be too costly to a data intermediary — because of the informational rent that is passed on to selling firms. We add to this line of result in our work stating that the information rent to be passed on to selling firms may not be high if the selling information is related to a community with high statistical correlation between consumer information. Research related to pricing information in a market setting has been proposed in [33][34][35][36][28][37][38]. However, none of these works mathematically address the notion of negative data externalities related to end-user information trading, and their impact on market performance, like we do.

Most existing technological works on privacy-aware mechanism design [39][40][41][42][43][44][45] [44][43][46][47][45] [48] [49] assume that there is a trusted data holder (e.g., platform, app) via the *trusted curator model* in differential privacy literature. The private data, i.e., personal consumer information, is either already kept safely to itself by the data holder, or is evoked using mechanisms that are designed with the aim of consensual truthfulness - i.e., What the data holder purchases is the "right" of using consumers' data in an announced way. The basic principle underlying these mentioned works is that the selling side (analogous to the end-user in our work) bid their costs of privacy loss caused by data usage by the buying side (the social community platform in our work), where each seller's privacy cost is modeled as a linear function of ϵ if its sold data is used in an ϵ-differentially private manner. The goal of the mechanism design here is to evoke truthful bids of seller cost functions. A major direction in which our work differs from this line of existing work is in considering (a) the trading market is not privacy-aware, i.e., trading prices is not a function of user privacy guarantees, and (b) that data holders (e.g., SNSs) though not trusted by consumers to keep their data private obeying consumer requirements, and may release it to agencies like ad-networks in return for benefits - however, the data holders pay the end-user for their data. Moreover, unlike us, the above-mentioned trading mechanisms do not mathematically incorporate statistical correlations between the selling users' data, and consequently does not capture negative externalities of possible data breaches. Recent efforts [50][51] have captured the role of privacy externalities in assessing market efficiency - however, they do not specifically model correlations among sold data by various sellers. In addition, they (in addition to all the aforementioned efforts) do not address market competition among data buyers, like we do. As a practical use-case of modeling privacy trading, the authors in [52] design a privacy trading mechanism for commercializing location privacy in mobile crowdsensing services. More specifically, they propose an auction-theoretic framework between workers and a single platform to trade location privacy data, given a differential privacy induced leakage budget. However, though they are similar in nature to our motivation in trading end-user information, we do not consider trading privacy leakage in our work. In addition, there is a significant fundamental difference between their contribution and ours: we formally model a duopoly market competition between established buyer firms being served by multiple data sellers; in contrast, the players (workers) in [52] are mobile end users distributed in a geographical locality thereby only interacting with a single platform through an auction, and not traditionally competing in a market.

IV. System Model

A simple example, such as the one aforementioned, clearly provides an intuition regarding the inefficiency of information trading in community settings with heterogeneous privacy valuations. In this section, en route to generalizing the validity (or invalidity) of our intuition, we propose a duopoly information trading market model consisting of n users split between two profit-maximizing community platforms (e.g., SNSs). However, before directly jumping to formalizing market competition, we first formalize (for the ease of exposition) the basic principle of trading operation between data suppliers and a single data buyer (platform) in a non-competitive setting, that remains invariant in a duopoly setting. Subsequently, we propose our duopoly market model to capture the strategy behind splitting a group of n end-users between two community platforms before trading operation begins and present a novel notion of market equilibrium for such a model.

A. Basic Principle of Operation Behind Trading

We consider n community 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 $\Sigma \in \mathbf{R}^{n \times n}$ 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 its type X_i, i.e., the type being the 'DNA' that drives the user's tastes (for example, based on its 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[5], i.e., $Z_i \sim \mathcal{N}(0, 1)$. For any

[5]This has taken various forms in the information privacy literature [53]. In practice Z_i can represent the perturbation output of the Local Differential Privacy (LDP) model used by Google and Apple [54]

user joining the community platform, the platform can derive additional revenue (e.g., due to benefits of targeted advertising) if it can predict the user's type. We simply assume that the community platform's revenue from each user is a deceasing 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. More specifically, the objective of the platform is to minimize

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

where $\mathbf{S}$ is the vector of data the platform acquires, $\hat{x}_i(\mathbf{S})$ 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 (be it explicit or implicit) to user i from the platform for their data (we ignore for simplicity any other transaction costs incurred by the platform).

Users value their privacy, which we also model in a reduced-form manner (reflecting both pecuniary and non-pecuniary[6] motives) as a function of the same mean square error.

We assume, specifically, that user i' s value of privacy is a parameter $v_i \geq 0$, and its payoff is

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

This expression and its comparison with objective (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 normalization, which ensures that if the platform acquires no additional information about the user and makes no payment to it, the payoff is zero. More specifically, the payoff for a user is the price it gets from the buyer for its data subtracted from the product of the intrinsic valuation of its personal data with the 'quality' of its inference (denoted through the MSE) by the buyer. We adopt this simple linear form of the payoff for analytical tractability. Clearly, 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 settings 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. An example of an environment without data externalities being collection agencies not gathering addresses locations. A simple market mechanism based on prices for data can implement this efficient outcome, in accordance to the traditional economic notion that more information implies better market efficiency. However, the situation could be very different in the presence of data externalities (e.g., online SN environments).

A key notion for our analysis is breached 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 $\mathbb{E}\left[\left(\hat{x}_i-X_i\right)^2\right]=\sigma_i^2$. As the platform receives data from this and other users, its estimate improves and the mean square error declines. The notion of breached information captures this reduction in mean square error (MSE). 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 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 breached information of (or about) user $i \in \mathcal{V}$ is the reduction in the MSE of the best estimator of the type of user i :

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

Notably, because of data externalities, breached information about user i depends not just on its 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 $p = (p_1, ..., p_n)$ as

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

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

We now characterize two important properties of the breached information function $\mathcal{I}_i : \{0,1\}^n \to \mathbf{R}$.

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

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

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

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

The monotonicity property states that as the set of community users who share their information expands, the breached 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 breached 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.

B. A Market Model for A Duopoly

Now that we have laid forth the basic principles of trading operation, in this section we introduce modeling elements of duopoly competition between social community platforms. The two main reasons to consider a duopoly instead of oligopoly are (i) simplicity, and (ii) in practice there are only a few major community platforms.

1) Game Induced by the Duopoly: Formally, we are dealing with a *three-stage game* in which first users strategically decide which platform to join (if any), then platforms simultaneously offer prices for data, and then finally all users

[6]As example, the fact that a user may receive a greater consumer surplus when the platform knows less about it or it may have a genuine demand for keeping its 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.

simultaneously decide whether to share their data. Note that the first step is followed by a popular two-stage Stackelberg game [55] setting for individual community platforms with their clients. Here, the end-users first observe the prices set by their platform, and then subsequently decide optimally whether to share their data with the former. The platform pre-assuming this optimal behavior on part of the end-users sets its optimal price vector in the first stage by solving a profit maximization problem with the optimal end-user behavior as constraints. Thus, we represent this three-stage strategic packaging as a novel game type and term it as an *embedded Stackelberg game* (ESG) - a *3-stage dynamic game* of perfect information between the community platforms and their end-users.

2) Game Essentials: 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 any of the platform, $b_i = 1$ means it 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. The payoff of a platform is a function of breached information about users, and the return payments to latter. So for platform $k \in \{1,2\}$, we have

$$U^{(k)}\left(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}\right) = \sum_{i \in J_k} \mathcal{I}_i\left(\mathbf{a}^{J_k}\right) - \sum_{i \in J_k : a_i^{J_k} = 1} p_i^{J_k} \quad (2)$$

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. In practice, it is true that users can join both platforms. However, our modeling decisions are mainly driven to capture the effect of user data (negative) externalities (due to statistical correlations between users and between platforms) that impact privacy. In our market model, we allow users to join neither platform as it allows extremely (privacy) risk-averse users to escape the fear of their data being traded for commercial purposes - no matter how sensitive their data is. We allow users to join either platform to capture between-platform externalities - that cannot be captured if the same user is in both the platforms. One could argue here that a user common to each platform would pose different externality effects on each of the platforms, but in our work the focus is not on how different the externalities are across platforms but whether the externalities will lead to an efficient or inefficient market economy. Choosing a model design where users can choose either platform (but not both) serves our purpose.

The payoff to a user has three parts. First, each user receives utility due to a valuable service it receives from the platform it joins. We term this "joining utility" and assume that it depends on who else joins the platform. Let $c_i\left(J_{b_i}\right)$ for user i joining platform b_i, denote the first part of the payoff, with the convention that $J_0 = \varnothing$. We also normalize $c_i(J) = 0$ for all $J \not\ni i$ and for all $i \in \mathcal{V}$. As is usual, as a second payoff component, the user incurs a disutility due to leaked information that negatively affects its privacy, and we again denote the value of privacy for user i by v_i. Finally, for the third part it receives benefits from any payments from the platform in return of the data it shares. Thus the net payoff user i accrues in joining platform $k \in \{1,2\}$ is written (conditioned on the value of a_i) as

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

The value of the payoff function is zero when the user chooses $b_i = 0$,. More specifically, when there is no data sharing decision a user's payment from the buyer is equal to zero - at the same time, the leaked information about the user is also equal to zero, and $c_i(\emptyset) = 0$. The sequence of steps in the duopoly market competition game is as follows.

1) Users are given a choice to simultaneously decide on their joining platform of choice (in the extreme case, they may choose to join no platform), i.e., $\mathbf{b} = \{b_i\}_{i \in \mathcal{V}}$, which determines J_1 and J_2.
2) Once J_1 and J_2 are determined from Step 1, the two platforms are given a choice to simultaneously offer price vectors $\mathbf{p}^{J_1}$ and $\mathbf{p}^{J_2}$.
3) The users choose their sharing decisions, $\mathbf{a} : \{0,1,2\}^n \times \mathbb{R}^n \to \{0,1\}^n$, once J_1 and J_2 are determined and the price vectors $\mathbf{p}^{J_1}$ and $\mathbf{p}^{J_2}$, are respectively set from Steps 1 and 2 above.

We assume that there is a common knowledge about Σ and the v_i's.

3) Game (ESG) Equilibrium Concept: We once again note that once the set of users joining a platform is determined in Step 1 above for the ESG, it is identical to a Stackelberg game. For such a game, for a given platform J_k (see Definition 1 for a paired platform setting), we denote the set of user equilibria at a given price vector $\mathbf{p}^{J_k}$ by $\mathcal{A}(\mathbf{p}^{J_k})$. A pair $\left(\mathbf{p}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}\right)$ of price and action vectors is a pure strategy *Stackelberg equilibrium* [55] if $\mathbf{a}^{\mathrm{E}} \in \mathcal{A}\left(\mathbf{p}^{\mathrm{E}}\right)$ and there is no profitable deviation for the platform, i.e.,

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

In addition, from the celebrated result due to Topkis [56], for any $\mathbf{p}$, the set $\mathcal{A}(\mathbf{p})$ is a complete lattice, and thus has a least and a greatest element. This implies that the set of user equilibria is always non-empty. Then, for the first stage of the ESG, we next define a *joining equilibrium*, as a profile of joining decisions anticipating the Stackelberg equilibrium from the second stage onward. More formally:

Definition 1: 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 $\mathbf{p}^{J_1,\mathbf{E}}$ and $\mathbf{p}^{J_2,\mathbf{E}}$ and action profiles $\left(\mathbf{a}^{J_1,\mathbf{E}}, \mathbf{a}^{J_2,\mathbf{E}}\right)$ such that $\mathbf{a}^{J_k,\mathbf{E}} \in \mathcal{A}\left(\mathbf{p}^{J_k,\mathbf{E}}\right)$ and $U^{(k)}\left(J_k, \mathbf{a}^{J_k,\mathrm{E}}, \mathbf{p}^{J_k,\mathrm{E}}\right) \geq U^{(k)}\left(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}\right)$, for all $\mathbf{p}^{J_k}$ and for all $\mathbf{a}^{J_k} \in \mathcal{A}\left(\mathbf{p}^{J_k}\right)$ 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 $i \in \mathcal{V}$

$$u_i\left(J_{b_i,}^{\mathrm{E}}, \mathbf{a}^{J_{b_i}^E,E}, \mathbf{p}^{J_{b_i}^E,E}\right) \geq u_i\left(J_k^{\mathrm{E}} \cup \{i\}, \mathbf{a}^{J_k^E \cup \{i\},E}, \mathbf{p}^{J_k^E \cup \{i\},E}\right),$$

for $k \neq b_i$ and

$$u_i\left(J^E_{b_i}, \mathbf{a}^{b_i,E}, \mathbf{p}^{J^E_b,E}\right) \geq 0, \forall i \in J^{\mathrm{E}}_{b_i}$$

Note that the first condition for the joining equilibrium ensures that each user prefers the platform she joins to the other platform. Given $J_0 = \emptyset$, the second condition makes sure that a user joining a platform receives non-negative payoff, since not joining either platform guarantees zero payoff.

V. Market Analysis

In this section, we analyse our proposed duopoly information trading market for characterizing market equilibria and their efficiency.

A. Characterizing Market Equilibria

Since our focus is on scenarios where users join community platforms and share their data, we impose that joining values are sufficiently large. More formally, we assume for each $i \in V$, 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) $c_i(\{i\}) > \max_{i\in\mathcal{V}} v_i\sigma_i^2$.

The first point in the assumption implies that users receive greater value-added services from a community platform when there are more users on the platform - a phenomena quite common to social networking platforms. The second point in the assumption imposes a pathological condition that even when there are no other users on a platform, the value of the services provided by the platform is greater than the cost of loss of privacy. This latter situation is a pathological extreme in theory (every commercial platform has quite a few users), put in place solely to derive important theoretical insights. Note here that the valuation can be either zero or positive (say the user gains some very specific utility in joining the platform even if no one joins), but the privacy loss is always less than the valuation as there is no correlation between users. This second aspect also directly yields the next lemma, which simplifies the rest of our analysis. Lemma 1 holding for this extreme pathological setting immediately implies its generality to the broader class of general settings. The proof of the lemma is in the online Appendix.

Lemma 1: Each user finds it incentive compatible to join one of the two community platforms. More precisely, $b_i = 1$ or 2 for all $i \in V$, at market equilibrium.

The next theorem characterizes the Stackelberg equilibrium of stages 2 and 3 of our ESG, given joining decisions.

Theorem 1: For a duopoly induced by an ESG, consider a joining profile $\mathbf{b}$ with the corresponding sets of users J_1 and J_2. Then a pure strategy Stackelberg equilibrium exists for stages 2 and 3 of the ESG, and satisfies

$$U^{(k)}\left(J_k, \mathbf{a}^{J_k,\mathbf{E}}, \mathbf{p}^{J_k,\mathbf{E}}\right) \geq U^{(k)}\left(J_k, \mathbf{a}^{J_k}, \mathbf{p}^{J_k}\right),$$
$$\forall \mathbf{p}^{J_k}, \mathbf{a}^{J_k} \in \mathcal{A}\left(\mathbf{p}^{J_k}\right), \text{and } k = 1, 2.$$

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

$$p_i^{J_k,\mathrm{E}} = v_i\left(\mathcal{I}_i\left(\mathbf{a}^{J_k,\mathbf{E}}\right) - \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_k,\mathbf{E}}\right)\right), \quad (3)$$

and

$$u_i\left(J_k^{\mathrm{E}}, \mathbf{a}^{J_k^{\mathrm{E}},\mathbf{E}}, \mathbf{p}^{J_k^{\mathrm{E}},\mathbf{E}}\right) = -v_i\mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_k,\mathrm{E}}\right) + c_i\left(J_k\right). \quad (4)$$

Implications - The implications of this theorem are indeed quite interesting. Although a pure strategy Stackelberg equilibrium always exists, a pure strategy joining equilibrium may not. The rationale behind this result lies in the negative data externalities generated through user data, which makes the latter sometimes go to the platform where less of their information will be breached. We emphasize here that there will be an improvement in social surplus via increasing the amount of breached information about *low-value users* (users with valuation v less than 1) - however, it will reduce the payoff for these users because it enables their platform to pay less for their data. As a result low-value users may prefer to choose a platform with fewer other low-value users to have less information breached about themselves.

Despite the non-guarantee of a pure strategy joining equilibrium of the ESG, we can show the perennial existence of a mixed strategy joining equilibrium. In order to do that we first resort to the following definition.

Definition 2: For any user $i \in \mathcal{V}$, 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 $\beta \in \Pi_{i\in\mathcal{V}}\mathcal{B}_i$ be a mixed strategy profile. Then β^{E} is a mixed strategy joining equilibrium if

$$u_i\left(\beta_i^{\mathrm{E}}, \beta_{-i}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}\right) \geq u_i\left(\beta_i, \beta_{-i}^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}, \mathbf{p}^{\mathrm{E}}\right), \quad \forall i \in \mathcal{V};\ \beta_i \in \mathcal{B}_i.$$

where

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

We now state the following result on the existence of a mixed-strategy equilibrium, the proof of which is in the online Appendix.

Theorem 2: There always exists a mixed strategy joining equilibrium of the duopoly characterized by an ESG in which all users join each community platform with probability $\frac{1}{2}$.

Implications - The theorem follows simply because each user will indifferently choose between the two platforms with probability $\frac{1}{2}$) at ESG equilibrium, while the other users choose one of the two platforms uniformly at random, at the ESG equilibrium. We also note that when the benefit from joining a more represented platform is sufficiently greater than the benefit from joining a relatively scarcely represented platform, a pure strategy equilibrium might exist in the ESG. As a matter of fact, if all users are of low-value, a potential game [57] results for which a pure strategy equilibrium exists.

B. Characterizing Market Inefficiency

The social surplus (SoS) of strategy profile $(\beta, \mathbf{a}, \mathbf{p})$ for the ESG, denoted as $\mathrm{Social\,Surplus}(\beta, \mathrm{a})$, is defined as

$$\mathbb{E}_{\mathrm{b}\sim\beta}\left[\sum_{i\in J_1}\left((1-v_i)I_i(\mathrm{a}^{J_1}) + c_i(J_1)\right) + \sum_{i\in J_2}\left((1-v_i)\mathcal{I}_i(\mathrm{a}^{J_2}) + c_i\left(J_2\right)\right),\right].$$

Here, the sets J_1 and J_2 are characterized as induced by random variable $\mathbf{b} \sim \beta$.

An intuitive hypothesis that naturally follows pertaining to our problem is that market inefficiencies can be alleviated using either (a) a pricing structure that internalizes the negative externalities that arise due to inference of user data from statistical correlations, or (b) enabling *high-value users* (users with valuation v greater than 1) to cluster to a platform where less of their information will be breached. The following theorem elucidates conditions under which such a hypothesis may hold true. The proof of the theorem is in the online Appendix.

Theorem 3: Let $\mathcal{V}^{(l)} = \{i \in \mathcal{V} : v_i \leq 1\}$ be the set of low-valued users, and $\mathcal{V}^{(h)} = \{i \in \mathcal{V} : v_i > 1\}$ be the set of high-valued users. Also denote by $\mathbf{v}^{(h)}$ and $\mathbf{v}^{(l)}$ to be the vectors of valuations of privacy for high-value and low-value users, respectively. Now

1) *Under the pathological assumption that the data for every high-value user in a social community is uncorrelated with all other users, the market equilibrium is efficient if and only if $c_i(\mathcal{V}) - c_i(\{i\}) \geq v_i \mathcal{I}_i\left(\mathcal{V}^{(l)} \backslash \{i\}\right)$ for all $i \in \mathcal{V}^{(l)}$.*
2) *Under the more realistic assumption that at least one high-value user is correlated (has a non-zero correlation coefficient) with a low-value user, there exists $\overline{\mathbf{v}} \in \mathbb{R}^{|\nu(h)|}$ and $\underline{\mathbf{v}} \in \mathbb{R}^{\left|\nu^{(l)}\right|}$ such that when $\mathbf{v}^{(h)} \geq \overline{\mathbf{v}}$ and $\mathbf{v}^{(l)} \geq \underline{\mathbf{v}}$, the market equilibrium is inefficient.*
3) *Finally, let us consider another pathological assumption where 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 market equilibrium is inefficient.*

Implications - The theorem provides the conditions under which market equilibrium in a duopoly community information trading setting is utilitarian welfare (in)efficient. According to the points in the theorem, it is evident that the information trading market is efficient is when high-value users are uncorrelated with all other users - something practically rare to achieve. Moreover, the condition for efficiency in the first part of the theorem is stringent due to additionally imposed restrictions on receiving direct benefits from joining platforms with different subsets of users. This is intuitive, given the fact that 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/breached to the platform by other low-value users. A common perception (from our model) is that it is socially beneficial (for data buying agencies) if information related to low-value users is breached - however, the price that low-value sellers get in return from the platforms at market equilibrium is too low to optimize social welfare overall. This fragmented allocation of users across platforms is costly if there are gains from being on larger community platforms. In all other cases, i.e, parts 2 (the more practical scenario) and 3 in the theorem, the information trading market is inefficient (SoS at equilibrium is not optimal). More specifically, part 2 of the theorem brings in privacy breach concerns as well, for the low-value users - which in a community population setting is very realistic. This part of the theorem rules out the possibility that low and high-value users go to separate platforms at market equilibrium, that may subsequently maximize social surplus even at the cost of the loss of direct benefits from forming a larger network on a single platform. However the former two-platform split is very hard to sustain in practice, especially when low-value users care about their privacy - simply because the latter will have an incentive to switch to the platform populated by high-value users where there will be no information leaked about them (as low correlations in their data with high-value users) and subsequently obtain higher prices for their data.

The next result provides an upper bound on social surplus, the proof of which is in the online Appendix.

Theorem 4: Let β^{E} be the uniform mixed joining strategy of the duopoly characterized by the ESG. Then

$$\mathrm{SoS}\left(\beta^{\mathrm{E}}, \mathbf{a}^{\mathrm{E}}\right) \leq \sum_{i \in \mathcal{V}^{(l)}} (1 - v_i)\, \mathcal{I}_i(\mathcal{V}) - \frac{1}{2} \sum_{i \in \mathcal{V}^{(h)}} (v_i - 1)\, \mathcal{I}_i\left(\mathcal{V}^{(l)}\right)$$

Implications - Theorem 4 implies that if

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

then it is beneficial to not have a market for information trading, i.e., having a market would lead to a negative social surplus. This happens when the value of privacy of high-value users is high and/or when the correlation between high-value and low-value users is large. This condition also 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 that the expected breach of user i's information from data sharing of low-value users is greater than 0.5 times the total breached information about this user with all the low-value users sharing their data on the same platform.

C. Capturing Externalities Between the Community Platforms

Till now, we have analyzed the duopoly information trading market accounting for negative externalities between users attached to a single platform. However, it is quite likely for externalities to percolate between platforms. In this subsection, we generalize our model to allow for between-platform externalities. This enables us to realize that a degree of privacy loss may indirectly occur for end-users who are not on the platform, but whose statistically correlated 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

$$u_i\left(J_k, a_i, \mathbf{a}_{-i}^{J_k}, \mathbf{p}^{J_k}\right) = \begin{cases} f_i^1\left(J_k, a_i = 1, \mathbf{a}_{-i}^{J_k}, \mathbf{p}^{J_k}\right), & k' \neq k \\ f_i^2\left(J_k, a_i = 0, \mathbf{a}_{-i}^{J_k}, \mathbf{p}^{J_k}\right) & k' \neq k \end{cases}$$

where

$$f_i^1(\cdot) = p_i^{J_k} - v_i \mathcal{I}_i\left(a_i = 1, \mathbf{a}_{-i}^{J_k}\right) - v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right) + c_i\left(J_k\right)$$

and

$$f_i^2(\cdot) = -v_i \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_k}\right) - v_i \alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right) + c_i\left(J_k\right).$$

Here, the term $\alpha \mathcal{I}_i\left(\mathbf{a}^{J_{k'}}\right)$ is breached information about user i available to the platform it has has not joined. The payoff of

users who have not joined either of the platforms is given by

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

$\alpha\in[0,1]$ in our analysis captures the externalities between the two platforms. When $\alpha=0$, we converge upon the model 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 in terms of payoff, the platforms benefit in an equal fashion from information breached about individuals who are not their customers. Subsequently, the the payoff of platform $k\in\{1,2\}$ is now

$$U^{(k)}\left(J_k,\mathbf{a}^{J_k},\mathbf{p}^{J_k}\right)=\sum_{i\in J_k}\mathcal{I}_i\left(\mathbf{a}^{J_k}\right)+\alpha\sum_{i\in J_{k'}}\mathcal{I}_i\left(\mathbf{a}^{J_k}\right)-\sum_{i\in J_k:a'_i=1}p_i, \tag{5}$$

where $k'\neq k$ and $\sum_{i\in J_V},\mathcal{I}_i\left(\mathbf{a}^{J_k}\right)$ is breached information about users on the other platform. We next show that our main results generalize to this case, the proof of which is in the online Appendix.

*Theorem 5: For any joining profile **b** with the corresponding sets of users J_1 and J_2, there exists a pure strategy Stackelberg equilibrium in the second stage of the duopoly characterized by the ESG, with market equilibrium prices as in equation (3) and equilibrium payoffs of users joining community platform k given by*

$$u_i\left(J_k^{\mathrm{E}},\mathbf{a}^{J_k^{\mathrm{E}},\mathbf{E}},\mathbf{p}^{J_k^{\mathrm{E}},\mathbf{E}}\right)=-v_i\mathcal{I}_i\left(0,\mathbf{a}_{-i}^{J_k,\mathrm{E}}\right)-v_i\alpha\mathcal{I}_i\left(\mathbf{a}^{J_{k'},\mathbf{E}}\right)+c_i\left(J_k\right) \tag{6}$$

In addition, for the first stage of the ESG, there always exists a mixed strategy joining equilibrium of the ESG, where all users join each platform with probability $\frac{1}{2}$.

Implications - We observe from the theorem that the equilibrium prices are the same as in equation (3) because the incremental effect of data sharing for a user is the same as before due to the fact that breached information is unaffected by the user's data sharing decision. However, now since the user's information is breached by the data shared by users on the other platform, its payoff is less than in equation (4).

The next theorem (see online Appendix for a proof) provides sufficient conditions for the inefficiency of the duopoly equilibrium in information trading markets in this case.

Theorem 6:

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 $c_i(\mathcal{V})-c_i(\{i\})\geq(1-\alpha)\mathcal{I}_i\left(\mathcal{V}^{(l)}\backslash\{i\}\right)\forall i\in\mathcal{V}^{(l)}$*
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 $\overline{\mathbf{v}}\in\mathbb{R}^{|\nu(h)|}$ such that for $\mathbf{v}^{(h)}\geq\overline{\mathbf{v}}$ 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.*

Implications - The basic implication of the theorem is the inefficiency of information trading markets in a community platform setting (similar to that in Theorem 3) with cross externalities. However, there are salient differences from Theorem 3. To start with, apart from the conditions in the first part of Theorem 3, we now additionally require the between/cross-platform spillovers not to be too large. The inability to satisfy this condition will result in the first best, i.e., social welfare optimal state, to involve splitting low-value users across the two platforms. Moreover, the conditions for inefficiency in the second part of Theorem 3 are slightly weaker because the cross-platform externalities increases the likelihood that the information of high-value users will be breached inefficiently.

The next result provides an upper bound on social surplus, the proof of which is in the online Appendix.

Theorem 7: Let β^{E} be the (uniform) mixed joining strategy of the market mechanism characterized by the ESG, at equilibrium. Then

$$SoS(\beta^{\mathrm{E}},\mathbf{a}^{\mathrm{E}})\leq(1+\alpha)\sum_{i\in\mathcal{V}(t)}(1-v_i)\mathcal{I}_i(\mathcal{V})$$

$$\frac{1+\alpha}{2}\sum_{i\in\mathcal{V}(h)}^{-}(v_i-1)\mathcal{I}_i\left(\mathcal{V}^{(l)}\right)$$

Implications - Theorem 7 implies that under the following condition,

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

not trading community data in a duopoly (competitive) market is once again beneficial, i.e., the social surplus can be negative if we do not (reasons similar to that in Theorem 4).

VI. Numerical Evaluation

In this section, we provide numerical examples to illustrate important practical market outcomes from our research. More specifically, we first show through *Example 1*, the role of negative externalities *to prevent the existence of a pure strategy joining equilibrium in the 3-stage ESG*, despite the existence of a pure strategy equilibrium in the Stackelberg game. We then illustrate via *Example 2* that in contrast to popular intuition, *market competition need not redress market inefficiencies* by either increasing or decreasing data prices, or allowing high-value users to go to a platform where there are less chances of their information being breached. Finally, via *Example 3*, we illustrate that at a mixed-strategy joining equilibrium of the ESG, the *market equilibrium entail negative social surplus.*

A. Illustrating Example 1

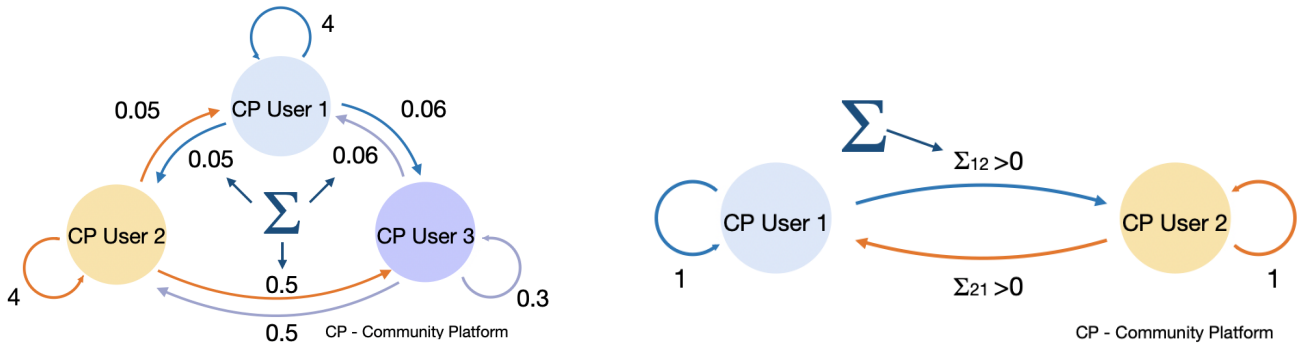
Consider a three-user covariance matrix Σ (see below, Figure 1a) characterizing the dependency between the data of the three users.

$$\Sigma=\begin{pmatrix}4 & .05 & .06\\ .05 & 4 & .5\\ .06 & .5 & .3\end{pmatrix},$$

and their intrinsic valuations for privacy given by

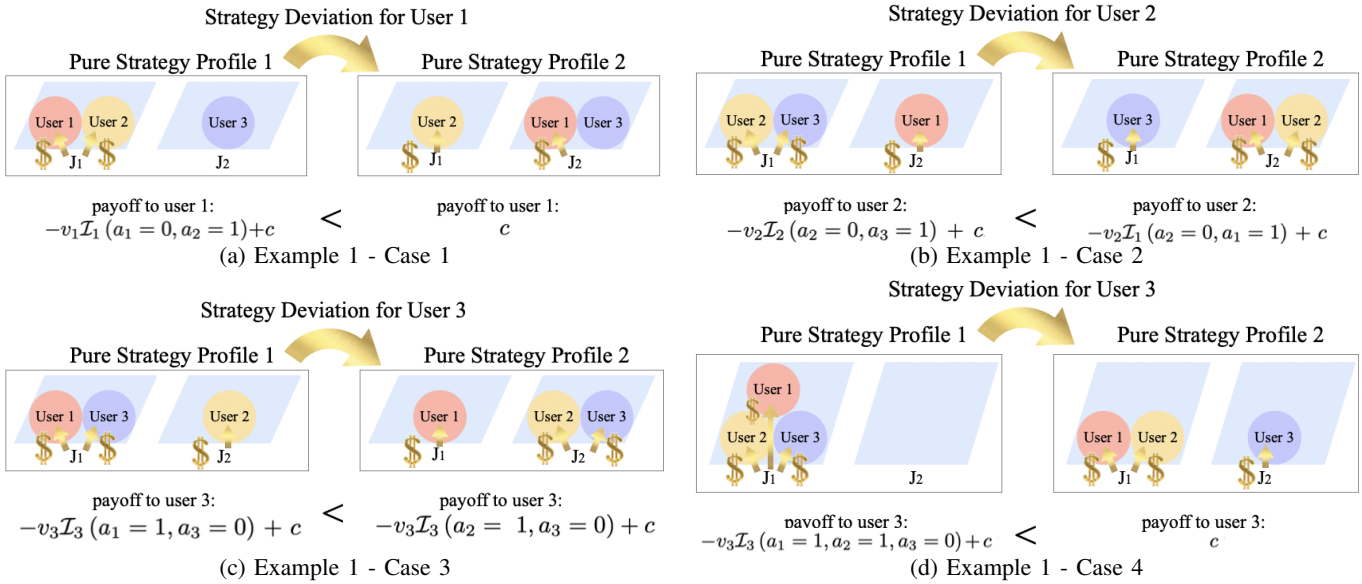
$$v_1=0.01,\quad v_2=0.99,\quad v_3=1.1.$$

We assume for the sake of simplicity that post joining a platform, all users receive the same benefit, i.e., c_i equals a



(a) Example 1 (b) Example 2

Fig. 1: A Figurative Illustration of The Covariance Matrix for Illustration Examples 1 and 2



(a) Example 1 - Case 1 (b) Example 1 - Case 2

(c) Example 1 - Case 3 (d) Example 1 - Case 4

Fig. 2: A Figurative Illustration of the Non-Equilibrium Nature of Possible Pure Strategy Profiles for Illustrative Example 1

constant c for all i. From the example covariance matrix and the intrinsic privacy valuations, we observe that user 3 has the highest value for privacy, but shares its data if it is on the same platform as user 2 (this is because user 2 data is quite correlated with that of user 3, and it is incentive compatible for user 3 to share with the platform despite putting a high intrinsic value on its privacy). However, either user 2 or user 3 will not find it incentive compatible to share their data if they are on the same platform as user 1 (due to the low covariance coefficient between their data and that of user 1). We next list the possible pure strategy joining profiles and show that none of them can be an equilibrium (see online Appendix *Detail-E1* for the derivation of the strategy closed forms).

1. **Pure Strategy Profile 1:** $J_1 = \{1, 2\}, J_2 = \{3\}$ - Figure 2a illustrates that this pure strategy joining profile is not an equilibrium, as it is beaten by the pure strategy joining profile on the right. In particular, from equation (3), the price of user 1's data will lead to a payoff of $-v_1\mathcal{I}_1\left(a_1 = 0, a_2 = 1\right)+c$ in this candidate equilibrium induced by joining profile $J_1 = \{1, 2\}, J_2 = \{3\}$. This implies that user 1 can profitably deviate to platform 2, which will result in the following candidate Stackelberg equilibrium profile $J_1 = \{2\}$ and $J_2 = \{1, 3\}$, that will give user 1 a payoff of c, verifying that the deviation is beneficial for the latter.
2. **Pure Strategy Profile 2:** $J_1 = \{2, 3\}, J_2 = \{1\}$ - With this joining profile, via a similar reasoning as for the case of Profile 1, at the resulting Stackelberg equilibrium, user 2's payoff is $-v_2\mathcal{I}_2\left(a_2 = 0, a_3 = 1\right) + c$. However, user 2 can profitably deviate to platform 2 (see Figure 2b), where it will receive a payoff of $-v_2\mathcal{I}_1\left(a_2 = 0, a_1 = 1\right)+c$, which exceeds its candidate equilibrium payoff induced by joining profile $J_1 = \{2, 3\}, J_2 = \{1\}$.
3. **Pure Strategy Profile 3:** $J_1 = \{1, 3\}, J_2 = \{2\}$ - With this joining profile, via a similar reasoning as for the case of Profiles 1 and 2, at the resulting Stackelberg equilibrium, user 3's payoff is $-v_3\mathcal{I}_3\left(a_1 = 1, a_3 = 0\right)+c$. However, user 3 can profitably deviate to platform 2 (see Figure 2c), where it will receive a greater payoff of $-v_3\mathcal{I}_3\left(a_2 = 1, a_3 = 0\right) + c$ when compared to that received from candidate equilibrium payoff induced by

joining profile $J_1 = \{1, 3\}, J_2 = \{2\}$.

4. **Pure Strategy Profile 4:** $J_1 = \{1, 2, 3\}, J_2 = \emptyset$ - With this joining profile, user 3 again has a profitable deviation and can increase its payoff from $-v_3\mathcal{I}_3\left(a_1 = 1, a_2 = 1, a_3 = 0\right) + c$ to c (see Figure 2d) by switching to platform 2. *The aforementioned joining profiles establish that there is no pure strategy equilibrium in this ESG characterizing a market duopoly.*

Our goal was to show a *single* existence of a duopoly market (induced via an ESG) scenario where there is no pure strategy market equilibrium. Thus our example above is mathematically sufficient to safely infer that ESGs pertaining to our problem setting need not necessarily have pure strategy equilibria.

B. Illustrating Example 2

In order to illustrate that competition need not reduce market inefficiencies in a community data economy, we consider a setting (see Figure 1b) 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. We first show that market competition improves user surplus, followed by the fact that it may also reduce it.

The Improvement Scenario - Consider a monopoly setting where $v_2 > 1$ is made sufficiently large for user 2 to share its data with user 1 at a market equilibrium. The monopoly equilibrium surplus under sharing decisions (derived at market equilibrium) $a_1^{\mathrm{E}} = 1$ and $a_2^{\mathrm{E}} = 0$ is

$$\begin{aligned}&(1 - v_1)\,\mathcal{I}_1\left(a_1 = 1, a_2 = 0\right)\\ &+ (1 - v_2)\,\mathcal{I}_2\left(a_1 = 1, a_2 = 0\right) + 2c.\end{aligned}$$

However, under market competition, the equilibrium surplus under joining decisions (derived at market equilibrium) $b_1^{\mathrm{E}} = 1, b_2^{\mathrm{E}} = 2$ and equilibrium data sharing decisions (also derived at market equilibrium) $a_1^{J_1,\mathrm{E}} = 1$ and $a_2^{J_2,\mathrm{E}} = 0$. is

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

which is *strictly greater than the equilibrium surplus under monopoly.*

The Non-Improvement Scenario - Consider a monopoly setting where $v_2 < 1$. The data sharing decisions at market equilibrium are given by $a_1^{\mathrm{E}} = 1$ and $a_2^{\mathrm{E}} = 1$, and the monopoly equilibrium surplus is

$$\begin{aligned}&(1 - v_1)\,\mathcal{I}_1\left(a_1 = 1, a_2 = 1\right)\\ &+ (1 - v_2)\,\mathcal{I}_2\left(a_1 = 1, a_2 = 1\right) + 2c.\end{aligned}$$

However, under market competition, the equilibrium surplus under joining decisions (derived at market equilibrium) $b_1^{\mathrm{E}} = 1, b_2^{\mathrm{E}} = 2, a_1^{J_1,\mathrm{E}} = 1$ and sharing decisions (also derived at market equilibrium) $a_2^{J_2,\mathrm{E}} = 1$ is

$$\begin{aligned}&(1 - v_1)\,\mathcal{I}_1\left(a_1 = 1, a_2 = 0\right)\\ &+ (1 - v_2)\,\mathcal{I}_2\left(a_1 = 0, a_2 = 1\right) + 2c,\end{aligned}$$

which is *strictly less than the surplus under monopoly.*

We again note here that our goal was to show *one* example that illustrates that market competition for user data among two social community platforms (as per our system model) might not redress inefficiency issues. Hence our example above is mathematically sufficient to safely infer that such a market competition is not always welfare beneficial.

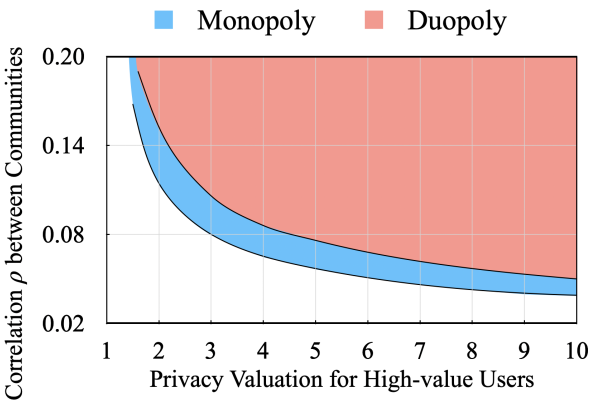


Fig. 3: Shaded area shows the pairs of (ρ, v_h) with negative equilibrium surplus in the monopoly and duopoly setting.

C. Illustrating Example 3

Intuitively, it is most likely to show the possibility of a negative social surplus in scenarios where the difference in the intrinsic privacy valuations of high-value and low-value users is not large (setting the stage for a large amount of negative externalities). Hence, without loss of generality, we consider a setting with two sets (say 1 and 2) of users, each of size 10, and two competing community platforms. We allow the users in each set to randomly choose a community platform of their choice. Assume that all users in set 1 are low-value and have a value of privacy equal to 0.9, while all users in set 2 are high-value (with $v_h > 1$). In the most conservative setting (acting as a barrier to generate a negative social surplus), we homogenize (a) user data to have a variance of 1, (b) any two users who belong to the same set to have a correlation coefficient of $\frac{1}{20}$, and (c) any two users who belong to different the correlation between any two users who belong to different sets to have a correlation coefficient of ρ. In Figure 3, we plot the equilibrium surplus as a function of the v_h and ρ variables for the monopoly and duopoly settings respectively. The plot for the monopoly setting (the lower curve) represents the set of points characterizing the combinations of these two variables (v_h and ρ) for which the social surplus is equal to zero. Moving in the northeast direction reduces equilibrium surplus and hence the shaded area has negative surplus. Consequently, there is an improvement in the utilitarian social welfare if data trading markets do not operate in the shaded area.

We also observe two important phenomena. First, relatively small values of the correlation coefficient ρ are sufficient for social surplus to be negative, indicating the relative ease with which running data markets can become welfare infeasible. Second, with v_h values getting 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. The northeast region of the top curve in Figure 3 represents the set of points characterizing the combinations of these two variables (v_h and ρ) for which the social surplus is negative, in the duopoly setting. The existence of this region follows from Theorem 4, when the correlation between high-value and low-value users is large. The derivation for closed form expressions pertaining to the example is detailed in online Appendix *Detail-E2*.

VII. Duopoly over Pre-Set Data Prices

The paper thus far has dealt with a duopoly market where users strategically decide to join a community platform based on an ESG analysis in which platform price setting is a strategic step. In practice, it could also be the case that the social platform announces prices in advance to attract end-users and these prices are not explicitly the outcome of a game analysis. This happens in current reality where the 'prices' are analogous to marketing benefits to attract consumer base, as done by platforms such as *Fabdog* (provide high-frequency retargeting ads), *Wayfair* (showcase their products and benefits with Facebook Carousel ads), *FabFitFun* (use shoppable pins on Pinterest for direct social platform click leads to shopping site), *Kohl's* (push workout videos through Facebook to attract customers), *Babyleggings* (offer Instagram discounts), etc. In the duopoly model analyzed so far in Section V, after the two platforms set data prices, users no longer had the option of switching to another platform. At the Stackelberg equilibrium, the platforms set prices by anticipating user choices and selected the most advantageous (pure strategy) user equilibrium for itself (when there were multiple user equilibria). This implies that at market equilibrium, the data price will ensure a (weakly) greater payoff for the platform than any other price for any other (pure strategy) user equilibrium. However, once users make their joining decisions after price offers, we will 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.

The exact sequence of events for such a market setting form a 2-stage dynamic game, and is given as follows.

1. Two community platforms advertise their price vectors $\mathbf{p}^1 \in \mathbf{R}^n$ and $\mathbf{p}^2 \in \mathbf{R}^n$ in the market.
2. Post the above advertisement, users simultaneously decide which platform to join (they may decide to join none of the community platforms), i.e., $\mathbf{b} = \{b_i\}_{i\in\mathcal{V}}$ (which determines J_1 and J_2) and also whether they would want to share their data, i.e., $\mathbf{a} = \{a_i\}_{i\in\mathcal{V}}$, with the platforms.

It is worth noting that due to the fiercer nature of competition between platforms over data and the number of users, converging to a pure strategy equilibrium is increasingly challenging for such a market. Our goal here is to investigate the existence of a mixed strategy for such market settings. It may seem obvious that every finite game has a mixed strategy Nash equilibrium (by Nash's celebrated theorem); however, only stage 2 is a finite game and will always have a mixed Nash equilibrium. *The challenge is show the existence of a mixed strategy equilibrium for the first stage game,* that is not a finite game.

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

$$u_i\left(\mathbf{a}, \mathbf{b}, \mathbf{p}^1, \mathbf{p}^2\right) = \begin{cases} p_i^{J_{b_i}} - v_i \mathcal{I}_i\left(a_i = 1, \mathbf{a}_{-i}^{J_{b_i}}\right) + c_i\left(J_{b_i}\right), \\ -v_i \mathcal{I}_i\left(a_i = 0, \mathbf{a}_{-i}^{J_{b_i}}\right) + c_i\left(J_{b_i}\right) \end{cases}$$

Here, a^{J_k} denotes the vector of sharing decisions in the set J_k for $k = 1, 2$. We ensure that every user joins one of the platforms, and this is ensured through the assumption set in Section V. The payoff of platform $k \in \{1, 2\}$ is

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

Definition 3: For given price vectors $\mathbf{p}^1 \in \mathbf{R}^n$ and $\mathbf{p}^2 \in \mathbf{R}^n$, the joining and sharing profiles $\mathbf{b}$ and $\mathbf{a}$ constitute a user equilibrium if for any $i \in \mathcal{V}$, we have

$$(a_i, b_i) \in argmax_{a,b} u_i\left(a, b, \mathbf{a}_{-i}, \mathbf{b}_{-i}, \mathbf{p}^1, \mathbf{p}^2\right)$$

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

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

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

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

We next define a mixed strategy equilibrium similarly to that in Section V. The equilibrium this time is formed of strategies that will be characterized by probability distributions over price vectors for the platforms and user actions.

Definition 4: Let $\mathcal{P}$ be the set of probability measures over $\mathbf{R}_+^n$. 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 $\mathbf{p}^1 \in \mathbf{R}^n$ and $\mathbf{p}^2 \in \mathbf{R}^n$, the joining and sharing profiles $\beta \in \Pi_{i\in\mathcal{V}}\mathcal{B}_i$ and $\alpha \in \Pi_{i\in\mathcal{V}}\mathcal{A}_i$ constitute a mixed user equilibrium if for any $i \in \mathcal{V}$, we have

$$u_i\left(\alpha, \beta, \mathbf{p}^1, \mathbf{p}^2\right) \geq u_i\left(\alpha_i', \beta_i', \alpha_{-i}, \beta_{-i}, \mathbf{p}^1, \mathbf{p}^2\right), \forall \alpha_i' \in \mathcal{A}_i;\ \beta_i' \in \mathcal{B}_i$$

where $u_i\left(\alpha_i, \beta_i, \alpha_{-i}, \beta_{-i}, \mathbf{p}^1, \mathbf{p}^2\right)$
=
$$\mathbb{E}_{a_i\sim\alpha_i, \mathbf{a}_{-i}\sim\alpha_{-i}, b_i\sim\beta_i, \mathbf{b}_{-i}\sim\beta_{-i}}\left[u_i\left(a_i, b_i, \mathbf{a}_{-i}, \mathbf{b}_{-i}, \mathbf{p}^1, \mathbf{p}^2\right)\right].$$

We let $\mathcal{A}\left(\pi^1, \pi^2\right)$ denote the set of mixed strategy user equilibria for given price strategies π^1 and π^2. Strategy price profiles $\pi^{k,\mathrm{E}} \in \mathcal{P}, k = 1, 2$, joining profile β^{E}, and sharing profile α^{E} constitute a mixed strategy equilibrium if $\left(\alpha^{\mathrm{E}}, \beta^{\mathrm{E}}\right) \in \mathcal{A}\left(\pi^{1,\mathrm{E}}, \pi^{2,\mathrm{E}}\right)$ and for any $\pi \in \mathcal{P}$ there exists $(\alpha, \beta) \in \mathcal{A}\left(\pi, \pi^{2,\mathrm{E}}\right)$ such that

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

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

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

We now have the following theorem that establishes the existence of the mixed strategy Nash equilibrium of the non-

finite duopoly market game with pre-set data prices. The proof of the theorem is in the online Appendix.

Theorem 8: There exists a mixed strategy equilibrium strategy to the duopoly market game with pre-set data prices.

Implications - The non-existence of a pure strategy equilibrium is due to the same reason as in the market scenario in Section V. However, this time, the existence is significantly more challenging because of fierce competition over data and users before the lock-in phase. The case of investigating the theoretical existence of an SPNE with pure $\mathbf{p}^1$, $\mathbf{p}^2$ and mixed a, b might be an interesting one, but in the interest of practical realization, a pure a, b is much desired at market equilibrium.

Practical Justification for a Pure Strategy - During the 1980s, the concept of mixed strategies came under heavy fire for being intuitively "problematic." Randomization, central in mixed strategies, lacks behavioral support. Seldom do people make their choices following a lottery. This behavioral problem is compounded by the cognitive difficulty that people are unable to generate random outcomes without the aid of a random or pseudo-random generator. In 1991, game theorist Ariel Rubinstein described alternative ways of understanding the concept. The first, due to Harsanyi in 1973, is called purification, and supposes that the mixed strategies interpretation merely reflects our lack of knowledge of the players' information and decision- making process. Apparently random choices are then seen as consequences of non-specified, payoff-irrelevant exogenous factors. However, it is unsatisfying to have results that hang on unspecified factors. Aumann and Brandenburger in 1995 re-interpreted Nash equilibrium as an equilibrium in beliefs, rather than actions. For instance, in the "rock-paper-scissors" game an equilibrium in beliefs would have each player believing the other was equally likely to play each strategy. This interpretation weakens the predictive power of Nash equilibrium, however, since it is possible in such an equilibrium for each player to actually play a pure strategy of Rock. Ever since, game theorists' attitude towards mixed strategies-based results have been ambivalent. *Mixed strategies are still widely used for their capacity to provide Nash equilibria in games where no equilibrium in pure strategies exist, but the model does not specify why and how players randomize their decisions.*

VIII. Tech-Policy Recommendations

Our research outcome raises the following recommendations in relation to trading data of users on social community platforms, assuming a realistic non-monopolistic platform setting:

- **Recommendation R1** - There should be a proper valuation of the supply-side data that is collected by the platforms, both from the buyer side and also from the seller side. It is quite likely that different data types will have different valuation-sensitivity tradeoffs, as perceived by both the buyer and the seller sides, and consequently should be traded in different tranches for improved market efficiency [51]. As an example, data types that are perceived to be less privacy-sensitive by community platform users should be traded in a different market than those that are deemed to be more privacy-sensitive. A failure to do this would result in significant negative externalities generated within a mixed market, that would be hard to internalize and will lead to low aggregate welfare in society.
- **Recommendation R2** - A proper valuation-sensitivity evaluation exercise, as mentioned in R1, should be necessitated via conducting social experiments in a global population of SNS users. One particular direction would be the use of institutional review board (IRB)-sanctioned randomized controlled trials (RCTs) [58] [59] [60] [61] that need to be conducted by the data buying platforms, prior to trading, to test the privacy-sensitivity of certain data attributes against different valuation amounts, with and without an intervention method (e.g., a privacy awareness program) applied to a significant population of global SNS users. Based on the outcome of the experiments, an appropriate regulated (privacy-preserving) trading model should be adopted to achieve appropriate privacy-utility tradeoffs. As an example, a market construct (such as the one proposed in [51]) based on the trusted curator differential privacy (DP) model can be adopted for low privacy-sensitive data, and a similar construct based on the local differential privacy (LDP) model [62] [63] [64] [65] can be adopted for higher privacy-sensitive data. For both these data types, the DP perturbation would assume different intervals for the respective data types.
- **Recommendation R3** - In case of the lack of use of privacy-preserving constructs to conduct trading markets (as in R2), 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 state, i.e., the state at which utilitarian social/privacy welfare is maximized. More interestingly, one could propose a scheme based on mediated-data sharing that can improve welfare. In particular, when equilibrium surplus is negative, shutting down data markets, for example with high uniform taxes on all data transactions, could improve welfare. But this prevents the sharing of the data of individuals with low value of privacy or high benefits from goods and services that depend on the platform accessing their data. One way to alleviate this problem is to first share individual data with a mediator which transforms them (like in a local differential privacy model) before revealing them to the platform, but also maintaining utility for the buyer. In this way the correlation of the data with the information of privacy-conscious users can be significantly eliminated, and this would improve welfare relative to the option of shutting off data markets altogether.

IX. Summary

In this paper, we mathematically argued that community data trading is not economically welfare efficient, at least in a duopoly market setting. Our insights point to an economically inefficient market state if profit-making data trading platforms were to form an oligopoly market. More specifically, we modeled our duopoly setting as a novel embedded Stackelberg

game (ESG) that is a three-level dynamic game with the first level being a platform joining game for users, followed by a two-layer Stackelberg game between end-users and their respective platforms. For both settings where negative data breach externalities affect users within and between platforms, we show that the duopoly information trading market exists but is inefficient, and subsequently characterize the upper bound of the market efficiency. We also considered a duopoly competition setting with pre-set prices to attract customers - an environment analogous to marketing ads adopted by various social platforms to attract consumers. This setting also entails a standard market Nash equilibrium that is inefficient. Our obtained results are in stark contrast to an existing general economic philosophy/intuition that increased amounts of end-user data signals in a market improves utilitarian social welfare, i.e., efficiency. The primary reason explaining our results is the significant negative externality (via user signal correlations) generated by privacy breaches in the information market that cannot be cancelled out via market equilibrium prices handed over to the users for their information. Such an externality may be compensated for in a non-community trading environment.

CONTRIBUTION STATEMENT

R. Pal, J. Crowcroft, and M. Liu conceptualized the system and designed the system model. R. Pal, J. Li and Y. Wang analyzed the system model. Y. Li and S. Tarkoma designed the numerical examples to illustrate the theory. R. Pal, M. Liu, and J. Crowcroft wrote the paper.

ACKNOWLEDGEMENT

The authors would like to thank Achilleas Anastasopoulos for his insightful analysis and comments on this work. This work is supported by the NSF under grants CNS-1616575, CNS-1939006, CNS-2012001, and ARO W911NF1810208.

REFERENCES

[1] R. Poser, "The right of privacy," *Georgia Law Review*, vol. 12, no. 3, 1978.
[2] R. Poser, "The economics of privacy," *American Economic Review*, vol. 71, no. 2.
[3] G. Stigler, "An introduction to privacy in economics and politics," *Journal of Legal Studies*, vol. 9, no. 4, 1978.
[4] K. C. Laudon, "Markets and privacy," *Commun. ACM*, vol. 39, pp. 92–104, Sept. 1996.
[5] A. Acquisti, C. Taylor, and L. Wagman, "The economics of privacy," *Journal of Economic Literature*, vol. 54, no. 2, pp. 442–92, 2016.
[6] A. Odlyzko, "Privacy, economics, and price discrimination on the internet," *Economics of Internet Security (Eds. Jean Camp, Stephen Lewis*, 2003.
[7] P. Samuelson, "Privacy as intellectual property?," *Stanford law review*, pp. 1125–1173, 2000.
[8] P. M. Schwartz, "Property, privacy, and personal data," *Harv. L. Rev.*, vol. 117, p. 2056, 2003.
[9] E. A. Posner and E. G. Weyl, *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press, 2018.
[10] H. R. Varian, "Economic aspects of personal privacy," in *Internet policy and economics*, pp. 101–109, Springer, 2009.
[11] M. Farboodi, R. Mihet, T. Philippon, and L. Veldkamp, "Big data and firm dynamics," in *AEA papers and proceedings*, vol. 109, pp. 38–42, 2019.
[12] R. Calo, "Privacy and markets: a love story," *Notre Dame L. Rev.*, vol. 91, p. 649, 2015.
[13] H. Varian, "Economics aspects of personal privacy," *Privacy and Self-Regulation in the Information Age*, 1997.
[14] A. Goldfarb and C. Tucker, "Shifts in privacy concerns," *American Economic Review*, vol. 102, no. 3, pp. 349–53, 2012.
[15] R. Montes, W. Sand-Zantman, and T. Valletti, "The value of personal information in online markets with endogenous privacy," *Management Science*, vol. 65, no. 3, pp. 1342–1362, 2019.
[16] J. Hirschleifer, "The private and social value of information and the reward to inventive activity," *American Economic Review*, vol. 61, no. 4, 1971.
[17] J. Hirschleifer, "Privacy: Its origin, function, and future," *Journal of Legal Studies*, vol. 9, no. 4, 1980.
[18] J. Burke, C. Taylor, and L. Wagman, "Information acquisition in competitive markets: An application to the us mortgage market," *American Economic Journal: Microeconomics*, vol. 4, no. 4, 2012.
[19] L. Wagman, "Good news or bad news?: Information acquisition and applicant screening in competitive labor markets," *SSRN*, 2014.
[20] A. Daughety and J. Reinganum, "Public goods, social pressure, and the choice between privacy and publicity," *American Economics Journal: Microeconomics*, vol. 2, no. 2, 2010.
[21] M. Spence, "Job market signalling," *Quarterly Journal of Economics*, vol. 2, no. 2, 2010.
[22] R. H. Coase, "The problem of social cost," in *Classic papers in natural resource economics*, pp. 87–137, Springer, 1960.
[23] P. Bolton and M. Dewatripont, *Contract Theory*. MIT Press, 2005.
[24] R. Pal and J. Crowcroft, "Privacy trading in the age of surveillance capitalism: Viewpoints on 'privacy-preserving' societal value creation," *ACM SIGCOMM Computer Communication Review*, vol. 49, no. 3, 2019.
[25] M. MacCarthy, "New directions in privacy: Disclosure, unfairness and externalities," *ISJLP*, vol. 6, p. 425, 2010.
[26] J. A. Fairfield and C. Engel, "Privacy as a public good," *Duke LJ*, vol. 65, p. 385, 2015.
[27] J. P. Choi, D.-S. Jeon, and B.-C. Kim, "Privacy and personal data collection with information externalities," *Journal of Public Economics*, vol. 173, pp. 113–124, 2019.
[28] D. Bergemann and A. Bonatti, "Markets for information: An introduction," *Annual Review of Economics*, vol. 11, pp. 85–107, 2019.
[29] A. de Corniere and R. De Nijs, "Online advertising and privacy," *SSRN*, 2014.
[30] J. Levin and P. Milgrom, "Online advertising: Heterogeneity and conflation in market design," *American Economic Review*, vol. 100, no. 2, 2010.
[31] D. Bergemann and A. Bonatti, "Targeting in advertising markets: Implications for offline versus online media," *RAND Journal of Economics*, vol. 42, no. 3, 2011.
[32] S. Cowan, "The welfare effects of third-degree price discrimination with non-linear demand functions," *RAND Journal of Economics*, vol. 38, no. 2, 2007.
[33] J. J. Anton and D. A. Yao, "The sale of ideas: Strategic disclosure, property rights, and contracting," *The Review of Economic Studies*, vol. 69, no. 3, pp. 513–531, 2002.
[34] M. Babaioff, R. Kleinberg, and R. Paes Leme, "Optimal mechanisms for selling information," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 92–109, 2012.
[35] P. Eső and B. Szentes, "Optimal information disclosure in auctions and the handicap auction," *The Review of Economic Studies*, vol. 74, no. 3, pp. 705–731, 2007.
[36] J. Hörner and A. Skrzypacz, "Selling information," *Journal of Political Economy*, vol. 124, no. 6, pp. 1515–1562, 2016.
[37] A. R. Admati and P. Pfleiderer, "Selling and trading on information in financial markets," *The American Economic Review*, vol. 78, no. 2, pp. 96–103, 1988.
[38] J. Begenau, M. Farboodi, and L. Veldkamp, "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, vol. 97, pp. 71–87, 2018.
[39] A. Ghosh and A. Roth, "Selling privacy at auction," *Games and Economic Behavior*, vol. 91, pp. 334–346, 2015.
[40] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 568–585, ACM, 2012.
[41] K. Ligett and A. Roth, "Take it or leave it: Running a survey when privacy comes at a cost," in *International Workshop on Internet and Network Economics*, pp. 378–391, Springer, 2012.
[42] A. Roth and G. Schoenebeck, "Conducting truthful surveys, cheaply," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 826–843, ACM, 2012.
[43] A. Ghosh and K. Ligett, "Privacy and coordination: computing on databases with endogenous participation," in *Proceedings of the fourteenth ACM conference on Electronic commerce*, pp. 543–560, ACM, 2013.
[44] K. Nissim, S. Vadhan, and D. Xiao, "Redrawing the boundaries on purchasing data from privacy-sensitive individuals," in *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 411–422, ACM, 2014.

[45] A. Ghosh, K. Ligett, A. Roth, and G. Schoenebeck, "Buying private data without verification," in *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 931–948, ACM, 2014.

[46] D. Xiao, "Is privacy compatible with truthfulness?," in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 67–86, ACM, 2013.

[47] Y. Chen, S. Chong, I. A. Kash, T. Moran, and S. Vadhan, "Truthful mechanisms for agents that value privacy," *ACM Transactions on Economics and Computation (TEAC)*, vol. 4, no. 3, p. 13, 2016.

[48] M. M. Khalili, X. Zhang, and M. Liu, "Contract design for purchasing private data using a biased differentially private algorithm," in *Proceedings of the 14th Workshop on the Economics of Networks, Systems and Computation*, p. 4, ACM, 2019.

[49] W. Wang, L. Ying, and J. Zhang, "The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, pp. 249–260, ACM, 2016.

[50] R. Pal and J. Crowcroft, "Privacy trading in the surveillance capitalism age viewpoints on'privacy-preserving'societal value creation," *ACM SIGCOMM Computer Communication Review*, vol. 49, no. 3, pp. 26–31, 2019.

[51] R. Pal, J. Crowcroft, Y. Wang, Y. Li, S. De, S. Tarkoma, M. Liu, B. Nag, A. Kumar, and P. Hui, "Preference-based privacy markets," *IEEE Access*, vol. 8, pp. 146006–146026, 2020.

[52] W. Jin, M. Xiao, M. Li, and L. Guo, "If you do not care about it, sell it: Trading location privacy in mobile crowd sensing," in *IEEE INFOCOM*, IEEE, 2019.

[53] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[54] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang, "Privacy at scale: Local differential privacy in practice," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 1655–1658, 2018.

[55] D. Fudenberg and J. Tirole, "Game theory," 1991.

[56] D. M. Topkis, "Minimizing a submodular function on a lattice," *Operations research*, vol. 26, no. 2, pp. 305–321, 1978.

[57] D. Monderer and L. S. Shapley, "Potential games," *Games and economic behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[58] C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, *et al.*, "Improving the quality of reporting of randomized controlled trials: the consort statement," *Jama*, vol. 276, no. 8, pp. 637–639, 1996.

[59] J. Kendall, "Designing a research project: randomised controlled trials and their principles," *Emergency medicine journal: EMJ*, vol. 20, no. 2, p. 164, 2003.

[60] T. R. Frieden, "Evidence for health decision making—beyond randomized, controlled trials," *New England Journal of Medicine*, vol. 377, no. 5, pp. 465–475, 2017.

[61] A. Deaton and N. Cartwright, "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, vol. 210, pp. 2–21, 2018.

[62] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits, "{BLENDER}: Enabling local search with a hybrid differential privacy model," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, pp. 747–764, 2017.

[63] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015.

[64] T. Dienlin and S. Trepte, "Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors," *European journal of social psychology*, vol. 45, no. 3, pp. 285–297, 2015.

[65] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 192–203, 2016.

Ranjan Pal is on the faculty of ECE at University of Michigan Ann Arbor. His primary research interest lies in engineering robust cyber-security and information privacy solutions using tools from decision and the applied mathematical sciences. Ranjan received his PhD in Computer Science from USC's Viterbi School of Engineering, and was a postdoctoral fellow at the University of Cambridge (CST, DPMMS). He is a member of IEEE and INFORMS.

Yixuan Wang is simultaneously working towards the bachelor's degree in computer science at the University of Michigan Ann Arbor. Her main research interests include AI, machine learning, privacy, and human computer interaction. Yixuan is a student member of the IEEE and the ACM.

Junhui Li is working towards a bachelor's degree in computer science at the University of Michigan Ann Arbor. Her primary research interests lie in machine learning, privacy and security, and human computer interaction. She is a student member of the IEEE.

Mingyan Liu is an entrepreneur and the Peter and Evelyn and Fuss Chair Professor of Electrical and Computer Engineering at University of Michigan Ann Arbor. Her current research interests lie in communication networks, decision theory, incentive design for cybersecurity and privacy, online learning, and experimental data science related to cybersecurity. She was a co-founder of the cybersecurity scoring startup Quadmetrics in 2014 that got acquired by FICO in 2016. She is a Fellow of the IEEE and a member of the ACM.

Jon Crowcroft is the Marconi Professor of Communications Systems in the Computer Laboratory at the University of Cambridge, and a member of the Alan Turing Institute. His current active research areas are Opportunistic Communications, Social Networks, Privacy Preserving Analytics, and techniques and algorithms to scale infrastructure-free mobile systems. Since 2016, he has been Programme Chair at the Alan Turing Institute. He is a Fellow of the Royal Society, ACM, the Royal Academy of Engineering and IEEE.

Yong Li received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the PhD degree in electronics engineering from Tsinghua University, Beijing, China, in 2012. Yong did his postdoctoral work at T-Labs, Berlin. He is currently a faculty member with the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of networking and communications.

Sasu Tarkoma is a Professor of Computer Science with the University of Helsinki and the Head of the Department of Computer Science. He is also affiliated with the Helsinki Institute for Information Technology. He has authored four textbooks and has published over 160 scientific articles. He has seven granted U.S. patents. His research interests are Internet technology, distributed systems, data analytics, and mobile and ubiquitous computing. He was a recipient of best paper awards in IEEE PerCom, ACM CCR, and ACM OSR.