Data Externalities and Data Acquisition by Online Platforms*

Jieteng Chen
The Chinese University of Hong Kong
jieteng.chen@link.cuhk.edu.hk

T. Tony Ke
The Chinese University of Hong Kong
tonyke@cuhk.edu.hk

June 2025

*Very preliminary. Comments welcome.

Acquiring Data with Externalities for Matchmaking

Abstract

In the digital era, platforms actively acquire consumer data to improve match efficiency between two sides. Under the prevalent privacy regulations, the platform can only obtain consumer data upon their consent. However, even if a consumer opts out of the data collection, their information can still be leaked by others' data sharing because a consumer's data are predictive of others' preferences, thereby generating data externalities. This paper investigates the platform's optimal data acquisition strategy under privacy rights and data externality. We find that the platform only needs to compensate consumers who share data based on the consumption utility difference between sharing and not sharing data, which is endogenously affected by others' data sharing. In equilibrium, the platform balances the benefit of data to optimize match efficiency through personalized recommendations against the cost of data acquisition. As information correlation increases, the benefit of individual data for learning this specific consumer's preference declines because the information could be more accurately predicted from others' data. Conversely, the value of individual data for predicting other ones' preferences is enhanced, and the costs of data acquisition are lower. Consequently, the platform may acquire data from more or fewer consumers as the information correlation rises. We also discuss the implications for platform profit, consumer surplus, and social welfare.

Keywords: data externality, data acquisition, consumer privacy, online platform

1 Introduction

From social media and e-commerce to various types of digital services, online platforms have become indispensable to our daily lives, fundamentally changing the way we communicate, socialize, shop, learn, travel, entertain, and work. Behind all these capabilities, the core functionality that an online platform provides is matchmaking—connecting users with the content, products, services, or people that meet their specific needs and preferences. In this view, modern online platforms can be considered an incarnation of the old business of matchmakers; however, the underlying technology that enables their scalable matchmaking capabilities is both distinct and unprecedented. Specifically, these online platforms rely on their ability to collect vast amounts of user data and use sophisticated algorithms to analyze this data, offering valuable information about potential matches between two sides of the marketplace. Data has become the new oil that powers these modern matchmakers.

Meanwhile, the relentless pursuit of personal data has sparked consumers' increasing concerns about privacy and data security. With the introduction of privacy and data protection laws such as GDPR, platforms can only access users' data after obtaining their explicit consent. The immense value of user data on one hand, together with users' privacy safeguard on the other hand, has motivated many platforms to proactively compensate their users so as to acquire their data. For example, it is a common practice for many online platforms, such as Netflix, Uber and Airbnb, to offer free trials or discounts as incentives to attract new users while simultaneously gaining access to their demographic and behavioral data. For another example, Amazon's Shopper Panel program rewards customers for sharing their purchase data generated outside the platform. Also in e-commerce, Rakuten provides cash or gift cards in return for completing surveys on consumer habits. Similarly, Google's Opinion Rewards program offers users credits for completing surveys; while Microsoft operates a reward program that offers points for using Bing, which users can exchange for gift cards or sweepstakes entries, while Microsoft collects data on users' search patterns.

These industry trend and business practices motivate our study of an online platform's optimal data acquisition strategy against the backdrop of consumers' privacy safeguard. Specifically, when the platform needs to incentivize consumers to share their data instead of freely using it for matchmaking, how should it balance the cost of data aquisition and the benefit of match efficiency improvement? How does a platform's dual roles as a matchmaker and a data aggregator interact? Furthermore, how the acquisition and aggregation of user data by the platform impacts the marketplace

and consequently, consumer and social welfare? These questions are becoming more and more relevant, as personal data is becoming increasingly valuable, and a growing number of consumers are becoming more aware of how their information is used for profit making, prompting a call for greater control and fair compensation. Indeed, new infrastructure is under way in reponse to this call. For example, Brave Browser provides its users with Basic Attention Tokens as a reward for allowing targeted ads, which users can redeem for various services or support their favorite content creators.

In studying the data acquisition and aggregation process by an online platform, an important consideration is the information correlation among consumers and the resulting data externalities. Specifically, when one consumer consents to share her personal data, she not only reveals information about her preference but also enables the platform to improve algorithms to predict other consumers' preferences. This interconnectedness from how the platform utilizes consumer data leads to data externalities. In particular, many e-commerce platforms, such as Amazon, personalize recommendations for users based others' choices such as "Customers also viewed" and "Customers also bought". Netflix famously pioneered the collaborative filtering algorithm, which predicts what movies a user likes based on other users' watch and rating data. How would data externalities moderate the platform's optimal data acquisition strategy? On one hand, data externalities allow the platform to infer new users' preferences from existing users' data, leading to improved match efficiency and resulting in higher commission; on the other hand, data externalities also impose a more subtle impact on the platform's data acquisition cost as they change the consumers' expected match as well as retail price depending on whether they consent to data sharing.

In this paper, we study an online platform's optimal data acquisition strategy amidst consumer privacy concerns and data externalities. We develop a theoretical model where a monopolistic platform intermediates the transactions between many sellers and a group of consumers who have correlated preferences toward product designs. To improve match efficiency and thereby the commission, the platform acquires data from some consumers by compensating both their intrinsic privacy costs and instrumental economic loss to obtain their consent. Then the platform uses collected data to predict remaining consumers' preferences, based on which, it recommends sellers' products to them. Lastly, recommended sellers price discriminate consumers based on the inferred consumer preferences.

We first establish that consumers who share their data impose a negative externality on those who do not share data. In fact, other consumers' data allow the platform to make more accurate prediction about a given consumer's preference, which enables

the platform to recommend a product that better matches with the consumer but also results in a higher price charged by the seller. It turns out that the resulting price hike always dominates the match improvement so that the consumer gets worse off as more of others share their data. For similar reasons, as the information correlation among consumers gets higher, a given consumer who does not share data will get worse off when more of others share data. Consequently, the platform could incentivize this consumer to share data by offering a lower price. This implies that the equilibrium price of data will decrease with the information correlation.

Moreover, information correlation influences the platform's data acquisition strategy through this fundamental tradeoff. First, by acquiring data from an additional consumer, the platform can improve the match efficiency by directly learning this consumer's preference and indirectly predicting other consumers' preferences. On the one hand, information correlation can substitute the value of individual data as this consumer's preference can be more accurately inferred from other consumers' data, reducing the value of individual data (*Substitution Effect*). On the other hand, the data can also be used to predict other people's preferences, and thus a higher information correlation amplifies the value of individual data in terms of prediction (*Prediction Effect*). Second, due to the negative data externalities, acquiring additional data also reduces the unit price of data for other consumers. In other words, under a higher information correlation, costs of data acquisition become lower, which encourages the platform to acquire more data (*Cost Reduction Effect*). To summarize, as information correlation rises, whether the platform acquires data from more or fewer consumers depends on the relative magnitude of these three effects. We find that if consumer privacy costs are low, it is profitable for the platform to acquire data from all consumers even without information correlation. In this case, when the market size is sufficiently large, the substitution effect can dominate so that the platform acquires data from fewer consumers as information correlation rises; otherwise, when the market size is small, the platform will always acquire data from all consumers. In contrast, if privacy costs are relatively high, the platform finds it unprofitable to acquire data in the absence of information correlation. As information correlation rises, the platform acquires data from more consumers because the prediction effect and cost reduction effect always dominate.

Additionally, we analyze the impact of information correlation on consumer surplus and social welfare. A higher information correlation affects consumer surplus and social welfare through two channels. First, conditional on the platform's data acquisition strategy, it improves social welfare by prompting more efficient transactions, but hurts consumer surplus by reducing surplus from product consumption and prices of

data. Second, it may influence the platform's data acquisition decision, which can enhance or harm consumer surplus and social welfare.

Finally, we extend the main model to a case where the market is only partially covered in equilibrium. When the consumer does not share data, the platform still recommends the most suitable product to consumers, and the recommended seller charges a fixed monopoly price. A higher information correlation generates positive externalities and benefits the consumer because the product price is fixed and the product match is improved. We find that the platform's data acquisition decision is still driven by the fundamental tradeoff over the benefit and cost of data acquisition and thus the main result still holds.

This paper is structured as follows. We review the relevant literature in Section 2 and present the model setup in Section 3. Then we solve the equilibrium in Section 4 and analyze the impact of information correlation on data acquisition, platform profit, consumer surplus, and social welfare in Section 5. Section 6 considers one extension of a partially covered market. Section 7 provides conclusions.

2 Literature Review

As Schneier (2015) succinctly puts in his book *Data and Goliath*, "Data is the pollution problem of the information age." This idea of negative data externalities has been formalized by Choi et al. (2019) and Acemoglu et al. (2022), who show that when other consumers' data sharing reveals sensitive information on a given consumer and thus partially compromises her privacy, excessive sharing of data happens in equilibrium despite of individual consumers' privacy concerns. This provides a plausible explanation for the empirically documented so-called digital privacy paradox (Norberg et al. 2007; Athey et al. 2017). Along this line, Miklós-Thal et al. (2024) allow a firm to learn about the correlation between users' sensitive and non-sentive data so that other consumers' data sharing will enable the firm to draw inferences about a given user's sensitive data based on her non-sensitive data. The paper predicts a polarization of users' data-sharing choices, where both users who share no data and those who share full data grow. In all these papers, data externalities originate from the revelation of consumers' sensitive information by others and the resulting privacy compromise cost. In contrast, we model an intrinsic consumer privacy cost that is unaffected by other consumers' data sharing; instead, data externalities originate from the revelation of consumers' match values by others, and consequently, data externailities could be either negative or positive.

Bergemann et al. (2022) also study data externalities on match information and thus is the most related paper to ours.[1] They consider a data intermediary that buys data from multiple consumers and sells their data to one firm. This is conceptually different from the setting we consider where an online platform matches multiple consumers with multiple firms and earns a commission from successful matches. Moreover, Bergemann et al. (2022) only consider the platform's choice between collecting either all or none of the consumers' data and they show that when the information correlation is high enough, the platform optimally collects all consumers' data. In contrast, we allow the platform to collect any subset of consumers' data and find that higher information correlation could increase or decrease the number of consumers to collect data from. Lastly, we come up with a micro-founded model that emulates the actual data collection process in practice, which naturally gives rise to the information correlation among consumers, while Bergemann et al. (2022) assume an exogenous correlated distribution on consumers' preferences directly.

Besides the papers on data externalities reviewed above, our paper also contributes broadly to three streams of literature. First, it contributes to the literature on the information or data markets. Some works focus on the businesses' incentives to collect user data and users' possible strategic reactions (e.g., Ichihashi 2023, Fainmesser et al. 2023), while others examine the sale of data by a monopolistic data provider (Bergemann and Bonatti 2015, Bergemann et al. 2018, Bounie et al. 2021, Yang 2022), or competing data intermediaries (Ichihashi 2021a). Our paper contributes to the literature by studying the platform's dual roles as a matchmaker and a data aggregator, so that the platform's benefit and cost of data acquisition are endogenized by the matchmaking process.

Second, this paper is related to the literature on economics of platform and two-sided markets (Armstrong 2006, Rochet and Tirole 2006). Existing works view the platform as a marketplace that profits from facilitating transactions and interactions between two sides and study the platform's optimal design problem, such as search design (Hagiu and Jullien 2011; Dukes and Liu 2016; Zhong 2023), rating system (Ke et al. 2024), recommendation system (Zhou and Zou 2023; Qian and Jain 2024), information design (Guda and Subramanian 2019; Ke and Zhu 2021; Romanyuk and Smolin 2019), targeted advertising (Ke et al. 2022; Bergemann and Bonatti 2024), reputation system (Shi et al. 2023), product ranking (Long and Liu 2024), matchmaking technology provision (Wu et al. 2018), and self-preferencing (Hagiu et al. 2022; Long and Amaldoss 2024), etc. Notably, the platform's superior match information is exogenously given in

[1] Ichihashi (2021b) also studies data externalities but focuses on the endogenous design of data externalities.

these papers but endogenized through data acquisition in our setting.

Lastly, this paper also enriches the literature on consumer privacy and regulation (see Acquisti et al. 2016 and Goldfarb and Que 2023 for recent surveys). Within this literature, the most relevant papers are those that study the opposing effects of match improvement and price discrimination upon consumers' sharing of data. De Corniere and De Nijs (2016) present a model where the platform has control over consumer information and decides whether to disclose it to advertisers, which can improve the match between advertisers and consumers and raise prices. Hidir and Vellodi (2021) find that consumers can achieve the optimal disclosure of their information by balancing the gain from product steering against the loss from price discrimination. Ali et al. (2023) demonstrate that consumers can benefit from disclosing data by intensifying competition. Ke and Sudhir (2023) evaluate the impact of GDPR by considering a two-period model where firms collect consumer data to personalize service and price discriminate. We contribute to the literature by studying the effect of data externalities, where a consumer's preference information may be involuntarily revealed by other consumers' data sharing.

3 Model Setup

Consider an online platform that intermediates transactions between a group of $N \geq 2$ consumers and one continuum of sellers, each offering a product. Both consumer preferences and product varieties are represented along a unit-length line. The consumer located at θ_i derives utility $u(\theta_i, x)$ from consuming product x,

$$u(\theta_i, x) = v - t|\theta_i - x| - p_x,$$

where t measures the degree of horizontal differentiation, and p_x is the price of product x, decided by seller x. Consumers have the option of not purchasing any product that provides utility u_0.

The consumer preference and product space can be divided into $K \geq 1$ segments of equal length, as shown in Figure 1. All consumers' preferences $\Theta = \{\theta_1, \cdots, \theta_N\}$ are random variables uniformly distributed in $[0, 1]$ and can belong to only one segment. Such segmentation may come from pre-registered information, such as name, email, IP address, demographics, or historical data before the inception of privacy regulation, which may enable the platform to cluster consumers into different groups based on their similar features. Thus, consumers within the same group may have common

or similar preferences for products. Essentially, we consider a group of consumers whose preferences are correlated through a common interest in product designs, and their common interest is initially unknown to the platform. The larger K, the finer the segmentation and the stronger the information correlation. To better understand this setup, let us consider two extreme cases. When $K = 1$, then the consumer preferences Θ are independent and identically distributed on $[0, 1]$; when K approaches infinity, consumer preferences are perfectly correlated.[2]

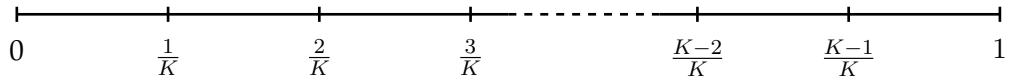


Figure 1: Illustration of Market Segment

By offering a payment of m_i to consumer i to compensate for her privacy cost and potential economic loss, the platform can obtain a signal s_i about θ_i after obtaining consumer i's consent. In general, m_i represents both direct payment and indirect payment, such as services or products provided by the platform in exchange for consumer data. For each consenting consumer i, the platform analyzes its collected data to generate a signal s_i that may reveal the consumer's true preference or not:

$$s_i = \begin{cases} \theta_i & \text{with probability } \sigma \\ \emptyset & \text{with probability } 1-\sigma \end{cases},$$

where $0 < \sigma < 1$ indicates that even if the platform acquires data from consumer i, it may not always successfully identify consumer i's preference. For the remaining consumers who do not share data, we can denote $s_i = \emptyset$ for simplicity.

Based on the acquired dataset $S = \{s_1, \cdots, s_N\}$, the platform recommends product $\hat{\theta}_i$ to consumer i, and consumer i purchases if $u(\theta_i, \hat{\theta}_i) \geq u_0$. The platform profits from the commission $\alpha\pi(\hat{\theta}_i)$, where α is the commission rate, and $\pi(\hat{\theta}_i)$ is the recommended seller's profit before the commission. Consequently, the platform's profit is

$$\Pi = \underbrace{\alpha \sum_{i\in\{1,\ldots,N\}} \pi(\hat{\theta}_i)}_{\text{Commission Revenue}} - \underbrace{\sum_{i\in\{1,\cdots,N\}} m_i}_{\text{Data Acquisition Cost}}.$$

[2]We provide the joint density function of Θ in Appendix. The correlation coefficient between any two consumers' preferences is $\rho(\theta_i, \theta_{i'}) = 1 - 1/K^2$, which increases with K.

The Timing of the Game

In the first stage, the platform decides how much to offer each consumer m_i for their data. In the second stage, consumers anticipate the consequences of sharing data and decide whether to share data with the platform. In the third stage, consumer preferences Θ realize, the platform recommends product $\hat{\theta}_i$ to each consumer, and the recommended sellers make pricing decisions. We assume that sellers and the platform have the same inference about consumer preference. In the last stage, consumers make purchase decisions, and the payoffs of all parties are realized.

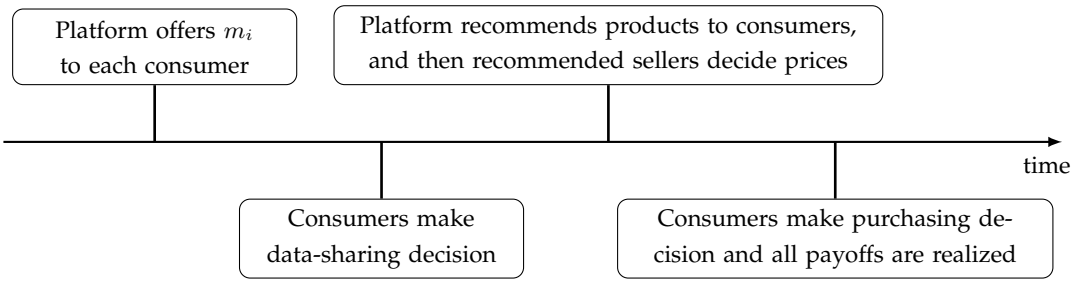


Figure 2: Timing of the Game

Before we proceed to analyze the game, we discuss several model assumptions here. First, when deciding to share data, consumers do not know their own data θ_i as they often need to choose whether to consent to the data collection policy before using a service. Second, how much data the platform collects and how it uses data to build a recommendation system are part of platform design, which the seller should know when setting prices. Third, multiple equilibria may exist in the data-sharing decision subgame due to data externality, so we follow Acemoglu et al. (2022) and focus on the best subgame equilibrium for the platform. We make the following technical assumption to focus on the most meaningful case.

Assumption 1. $0 < t < v$

This assumption ensures that the market is always fully covered in equilibrium. We will consider the case where the market is not fully covered in one extension.

Assumption 2. $\frac{1}{N} < \alpha < 1$

Notice that the platform bears the data acquisition cost, while only α fraction of the benefit from data acquisition is captured by the platform through a fixed commission rate. This assumption ensures the commission rate is sufficiently high so that there

always exists a non-empty parameter range where the platform will acquire data from consumers.

4 Equilibrium Analysis

We solve the game by backward induction. To begin with, we analyze the platform's recommendation strategy and sellers' pricing decisions. Then, we examine consumer data-sharing decisions. Lastly, we solve the platform's optimal data acquisition strategy and derive the equilibrium outcome.

4.1 Platform's Recommendation Strategy and Sellers' Pricing Decision

In this stage, the platform designs the recommendation strategy based on the acquired dataset S. The platform can personalize the recommendations for different consumers based on their data. Without loss of generality, it is convenient to focus on the platform's recommendation strategy for an individual consumer i.

First, we consider the case where the consumer i's preference is perfectly revealed to the platform by herself $s_i = \theta_i$. Suppose the platform recommends seller x to consumer i. As consumer i's location is perfectly known, seller x will charge consumer i's willingness to pay and fully extract her surplus by pricing at $p_x^* = v - t|\theta_i - x|$, leading to profit $\pi^*(x) = v - t|\theta_i - x|$. The platform's optimal recommendation strategy that maximizes expected commission revenue from consumer i is

$$\hat{\theta}_i = \arg\max_x \alpha\pi^*(x) = \theta_i.$$

Second, we consider the scenario where the consumer i's preference is involuntarily leaked by others' data, which occurs when the platform has the other ones' data and uses it to predict consumer i's preference. Specifically, if there is another consumer $-i$'s preference on k th segment with $1 \leq k \leq K$, the platform can infer that consumer i's preference θ_i is uniformly distributed on the same segment since

$$f\left(\theta_i \middle| s_{-i} \in \left[\frac{k-1}{K}, \frac{k}{K}\right)\right) = \begin{cases} K & \text{if } \theta_i \in \left[\frac{k-1}{K}, \frac{k}{K}\right), \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, given that all consumers can only be located in one segment, as long as one of them's location is revealed to the platform, the platform can know that the remaining ones are also in the same segment. Suppose the platform recommends product x to

consumer i. She will buy if and only if

$$v - t|\theta_i - x| - p_x \geq 0 \Leftrightarrow x - \frac{v - p_x}{t} \leq \theta_i \leq x + \frac{v - p_x}{t},$$

which implies that seller x's expected demand is

$$\begin{aligned} D_x(p_x) =& \Pr\left(x - \frac{v - p_x}{t} \leq \theta_i \leq x + \frac{v - p_x}{t} \middle| \frac{k-1}{K} \leq \theta_i < \frac{k}{K}\right) \\ =& K\left\{\min\left\{\frac{k}{K}, x + \frac{v - p_x}{t}\right\} - \max\left\{\frac{k-1}{K}, x - \frac{v - p_x}{t}\right\}\right\}. \end{aligned}$$

By solving the seller x's profit maximization problem, we have $p_x^* = \arg\max_{p_x} p_x D_x(p_x)$ and thus $\pi^*(x) = p_x^* D_x(p_x^*)$. Consequently, the platform's optimal recommendation strategy that maximizes the expected commission revenue is

$$\hat{\theta}_i = \arg\max_x \alpha\pi^*(x) = \arg\max_x \pi^*(x) = \frac{2k-1}{2K}. \tag{1}$$

Third, when consumer i's preference is not revealed to the platform, which happens if the platform does not acquire data from consumers or if the acquired data does not reveal any useful information, the platform and sellers hold the same inference that consumer i's preference is uniformly distributed on $[0, 1]$. This is a special case of the above scenario under $K = 1$. As a result, the optimal recommendation strategy is the same as Equation (1) by setting $k = K = 1$. We summarize the optimal recommendation strategy, the recommended seller's profit before commission, and consumer surplus from product consumption under different information environments in Proposition 1.

Proposition 1 (Optimal Recommendation Strategy)**.**

(i) If consumer i's preference is perfectly known by the platform, then the platform's optimal recommendation strategy is

$$\hat{\theta}_i = s_i,$$

the recommended seller's price and profit are v, and the consumer surplus from product consumption is zero.

(ii) If consumer i's preference is imperfectly known by the platform such that the platform expects that θ_i is uniformly distributed on $\left[\frac{k-1}{K}, \frac{k}{K}\right)$, the platform's optimal recommendation

strategy is given by Equation (1), the recommended seller's price and profit are

$$\pi(K) = v - \frac{t}{2K},$$

and the consumer surplus from product consumption is

$$V(K) = \frac{t}{4K}.$$

Proposition 1 shows that the platform always recommends the product most likely to match consumer preferences, as the platform and sellers share profits through a linear commission contract. Moreover, we find that consumer surplus from product consumption is related to the platform's information. If the platform perfectly knows a consumer's preference, it recommends the most suitable product and the consumer is charged her full willingness to pay, resulting in zero surplus from product consumption. If the platform's information on consumer preferences is imperfect, as described in Proposition 1 (ii), the consumer can obtain a positive surplus from product consumption. In this case, as the K increases, the platform makes a more precise prediction about consumer preference, which not only facilitates transactions involving better-matched products but also enables the recommended seller to extract consumer surplus through higher pricing. We find that consumer surplus from product consumption $V(K)$ decreases with K and eventually diminishes to zero. This is because, with K rises, the recommended seller can charge consumers a price much closer to their maximum willingness to pay v and thereby appropriate a much greater proportion of consumer surplus. [3]

4.2 Consumer Data-Sharing Decision and Price of Data

We now analyze the consumers' data-sharing decisions and the equilibrium price of data. Denote consumer i's data sharing decision by $a_i \in \{0, 1\}$, the profile of consumers' data sharing decisions by $\mathcal{A} = \{a_1, ..., a_N\}$, and decisions of consumers other than i by $\mathcal{A}_{-i}$. Due to information correlation, the platform can infer consumer i's preference not only from her data but also from other consumers' data.

Suppose the platform acquires data from a set of n consumers. For any consumer

[3] A similar intuition is discussed in Pepall et al. (2014) Chapter 7, where a seller offers multiple product varieties to serve different consumer segments on a Hoteling line. When one new product variety is offered, the consumer located near it may benefit from reduced transportation costs, while the seller charges a higher price to all consumers, leading to a reduction in consumer surplus.

i in this set, given that other $n-1$ consumers share data, there are two scenarios to consider: (1) With probability $(1-\sigma)^{n-1}$, the data from those $n-1$ consumers does not reveal their preferences, and thus the platform does not know the consumer i's preference. This is a special case with $K=1$ in Proposition 1(ii). As a result, consumer i's surplus from product consumption is $V(1)$; (2) With remaining probability $1-(1-\sigma)^{n-1}$, at least one of the $n-1$ consumer's data reveals her preference, and thus the platform knows that consumer i is also on the same segment, resulting a surplus $V(K)$ for consumer i. As a result, if consumer i does not share data, her surplus from product consumption is a weighted average of these two possibilities:

$$(1-\sigma)^{n-1}V(1)+\left[1-(1-\sigma)^{n-1}\right]V(K). \tag{2}$$

If the consumer i shares data, then with probability σ, her data is perfectly revealed to the platform, leading to a zero surplus; with probability $1-\sigma$, her surplus from product consumption is the same as if she does not share data, as shown in Equation (2).

The platform incentivizes consumers to share data by offering a payment of m_i to compensate for privacy costs and economic loss for each consumer in this set. Given that $n-1$ other consumers share data, consumer i's total utility is

$$U_i(a_i,\mathcal{A}_{-i})=\begin{cases} m_i-c+(1-\sigma)\left\{(1-\sigma)^{n-1}V(1)+[1-(1-\sigma)^{n-1}]V(K)\right\} & \text{if } \ a_i=1,\\ (1-\sigma)^{n-1}V(1)+[1-(1-\sigma)^{n-1}]V(K) & \text{if } \ a_i=0,\end{cases}$$

which implies that consumer i will share data if and only if

$$U_i(1,\mathcal{A}_{-i})\geq U_i(0,\mathcal{A}_{-i})\Leftrightarrow m_i\geq m(n)\equiv c+\sigma\left\{(1-\sigma)^{n-1}V(1)+[1-(1-\sigma)^{n-1}]V(K)\right\}. \tag{3}$$

Notice that $m(n)$ is the minimum unit price of data when the platform acquires data from a set of n consumers. To acquire data from n consumers, the platform can offer $m_i=m(n)$ to a set of n consumers, and $m_i=0$ to the rest.

We examine how information correlation influences the unit price of consumer data, $m(n)$ for $n\in\{1,\ldots,N\}$ in the following Lemma 1.

Lemma 1 (Effect of Information Correlation on Price of Data)**.**

(i) For $n=1$, $m(1)$ is invariant in K.

(ii) For $n\geq 2$, $m(n)$ decreases with K and $m(n)-m(n-1)\leq 0$.

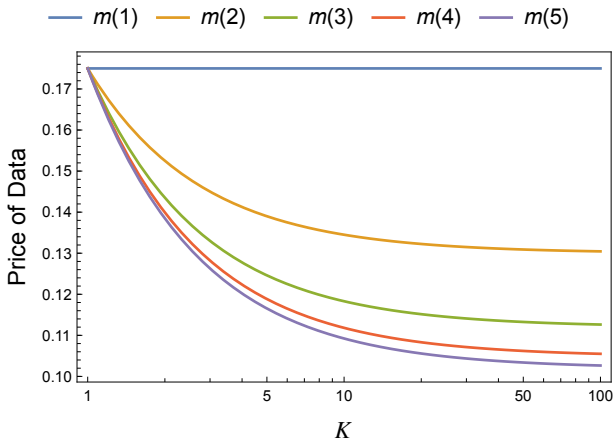


Figure 3: Effect of Information Correlation on Price of Data
($\sigma = 0.6, t = 0.5, c = 0.1$)

The price of data is influenced by the consumer's outside option value of refusing to share data and privacy cost. Consumers are more willing to share data when their outside options are of low value and privacy costs are low. When the platform acquires data from only one consumer, her preference can never be inferred from others and thus her outside option is independent of data externality. Consequently, the price of data $m(1)$ is independent of the information correlation.

When the platform acquires data from multiple consumers, data externality arises and plays an important role in determining the price of data. In this case, the platform can use data from other consumers to predict consumer i's preference, leaving consumer i the outside option of value $V(K)$ with positive probability. First, as the information correlation becomes higher, the prediction about consumer i's preference is more accurate, which not only helps consumer i to get a more suitable product but also enables the recommended seller to charge a higher personalized price. The latter one is more prominent, leading to a reduction in $V(K)$. In other words, the information correlation leads to a negative data externality among consumers. As a result, the unit price of data decreases with K. Second, when more other consumers share data, consumer i's preference is more likely to be leaked. As a consequence, the price of data also declines with the number of consumers who share data, as shown in Figure 3.

4.3 Platform's Data Acquisition Decision

Now, we investigate the platform's optimal data acquisition strategy. When the platform acquires data from n consumers, the number of consumers whose preferences

are perfectly revealed, denoted by l, follows a Binomial distribution with parameters n and σ. If $l = 0$, the platform does not have information about consumer preferences, and the sellers' total profits under optimal recommendation are $N\pi(1)$. If $l \geq 1$, the platform leverages the information of those whose preferences are known to predict other consumers' preferences. Specifically, for consumers whose preferences are revealed, the recommended seller's profit is v. In contrast, for other $N - l$ consumers, the platform knows that they belong to one of the K segments, and the recommended sellers' profit is $\pi(K)$. Thus, the total profit under optimal recommendation is $lv + (N - l)\pi(K)$. As a result, the platform's expected profit when acquiring data from n consumers is

$$\Pi(n) = \alpha \left\{ (1-\sigma)^n N\pi(1) + \sum_{l=1}^{n} \binom{n}{l} \sigma^l (1-\sigma)^{n-l} \left[lv + (N-l)\pi(K) \right] \right\} - n \cdot m(n),$$

which highlights the benefit of consumer data in improving match efficiency. This improvement comes from two aspects: (1) directly learning the preferences of those who share data; (2) predicting the preferences of consumers who do not share data. More subtly, when the platform has data from more consumers, the unit price of data also becomes lower because the information correlation translates into a negative data externality among consumers, and thus makes consumers more willing to share data. In other words, acquiring data from more consumers enables the platform to obtain data at a lower unit price.

To better understand the platform's incentive in data acquisition, we decompose the marginal value of acquiring data from one additional consumer $\Delta\Pi(n) = \Pi(n) - \Pi(n-1)$ for $n \in \{1, \ldots, N\}$ into the following parts:

$$\begin{aligned} &\Delta\Pi(n) \\ =& \underbrace{\alpha\sigma \left\{ v - [(1-\sigma)^{n-1}\pi(1) + [1-(1-\sigma)^{n-1}]\pi(K)] \right\}}_{\text{Improving Profit from } n\text{th Consumer}} + \underbrace{\alpha\sigma(1-\sigma)^{n-1}(\pi(K) - \pi(1))(N-1)}_{\text{Improving Profits from } N-1 \text{ Consumers}} \\ &+ \underbrace{[m(n-1) - m(n)]\,(n-1)}_{\text{Reducing the Unit of Price of Data for } (n-1) \text{ Consumers}} - \underbrace{m(n)}_{\text{Unit Price of Data}} . \end{aligned} \tag{4}$$

This highlights that one consumer's data not only improves the match efficiency by enhancing the prediction of her own preferences and other consumers' preferences but also imposes a negative externality on other consumers, thus reducing the unit price of data for other consumers. In equilibrium, the platform balances the benefits and costs of data acquisition. Before we solve the platform's profit maximization problem, we

provide an important property of the platform's profit function in Lemma 2.

Lemma 2. *Denote* $\Delta\Pi(n) = \Pi(n) - \Pi(n-1)$ *for* $n \in \{1, \ldots, N\}$.

(i) If $K = 1$, $\Delta\Pi(n)$ *is invariant in* n*, i.e.,*

$$\Delta\Pi(1) = \Delta\Pi(2) = \cdots = \Delta\Pi(N).$$

(ii) otherwise, if $K > 1$, $\Delta\Pi(n)$ *is strictly decreasing in* n*, i.e.,*

$$\Delta\Pi(1) > \Delta\Pi(2) > \cdots > \Delta\Pi(N).$$

Lemma 2 indicates that the marginal value of acquiring data from one additional consumer is influenced by the strength of information correlation. When $K = 1$, consumer preferences are independent. Having one's data does not affect the platform's inference of other consumers' preferences, and thus the marginal value of individual data remains constant. In this case, the platform acquires data from either all consumers or none in equilibrium. When $K > 1$, the marginal value of acquiring additional data is decreasing in the amount of data acquired by the platform. Intuitively, this follows from the fact that when other consumers' data reveal more information, there is less to be revealed by any individual's data, and thus the marginal value of acquiring additional data is reduced.

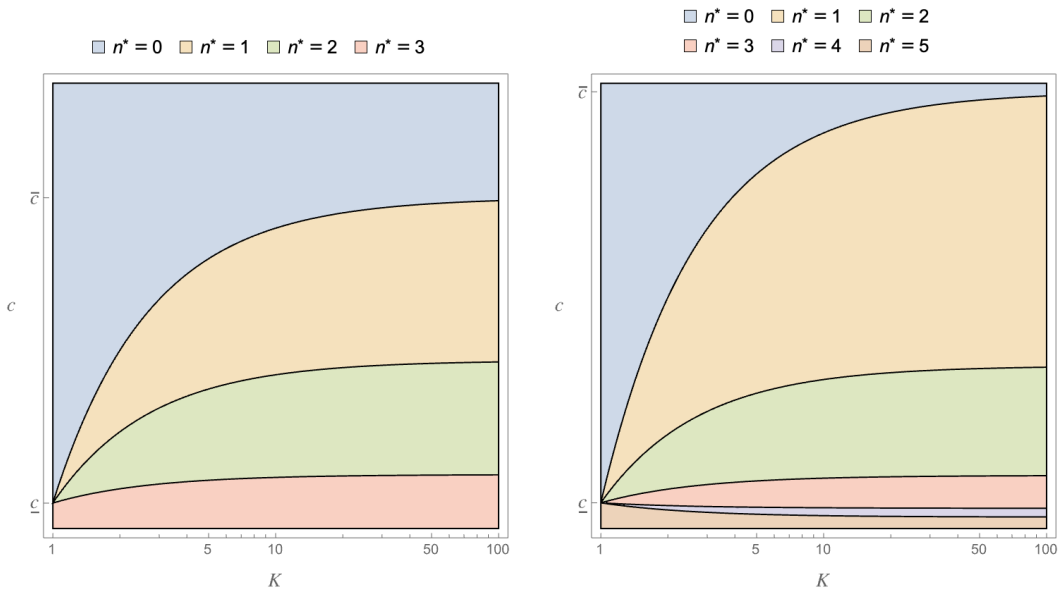
Finally, we characterize the equilibrium by solving the platform's profit maximization problem:

$$n^* = \underset{n \in \{0,1,\ldots,N\}}{\arg\max} \ \Pi(n). \tag{5}$$

Proposition 2 (Equilibrium Characterization)**.** *In equilibrium, the platform acquires data from* n^* *consumers by offering* $m_i = m(n^*)$ *to a set of* n^* *consumers and* $m_i = 0$ *to the rest* $(N - n)$ *consumers, where*

$$n^* = \begin{cases} 0 & \text{if } \ \Delta\Pi(1) < 0, \\ n & \text{if } \ 1 \le n \le N-1 \text{ and } \Delta\Pi(n) \ge 0 \ge \Delta\Pi(n+1, 0), \\ N & \text{if } \ \Delta\Pi(N) \ge 0. \end{cases}$$

The platform's recommendation strategy and the seller's pricing decision are given by Proposition 1.



(a): $(\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, N = 3)$ (b): $(\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, N = 5)$

Figure 4: Equilibrium Characterization

Proposition 2 derives the platform's equilibrium data acquisition strategy by determining the optimal number of consumer data to acquire. Since all consumers are identical in the first stage, the platform can acquire data from n^* consumers by offering $m(n^*)$ to any n^* consumers, whereas the remaining $N - n$ ones will not share data because they do not receive compensation. Figure 4 depicts the equilibrium structure and shows that the platform acquires fewer consumer data as the privacy cost c increases. An interesting observation occurs when $K = 1$. In this scenario, since all consumers' preferences are independent and the marginal value of acquiring additional data remains constant, the platform acquires data from all consumers when the privacy cost is low, and does not acquire data when the privacy cost is high. In contrast, when $K > 1$, the platform optimally balances the benefits and costs of acquiring data, resulting in an optimal number of consumers to acquire data that is neither zero nor the entire population, as shown in Figure 4.

5 Effect of Information Correlation

In this section, we examine the effects of information correlation on the equilibrium outcome in a series of propositions.

5.1 Effect of Information Correlation on Data Acquization

First, we are interested in how information correlation influences the platform's data acquisition strategy. To understand the relationship between n^* and K, it is worthwhile to take a closer examination on the platform' incentive to acquire data by decomposing the effect of K on $\Delta\Pi(n)$ into three components:

$$\frac{d\Delta\Pi(n)}{dK} = \underbrace{-\alpha\sigma\left[1-(1-\sigma)^{n-1}\right]\frac{d\pi(K)}{dK}}_{\text{Substitution Effect}(-)} + \underbrace{\alpha\sigma(1-\sigma)^{n-1}(N-1)\frac{d\pi(K)}{dK}}_{\text{Prediction Effect}(+)} + \underbrace{\left\{\sigma\left[1-(1-\sigma)^{n-1}\right]+\sigma^2(1-\sigma)^{n-2}(n-1)\right\}\left(-\frac{dV(K)}{dK}\right)}_{\text{Cost Reduction Effect}(+)}. \quad (6)$$

Here, the first two terms represent two opposing effects of K on the improvement of match efficiency from acquiring additional data. As K increases, the profit improvement from the individual consumer decreases because her preference can be more accurately inferred from the data of other consumers. In other words, stronger information can be a substitute for consumer data, reducing the benefit of learning individual consumer preference (*Substitution Effect*). On the other hand, data from individual consumer can improve the prediction of other $N-1$ consumers' preferences, leading to a higher profit (*Prediction Effect*).

The last part reflects the effect of information correlation K on $\Delta\Pi(n)$ through data acquisition cost. On one hand, a higher information correlation reduces the consumers' outside option value, leading to a lower price of data. On the other hand, one individual's data enables the platform to acquire other $n-1$ consumers' data at a lower price, and this effect is amplified under a higher information correlation. In short, a higher information correlation incentivizes the platform to acquire more data by influencing data acquisition cost (*Cost Reduction Effect*).

In equilibrium, the impact of information correlation K on the platform's data acquisition strategy depends on the relative magnitude of three effects outlined in Equation (6). Let us consider two extreme cases. When the platform acquires data from one consumer $n=1$, only the positive prediction effect exists since this consumer's preference can only be revealed by her data and thus the negative substitution effect and cost reduction effect are muted. In contrast, when n is sufficiently large, the magnitude of the negative substitution effect becomes larger as the consumer nth's preference is more likely to be inferred from other $n-1$ consumers' data and thus it is possible for

the substitution effect to dominate prediction effect and cost reduction effect. In equilibrium, the optimal number of data points to acquire, n^*, increases with K if and only if $\frac{d\Delta\Pi(n)}{dK}\big|_{n=n^*} \geq 0$. Proposition 3 summarizes the effect of K on n^*.

Proposition 3 (Effect of Information Correlation on Data Acquization)**.** *There exists thresholds $\underline{c}$ and $\bar{c}$ such that*

(i) if $0 \leq c < \underline{c}$,

(a) when $2 \leq N < N^*$, $n^* = N$ for all $K \geq 1$,

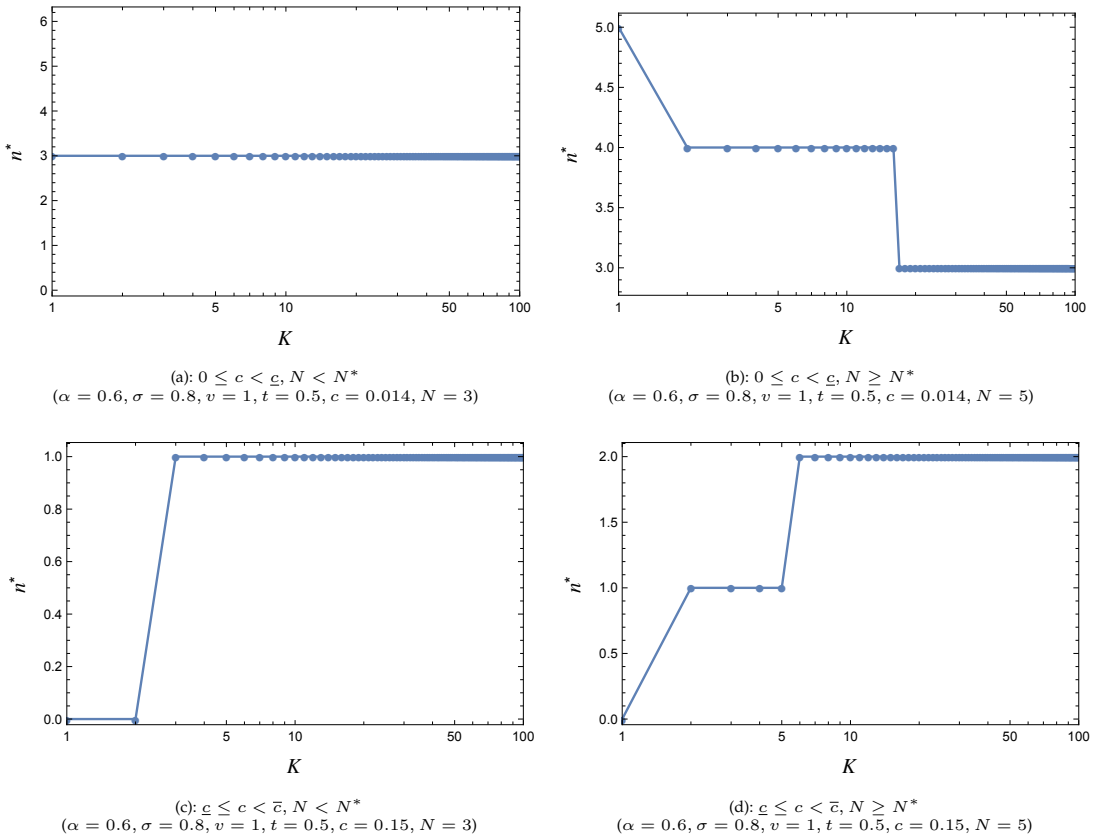
(b) when $N \geq N^*$, n^* decreases with K;

(ii) if $\underline{c} \leq c < \bar{c}$, n^* increases with K.

(iii) If $c \geq \bar{c}$, $n^* = 0$ for all $K \geq 1$.

If the privacy cost is small with $0 \leq c < \underline{c}$, the platform acquires all consumers' data in the absence of information correlation. When N is relatively small $2 \leq N < N^*$, the prediction effect and cost reduction effect can always dominate the substitution effect in Equation (6) such that $\frac{d\Delta\Pi(n)}{dK} > 0$ for all $n \in \{1, \ldots, N\}$. Consequently, the marginal value of acquiring additional data is universally increasing in K, and the platform always acquires data from N consumers, as shown in Figure (5) (a). In contrast, when N is large $N \geq N^*$, the substitution effect can override the prediction effect and cost reduction effect, leading to $\frac{d\Delta\Pi(n)}{dK} < 0$ if n is sufficiently high. As a result, the platform acquires data from fewer consumers due to the reduced marginal value of acquiring data, as illustrated in Figure (5) (b).

If the privacy is intermediate $\underline{c} \leq c < \bar{c}$, the platform refrains from acquiring data without information correlation. However, the marginal value of acquiring additional data $\Delta\Pi(n)$ increases with K if n is small (e.g. $n = 1$), expanding the parameter range that the platform acquires data. So it is profitable for the platform to acquire data only when K is sufficiently high, as shown in Figure 4(c) and (d). If c is sufficiently high $c \geq \bar{c}$, the platform never acquires data from consumers to the prohibitively high costs.



(a): $0 \leq c < \underline{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 3$)

(b): $0 \leq c < \underline{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 5$)

(c): $\underline{c} \leq c < \bar{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 3$)

(d): $\underline{c} \leq c < \bar{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 5$)

Figure 5: Effect of K on n^*

5.2 Effect of Information Correlation on Platform Profit

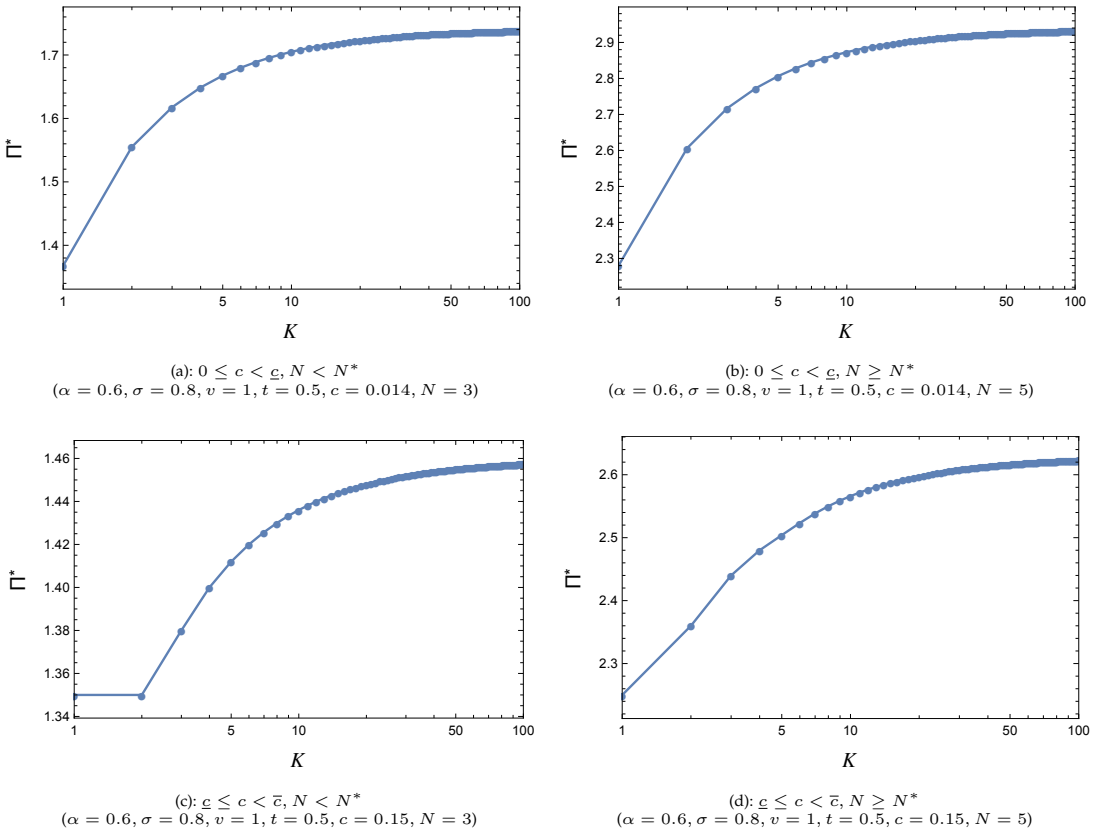
Next, we study how information correlation affects the platform profit in Proposition 4. When the platform acquires data from $n \geq 1$ consumers, the platform's profit $\Pi(n)$ increases with K because a higher information correlation prompts the platform's commission revenue by improving match efficiency and reduces the price of data.

If $0 \leq c \leq \underline{c}$, the platform always acquires consumer data, and thus the platform's prifot strictly increases with K, as illustrated in Figures 6 (a) and (b). If $\underline{c} \leq c < \bar{c}$, the platform acquires consumer data only when K is relatively high, under which the platform profit increases with K. Consequently, the platform's profit can be first invariant in K and then increase with K, or always increase with K, as shown in Figures 6 (c) and (d). If $c \geq \bar{c}$, the platform does not acquire consumer data, so its profit remains constant in K.

Proposition 4 (Effect of Informatino Correlation on Platform Profit)**.**

(i) If $0 \leq c < \underline{c}$, Π^* *increases with* K.

(ii) If $c \geq \overline{c}$, Π^* *is invariant in* K.



(a): $0 \leq c < \underline{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 3$)

(b): $0 \leq c < \underline{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 5$)

(c): $\underline{c} \leq c < \overline{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 3$)

(d): $\underline{c} \leq c < \overline{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 5$)

Figure 6: Effect of K on Platform Profit

5.3 Effect of Information Correlation on Consumer Surplus and Social Welfare

Lastly, we analyze the effect of information correlation on consumer surplus and social welfare in Proposition 5 and Proposition 6.

Proposition 5 (Effect of Information Correlation on Consumer Surplus)**.**

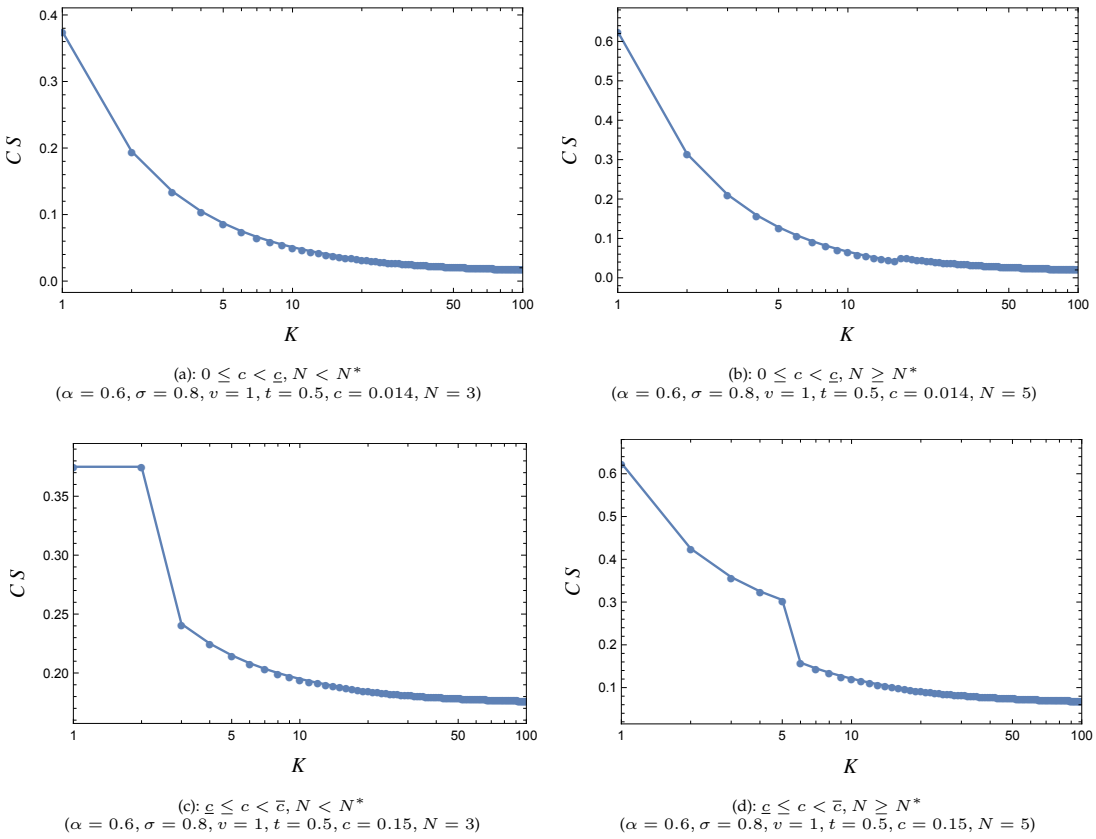
(i) If $0 \leq c < \underline{c}$,

(a) when $2 \leq N < N^*$, CS *decreases with* K;

(b) when $N \geq N^*$, CS *decreases with* K *in general but can discretely jump if the platform acquires data from fewer consumers.*

(ii) If $\underline{c} \leq c < \bar{c}$, CS decreases with K.

(iii) If $c \geq \underline{c}$, CS is invariant in K for $K \geq 1$.



(a): $0 \leq c < \underline{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 3$)

(b): $0 \leq c < \underline{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 5$)

(c): $\underline{c} \leq c < \bar{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 3$)

(d): $\underline{c} \leq c < \bar{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 5$)

Figure 7: Effect of n on Consumer Surplus

For consumers who share data with the platform, their total surplus is the sum of surplus from product consumption, the price of data paid by the platform, and the privacy cost. Essentially, the total surplus is equal to their utility of refusing to share data. For those who do not share data with the platform, their surplus is merely the surplus from product consumption. In equilibrium, consumer surplus is calculated by summing over all consumers' total surplus. A higher information correlation influences consumer surplus through two ways: (1) Given the platform's data acquisition strategy n^*, it negatively impacts the surplus from product consumption and reduces the price of data, thereby hurting consumer surplus. (2) It induces the platform to acquire data from more (fewer) consumers, which can harm (enhance) consumer surplus.

If $0 \leq c < \underline{c}$, the relationship between consumer surplus and information corre-

lation depends on N. When $2 \leq N < N^*$, the platform always acquires data from all consumers. Rasing information correlation leads to a lower consumer surplus (that is, the second effect is muted), as shown in Figure 7(a). When $N \geq N^*$, although consumer surplus still declines with K when n^* remains unchanged, the platform acquires data from fewer consumers as K increases, which could lead to a discrete jump in consumer surplus, as shown in Figure 7(b). If $\underline{c} \leq c < \bar{c}$, consumer surplus remains unchanged with K when K is low since the platform does not acquire data. As K increases, the platform acquires more data, leading to a decline in consumer surplus. This is illustrated by the lower two panels in Figure 7. If $c \geq \bar{c}$, consumer surplus keeps invariant in K because the platform never acquires any data.

Proposition 6 (Effect of Information Correlation on Social Welfare)**.**

(i) If $0 \leq c < \underline{c}$*,*

(a) when $2 \leq N < N^*$*, SW increases with* K*;*

(b) when $N \geq N^*$*, SW can increase with* K *or be non-monotonic in* K*;*

(ii) If $\underline{c} \leq c < \bar{c}$*, SW can increase with* K *or be non-monotonic in* K*.*

(iii) If $c \geq \underline{c}$*, SW is invariant in* K *for* $K \geq 1$*.*

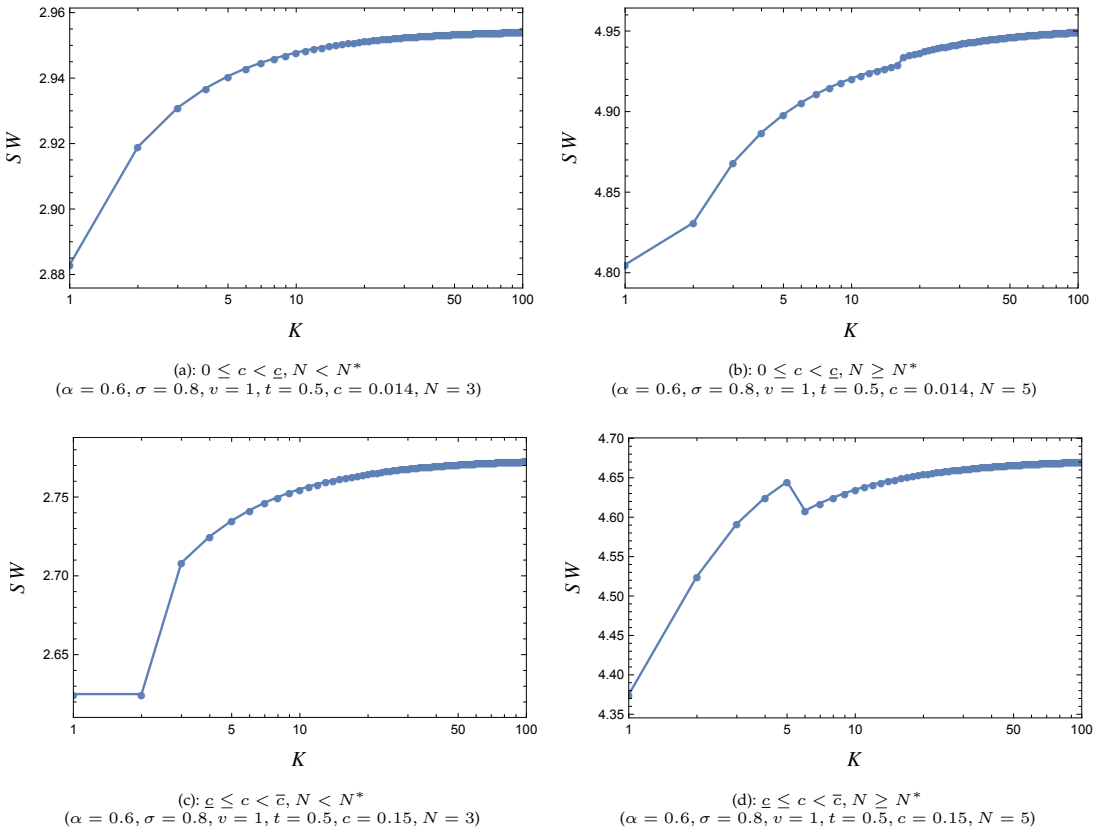
Social welfare compromises consumer surplus, the platform's profit, and the seller's profit. Since the commission and the price of data are just monetary transfers among platform, sellers, and consumers, we can calculate social welfare when the platform acquires data from n consumer as follows:

$$SW(n, K) = \left\{ Nv - \frac{[1 + (K-1)(1-\sigma)^n]N - n\sigma}{4K} t \right\} - nc$$

where the first part represents the welfare generated from product consumption and the second part accounts for consumers' intrinsic privacy cost. On the one hand, conditional on the platform's data acquisition strategy, a higher information correlation enhances social welfare by improving product matches. On the other hand, a higher information correlation may incentivize the platform to acquire more data and thus hurt consumer surplus, thereby reducing social welfare.

If $0 \leq c < \underline{c}$, a higher information correlation generally increases social welfare. When $2 \leq N < N^*$, the platform always acquires data from all consumers, and an increase in information correlation only leads to more efficient transactions and thus higher social welfare, as depicted in Figure 8(a). When $N \geq N^*$, the platform acquires

data from fewer consumers as K increases, which can lead to less efficient trade but a higher consumer surplus. However, social welfare can still increase with K, as shown in Figure 8 (b). If $\underline{c} \leq c < \bar{c}$, the platform always acquires more data as K increases, which not only improves match efficiency but also hurts consumer surplus. As a result, social welfare increases with K when the platform's data acquisition remains unchanged, but can discretely drop with K when the platform acquires additional data, as illustrated in the lower panels in Figure 8.



(a): $0 \leq c < \underline{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 3$)

(b): $0 \leq c < \underline{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.014, N = 5$)

(c): $\underline{c} \leq c < \bar{c}$, $N < N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 3$)

(d): $\underline{c} \leq c < \bar{c}$, $N \geq N^*$
($\alpha = 0.6, \sigma = 0.8, v = 1, t = 0.5, c = 0.15, N = 5$)

Figure 8: Effect of K on Social Welfare

6 Extension: Partially Covered Market

This section considers a case in which the market is partially covered in equilibrium. First, we analyze the platform's recommendation strategy and sellers' pricing. Following the same logic in Proposition 1, the platform still recommends the most suitable product to consumers. However, when the consumer's preference is only imper-

fectly revealed by others' data, the recommended seller's pricing becomes different Essentially, the recommended seller faces consumers in one segment with a length of $1/K$. When t is sufficiently high $t > Kv$, the seller finds it unprofitable to serve all consumers by charging a low price, and thus resorts to a monopoly price of $v/2$. As a result, only consumers with realized preferences close to the recommended seller will make purchases, resulting in a partially covered market. In this case, consumer surplus from product consumption $V(K) = Kv^2/(4t)$ increases with K. This is because the consumer can benefit from better matches as K increases and the recommended seller's price is fixed at $v/2$. In other words, this extension also helps us understand the situation in which data-sharing is beneficial to other consumers and generates positive externalities among consumers. Proposition 7 states the effect of information correlation on the platform's data acquisition strategy and shows that our results still hold.

Proposition 7. *Suppose $\sigma < \frac{2\alpha}{1+2\alpha}$ such that $\Delta\Pi(n)$ decreases with n. There exists threshold of c such that for $1 \leq K < t/v$,*

(i) if $0 \leq c < \underline{c}_p$,

(a) when $2 \leq N < N_p^*$, $n^* = N$,

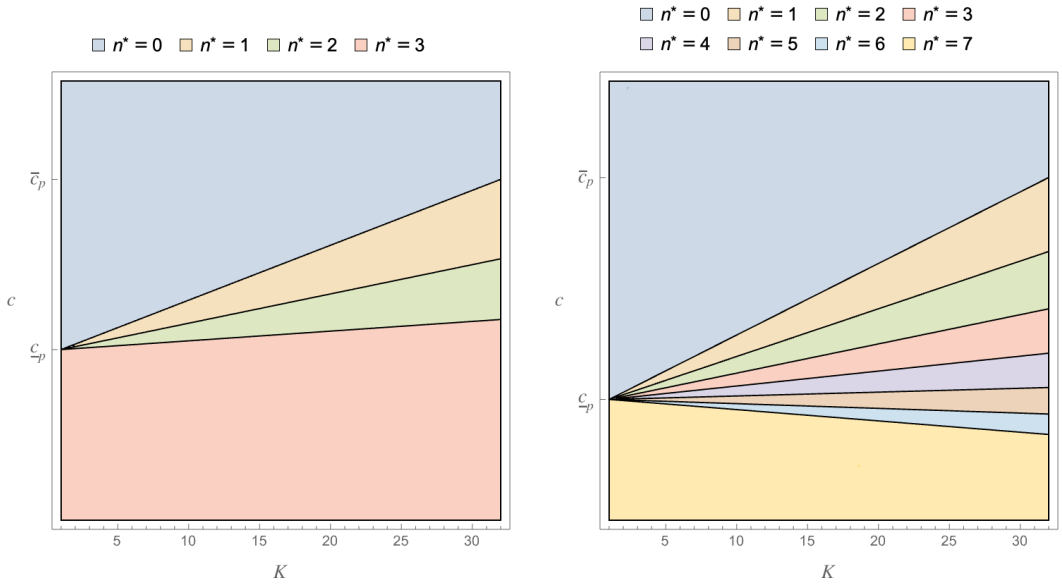
(b) when $N \geq N_p^*$, n^* decreases with K;

(ii) if $\underline{c}_p \leq c < \bar{c}_p$, n^* increases with K ;

(iii) if $c \geq \bar{c}_p$, $n^* = 0$.

A key distinction between the main model and this extension lies in whether data externality is positive or negative. Under a partially covered market, stronger information correlation enables the platform to recommend better-matched products to the consumer without affecting the product price even if she does not share data, leading to a higher surplus from product consumption. Such positive externality raises the outside option value and makes consumers less willing to share data as information correlation increases. This further changes the platform's incentive when deciding to acquire data. Since consumers benefit from other one's data sharing and higher information correlation, the platform has to pay a higher price of data for consumers. As a result, the cost reduction effect in Equation (6) is reversed under a partially covered market. As K increases, the platform acquires data from more consumers if and only if the prediction effect can dominate; otherwise, the platform acquires data from fewer consumers.

If the privacy is small $0 \leq c < \underline{c}_p$, the platform acquires all consumers' data without information correlation. Proposition 7 shows that when N is sufficiently high, the prediction effect can be dominated by the other two effects so that n^* declines with K, as illustrated by Figure 9 (b). Otherwise, when $0 \leq 2 < N_p^*$, the platform always acquires all consumers' data. If the privacy cost is intermediate $\underline{c}_p \leq c < \bar{c}_p$, the platform refuses to acquire data without information correlation. Since $\Delta\Pi(n)$ increases with K due to the prediction effect, the platform acquires more data as K increases. If the privacy cost is sufficiently high $c \geq \bar{c}_p$, the platform does not acquire consumer data because of high costs.



(a): $2 \leq N < N_p^*$
($\alpha = 0.6, \sigma = 0.2, v = 1, t = 32, N = 3$)

(b): $N \geq N_p^*$
($\alpha = 0.6, \sigma = 0.2, v = 1, t = 32, N = 7$)

Figure 9: Equilibrium Structure under Partially Covered Market

7 Conclusion

The growing proliferation of data flows from consumers has driven the development of many platforms and they are actively collecting data from consumers. The impact of platforms' data practice and the balance between utilizing consumer data for economic gain and protecting user privacy remain important issues in this digital age.

Through a theoretical model, we investigate the platform's acquisition and use of consumer data under information correlation among consumer preferences and pri-

vacy rights. We highlight the platform's fundamental trade-off between the benefits and costs of data acquisition. While big data can improve match efficiency and thus the platform's commission revenue, it also raises data acquisition costs. In equilibrium, the platform optimally balances the benefits and costs of data acquisition. We further examine the impacts of information correlation and the platform's data acquisition on consumer surplus as well as social welfare.

There are still several limitations in our research. First, we consider a specific distribution of consumer preferences such that consumers have common interests but can still retain heterogeneities toward product design. This specification approximates the situation in reality where the platform can segment consumer bases into multiple groups, with consumers in the same group having similar preferences, even if such preferences are unknown by the platform. Second, the model assumes that the platform can monetize customer data solely through personalized recommendations by a linear commission rate, which aligns with the online platform literature and industry practice. However, the platform can utilize customer data in various other ways, such as sponsor advertising and marketing analytics, which we leave for further directions.

Appendix

Probability Density Function of Consumer Preferences

Consumer preferences $\Theta = \{\theta_1, \ldots, \theta_N\}$ are correlated and follow a joint density function:

$$f(\Theta) = \begin{cases} K^{N-1} & \text{if } \Theta \in \left[\frac{k-1}{K}, \frac{k}{K}\right)^N, \ \forall\, k \in \{1, \cdots, K\}, \\ 0 & \text{otherwise,} \end{cases}$$

Consider $\theta_i \in \left[\frac{k-1}{K}, \frac{k}{K}\right)$, where $k \in \{1, \ldots, K\}$, we have the marginal density function for individual consumer's preference as follows:

$$f(\theta_i) = \int_{\theta_{-i} \in \left[\frac{k-1}{K}, \frac{k}{K}\right)^{N-1}} f(\Theta) d\theta_{-i} = 1$$

Essentailly, we can have $f(\theta_i) = 1$ for $0 \leq \theta_i \leq 1$, i.e., the marginal distribution of each consumer's preference θ_i is still uniformly on $[0, 1]$.

Proof of Proposition 1

Proof. We consider seller j's profit maximization when $\theta_i \in \left[\frac{k-1}{K}, \frac{k}{K}\right)$. For convenience, we denote $\underline{d} = \min\left\{j - \frac{k-1}{K}, \frac{k}{K} - j\right\}$ and $\overline{d} = \max\left\{j - \frac{k-1}{K}, \frac{k}{K} - j\right\}$.
(i). For $0 \leq j < \frac{k-1}{K}$, seller j's profit function is

$$\pi_j(p_j) = \begin{cases} p_j & \text{if } 0 \leq p_j \leq v - \left(\frac{k}{K} - j\right) t, \\ K p_j \left(j + \frac{v - p_j}{t} - \frac{k-1}{K}\right) & \text{if } v - \left(\frac{k}{K} - j\right) t < p_j \leq v - \left(\frac{k-1}{K} - j\right) t, \\ 0 & \text{if } p_j > v - \left(\frac{k-1}{K} - j\right) t. \end{cases}$$

By solving the profit maximization problem, we have

$$p^*(j) = \begin{cases} \frac{1}{2}\left[v - \left(\frac{k-1}{K} - j\right) t\right] & \text{if } 0 \leq j < \frac{k+1}{K} - \frac{v}{t} \\ v - \left(\frac{k}{K} - j\right) t & \text{if } \frac{k+1}{K} - \frac{v}{t} \leq j < \frac{k-1}{K} \end{cases},$$

and

$$\pi^*(j) = \begin{cases} \frac{K}{4t}\left[v - \left(\frac{k-1}{K} - j\right) t\right]^2 & \text{if } 0 \leq j < \frac{k+1}{K} - \frac{v}{t} \\ v - \left(\frac{k}{K} - j\right) t & \text{if } \frac{k+1}{K} - \frac{v}{t} \leq j < \frac{k-1}{K} \end{cases},$$

which implies that $\pi^*(j)$ increases with j for $0 \leq j < \frac{k-1}{K}$.

(ii). For $\frac{k-1}{K} \leq j < \frac{k}{K}$, seller j's profit function is:

$$\pi_j(p_j) = \begin{cases} p_j & \text{if } 0 \leq p_j \leq v - \overline{d}t \\ Kp_j\left(\underline{d} + \frac{v-p_j}{t}\right) & \text{if } v - \overline{d}t < p_j \leq v - \underline{d}t, \\ 2Kp_j\frac{v-p_j}{t} & \text{if } p_j > v - \underline{d}t. \end{cases}$$

By solving this profit maximization problem, we have the optimal price and profits:

(a) if $0 < t < \frac{Kv}{2}$, $p_j^* = v - \overline{d}t$ and $\pi^*(j) = v - \overline{d}t$;

(b) if $\frac{Kv}{2} \leq t < \frac{2Kv}{3}$,

$$p^*(j) = \begin{cases} \frac{1}{2}(v + \underline{d}t) & \text{if } 0 \leq \underline{d} \leq \frac{2}{K} - \frac{v}{t} \\ v - \left(\frac{1}{K} - \underline{d}\right)t & \text{if } \frac{2}{K} - \frac{v}{t} < \underline{d} \leq \frac{1}{2K} \end{cases} \text{ and } \pi^*(j) = \begin{cases} \frac{K}{4t}(v + \underline{d}t)^2 & \text{if } 0 \leq \underline{d} \leq \frac{2}{K} - \frac{v}{t} \\ v - \left(\frac{1}{K} - \underline{d}\right)t & \text{if } \frac{2}{K} - \frac{v}{t} < \underline{d} \leq \frac{1}{2K} \end{cases};$$

(c) if $\frac{2Kv}{3} \leq t < Kv$,

$$p^*(j) = \begin{cases} \frac{1}{2}(v + \underline{d}t) & \text{if } 0 \leq \underline{d} \leq \frac{v}{3t} \\ v - \underline{d}t & \text{if } \frac{v}{3t} < \underline{d} \leq \frac{1}{2K} \end{cases} \text{ and } \pi^*(j) = \begin{cases} \frac{K}{4t}(v + \underline{d}t)^2 & \text{if } 0 \leq \underline{d} \leq \frac{v}{3t} \\ 2K(v - \underline{d}t)\underline{d} & \text{if } \frac{v}{3t} < \underline{d} \leq \frac{1}{2K} \end{cases}.$$

Notice that $\underline{d}$ increases with j for $\frac{k-1}{K} \leq j \leq \frac{2k-1}{2K}$ and decreases with j for $\frac{2k-1}{K} < j < \frac{k}{K}$. Since $\pi^*(j)$ increases with $\underline{d}$, $\pi^*(j)$ also increases with j for $\frac{k-1}{K} \leq j \leq \frac{2k-1}{2K}$ and decreases with j for $\frac{2k-1}{2K} < j < \frac{k}{K}$.

(iii). For $\frac{k}{K} \leq j \leq 1$, seller j's profit function is

$$\pi_j(p_j) = \begin{cases} p_j & \text{if } 0 \leq p_j \leq v - \left(j - \frac{k-1}{K}\right)t, \\ Kp_j\left[\frac{k}{K} - \left(j - \frac{v-p_j}{t}\right)\right] & \text{if } v - \left(j - \frac{k-1}{K}\right)t < p_j \leq v - \left(j - \frac{k}{K}\right)t, \\ 0 & \text{if } p_j > v - \left(j - \frac{k}{K}\right)t \end{cases}$$

By solving the profit maximization problem, we have

$$p^*(j) = \begin{cases} v - \left(j - \frac{k-1}{K}\right)t & \text{if } \frac{k}{K} \leq j < \frac{v}{t} + \frac{k-2}{K} \\ \frac{1}{2}\left[v - \left(j - \frac{k}{K}\right)t\right] & \text{if } \frac{v}{t} + \frac{k-2}{K} \leq j \leq 1 \end{cases}$$

and

$$\pi^*(j) = \begin{cases} v - \left(j - \frac{k-1}{K}\right)t & \text{if } \frac{k}{K} \leq j < \frac{v}{t} + \frac{k-2}{K} \\ \frac{K}{4t}\left[v - \left(j - \frac{k}{K}\right)t\right]^2 & \text{if } \frac{v}{t} + \frac{k-2}{K} \leq j \leq 1 \end{cases},$$

which implies that $\pi^*(j)$ decreases with j for $\frac{k}{K} \leq j \leq 1$.

Combining (i), (ii), and (iii), we have that $\pi^*(j)$ increases with j for $0 \leq j \leq \frac{2k-1}{2K}$ and decreases with j for $\frac{2k-1}{2K} < j \leq 1$. This implies that the optimal recommendation strategy is

$$\hat{\theta}_i = \arg\max_j \alpha\pi(j) = \arg\max_j \pi(j) = \frac{2k-1}{2K}.$$

The recommended seller's price and profit are

$$p(K) = p^*(\hat{\theta}_i) = v - \frac{t}{2K}, \pi(K) = \pi^*(\hat{\theta}_i) = v - \frac{t}{2K}.$$

Consumer surplus from product consumption is

$$V(K) = \int_{\frac{k-1}{K}}^{\frac{k}{K}} K\left[v - t|\theta_i - \hat{\theta}_i| - p^*(\hat{\theta}_i)\right] d\theta_i = \frac{t}{4K}.$$

□

Proof of Lemma 1

Proof. First, we have $m(1) = c + \frac{\sigma t}{4}$, which is invariant in K. Second, for $n \geq 2$, we have

$$\frac{dm(n)}{dK} = -\sigma[1-(1-\sigma)^{n-1}]\frac{t}{4K^2} < 0.$$

We also have

$$m(n) - m(n-1) = -\frac{(K-1)\sigma^2(1-\sigma)^{n-2}t}{4K^2} \leq 0.$$

□

Proof of Lemma 2

Proof. We have $\Delta\Pi(n) = \left\{2\alpha - 1 + (K-1)(1-\sigma)^{n-2}\left[2\alpha(1-\sigma)N + \sigma n - 1\right]\right\}\frac{\sigma t}{4K} - c$, and

$$\Delta\Pi(n) - \Delta\Pi(n+1) = \frac{(K-1)\sigma^2(1-\sigma)^{n-2}t}{4K}\left[2\alpha(1-\sigma)N + n\sigma - (2-\sigma)\right],$$

which is zero if $K = 1$. If $K > 1$, we have $\Delta\Pi(n) - \Delta\Pi(n+1) > 0$ since

$$2\alpha(1-\sigma)N + n\sigma - (2-\sigma) > 2(1-\sigma) + \sigma - (2-\sigma) = 0,$$

where the $>$ is due to $\alpha > \frac{1}{N}$ and $n \geq 1$. □

Proof of Proposition 2

Proof. This proof mainly follows the results in Lemma 2. If $\Delta\Pi(1) < 0$, we have $\Pi(0) > \Pi(1) > \cdots > \Pi(N)$. If $\Delta\Pi(N) \geq 0$, we have $\Pi(N) \geq \Pi(N-1) \geq \cdots \geq \Pi(0)$. Otherwise, an interior solution $n^* \in \{1, \ldots, N-1\}$ to the platform's maximization problem in Equation (5) can exist if and only

$$\Delta\Pi(n^*) \geq 0 \geq \Delta\Pi(n^*+1, 0),$$

where at least one of the two weak inequalities must be a strict inequality due to the strict monotonicity of $\Delta\Pi(n)$ when $K > 1$. □

Proof of Proposition 3

Proof. We have $\frac{d\Delta\Pi(n)}{dK} = \frac{\sigma t}{4K^2}\left\{(1-\sigma)^{n-2}\left[-1+2\alpha(1-\sigma)N+\sigma n\right]+1-2\alpha\right\}$ and $\frac{d\Delta\Pi(1)}{dK} = \frac{\alpha\sigma t(N-1)}{2K^2} > 0$. Next, we have

$$\frac{d\Delta\Pi(n)}{dK} - \frac{d\Delta\Pi(n+1)}{dK} = \frac{\sigma^2(1-\sigma)^{n-2}t}{4K^2}[2\alpha(1-\sigma)N + n\sigma - (2-\sigma)] > 0,$$

which implies that $\frac{d\Delta\Pi(n)}{dK}$ is strictly decreasing in n as follows:

$$\frac{d\Delta\Pi(1)}{dK} > \frac{d\Delta\Pi(2)}{dK} > \cdots > \frac{d\Delta\Pi(N)}{dK}.$$

(i) If $\Delta\Pi(N)\big|_{K=1} \geq 0 \Leftrightarrow c \leq \underline{c} \equiv (2\alpha-1)\frac{\sigma t}{4}$, the platform acquires data from all consumers when $K = 1$. Let us consider two cases.

- When $\frac{d\Delta\Pi(N)}{dK} \geq 0$, we can directly have $\frac{d\Delta\Pi(n)}{dK} \geq 0$ for all $n \in \{1, \ldots, N\}$ and thus
$$\Delta\Pi(1) \geq \Delta\Pi(2) \geq \cdots \geq \Delta\Pi(N) \geq 0,$$
which implies that the platform always acquires data from all consumers, $n^* = N$.
- When $\frac{d\Delta\Pi(N)}{dK} < 0$, there must exist a threshold $\tilde{n}$ such that
$$\frac{d\Delta\Pi(1)}{dK} > \cdots > \frac{d\Delta\Pi(\tilde{n}-1)}{dK} > 0 > \frac{d\Delta\Pi(\tilde{n})}{dK} > \cdots > \frac{d\Delta\Pi(N)}{dK},$$
which implies that for $n \in \{\tilde{n}, \ldots, N\}$, $\Delta\Pi(n)$ decreases with K. As K increases,

n^* will decreases from N to $\tilde{n}$ if $\lim_{K\to+\infty}\Delta\Pi(\tilde{n})\geq 0$

To delimit the existence conditions of the above two cases, we only need to show that $\frac{d\Delta\Pi(N)}{dK}<0$ if and only if $N>N^*$. We have

$$\begin{aligned}\frac{\partial\Delta\Pi(N,K)}{\partial K}&=\frac{\sigma t}{4K^2}\left\{(1-\sigma)^{N-2}[2\alpha(1-\sigma)N+\sigma N-1]+1-2\alpha\right\}\\&\propto(1-\sigma)^{N-2}[2\alpha(1-\sigma)N+\sigma N-1]+1-2\alpha\equiv G(N).\end{aligned}$$

We can prove that $G(N)<0$ if and only if $N>N^*>2$. First, given $G(2)>0$, $\lim_{N\to\infty}G(N)<0$ and the continuity of $G(N)$ on N, there must exist as least a threshold N^* such that $G(N^*)=0$. Second, we have

$$\begin{aligned}G'(N)&=\{2\alpha(1-\sigma)+\sigma-\log(1-\sigma)+[2(1-\sigma)\alpha+\sigma]N\log(1-\sigma)\}(1-\sigma)^{N-2}\\&\propto 2\alpha(1-\sigma)+\sigma-\log(1-\sigma)+[2(1-\sigma)\alpha+\sigma]N\log(1-\sigma),\end{aligned}$$

which decreases with N and must be negative when N is sufficiently high. This means that if $G'(2)<0$, then $G'(N)<0$ for all $N\geq 2$; if $G'(2)>0$, then $G(N)$ can first increase and then decrease with N. Put together, the uniqueness of N^* can be established.
(ii) If $c>\underline{c}$, the platform does not acquire data when $K=1$, which means that $\Delta\Pi(N)\big|_{K=1}<0$. When $N\leq N^*$, we have $\frac{d\Delta\Pi(1)}{dK}>\frac{d\Delta\Pi(2)}{dK}>\cdots>\frac{d\Delta\Pi(N)}{dK}\geq 0$, which implies that for $n\in\{1,\ldots,N\}$, $\Delta\Pi(n)$ increases with K. As K increases, n^* will increase from 0 to n if $\lim_{K\to\infty}\Delta\Pi(n)\geq 0$. Similarly, When $N>N^*$, we have $\frac{d\Delta\Pi(1)}{dK}>\frac{d\Delta\Pi(2)}{dK}>\cdots>\frac{d\Delta\Pi(\tilde{n}-1)}{dK}\geq 0$, which implies that for $n\in\{1,\ldots,\tilde{n}-1\}$, $\Delta\Pi(n)$ increases with K. As K increases, n^* will increase from 0 to n if $\lim_{K\to\infty}\Delta\Pi(n)\geq 0\lim_{K\to\infty}\Delta\Pi(n)\geq 0$. As a result, if $\lim_{K\to\infty}\Delta\Pi(1)\geq 0\Leftrightarrow c\leq\bar{c}=(2\alpha N-1)\frac{\sigma t}{4}$. As K rises, n^* increases from 0 to a positive integer. Otherwise, if $c\geq\bar{c}$, $n^*=0$ for any $K\geq 1$. □

Proof of Proposition 4

Proof. When $n^*\geq 1$, we have

$$\frac{d\Pi^*}{dK}=\frac{\partial\Pi(n^*,K)}{\partial K}=\alpha\frac{[1-(1-\sigma)^{n^*}]N-n^*\sigma}{2K^2}t+\frac{n^*\sigma t[1-(1-\sigma)^{n^*-1}]}{4K^2},$$

which is strictly positive if $n^*\geq 1$ and equal to 0 if $n^*=0$. □

Proof of Proposition 5

Proof. When the platform acquires data from n consumer, CS is

$$CS(n,K) = n\frac{\left[(1-\sigma)^{n-1}K + 1 - (1-\sigma)^{n-1}\right]t}{4K} + (N-n)\frac{\left[(1-\sigma)^{n}K + 1 - (1-\sigma)^{n}\right]t}{4K},$$

which further implies that for $n \geq 1$, we have

$$\begin{aligned}\frac{dCS(n,K)}{dn} &= \frac{(K-1)(1-\sigma)^{n-1}t}{4K}[\sigma + ((1-\sigma)N + n\sigma)\log(1-\sigma)] \\ &\leq \frac{(K-1)(1-\sigma)^{n-1}t}{4K}[\sigma + (2(1-\sigma)+\sigma)\log(1-\sigma)] \\ &< 0.\end{aligned}$$

Combined with $CS(0,K) - CS(1,K) = \frac{(N-1)(K-1)\sigma t}{4K} > 0$, we can conclude that $CS(n,K)$ decreases with n, i.e., the platform's data acquisition always hurts consumers. We further find that

$$\frac{dCS(n,K)}{dK} = -\frac{n[1-(1-\sigma)^{n-1}]t}{4K^2} - \frac{(N-n)[1-(1-\sigma)^{n}]t}{4K^2} < 0,$$

which means that CS decreases with K when n^* does not change. □

Proof of Proposition 6

Proof. When the platform acquires data from n consumers, SW is

$$SW(n,K) = Nv - \frac{[1+(K-1)(1-\sigma)^{n}]N - n\sigma}{4K}t - nc$$

$$\frac{\partial SW(n,K)}{\partial K} = \frac{[1-(1-\sigma)^{n}]N - \sigma n}{4K^2}t > 0$$

$$\frac{\partial SW(n,K)}{\partial n} = \frac{\sigma - (K-1)(1-\sigma)^{n}\log(1-\sigma)N}{4K}t - c,$$

which can be positive or negative. □

Proof of Proposition 7

Proof. In proof of this extension, we slightly abuse notation and thus still denote $V(K) = \frac{Kv^2}{4t}$ and $\pi(K) = \frac{Kv^2}{2t}$. Then, applying the expression of $\Pi(n)$ in the main model, we can have the marginal value of acquiring additional data $\Delta\Pi(n) = \Pi(n) - \Pi(n-1)$ for

$n \in \{1, \ldots, N\}$ as follows:

$$\Delta\Pi(n) = \frac{\sigma v}{4t}\left\{4\alpha t - (1+2\alpha)Kv + (K-1)v(1-\sigma)^{n-2}\left[1+2\alpha(1-\sigma)N - n\sigma\right]\right\} - c.$$

To ensure the monotonicity of $\Delta\Pi(n)$ in Lemma 2 can hold in this extension, we assume $\sigma < \frac{2\alpha}{1+2\alpha}$ such that for $n \in \{2, \ldots, N\}$,

$$\Delta\Pi(n) - \Delta\Pi(n-1) = -\frac{(K-1)v^2\sigma^2(1-\sigma)^{n-3}\left[2+2\alpha(1-\sigma)N - n\sigma\right]}{4t} < 0.$$

Then, we can still have the monotonicity of $\Delta\Pi(n)$ on n for $n \in \{1, \ldots, N\}$, as shown in Lemma 2. Following the statement in Proposition 2, we can characterize the platform's equilibrium data acquisition by determining the optimal number of consumers to acquire data.

Next, we have $\frac{d\Delta\Pi(1)}{dK} = \frac{\alpha\sigma(N-1)v^2}{2t} > 0$ and

$$\frac{d\Delta\Pi(1)}{dK} > \frac{d\Delta\Pi(2)}{dK} > \cdots > \frac{d\Delta\Pi(N)}{dK}.$$

(i) If $\Delta\Pi(N)\big|_{K=1} \geq 0 \Leftrightarrow c \leq \underline{c}_p \equiv [2(2t-v)\alpha - v]\frac{\sigma v}{4t}$, the platform acquires data from all consumers when $K = 1$. Let us consider two cases.

- When $\frac{d\Delta\Pi(N)}{dK} \geq 0$, we can directly have $\frac{d\Delta\Pi(n)}{dK} \geq 0$ for all $n \in \{1, \ldots, N\}$ and thus
$$\Delta\Pi(1) \geq \Delta\Pi(2) \geq \cdots \geq \Delta\Pi(N) \geq 0,$$
which implies that the platform always acquires data from all consumers, $n^* = N$.
- When $\frac{d\Delta\Pi(N)}{dK} < 0$, there must exist a threshold $\tilde{n}$ such that
$$\frac{d\Delta\Pi(1)}{dK} > \cdots > \frac{d\Delta\Pi(\tilde{n}-1)}{dK} > 0 > \frac{d\Delta\Pi(\tilde{n})}{dK} > \cdots > \frac{d\Delta\Pi(N)}{dK},$$
which implies that for $n \in \{\tilde{n}, \ldots, N\}$, $\Delta\Pi(n)$ decreases with K. As K increases, n^* will decreases from N to $\tilde{n}$ if $\lim_{K\to+\infty}\Delta\Pi(\tilde{n}) \geq 0$

To delimit the existence conditions of the above two cases, we only need to show that $\frac{d\Delta\Pi(N)}{dK} < 0$ if and only if $N > N_p^*$. We have

$$\begin{aligned}\frac{d\Delta\Pi(N)}{dK} &= \frac{\sigma v^2}{4t}\left\{(1-\sigma)^{N-2}[2\alpha(1-\sigma)N - \sigma N + 1] - 1 - 2\alpha\right\} \\ &\propto (1-\sigma)^{N-2}[2\alpha(1-\sigma)N - \sigma N + 1] - 1 - 2\alpha \equiv H(N).\end{aligned}$$

We can prove that $H(N) < 0$ if and only if $N > N_p^* > 2$. First, given $H(2) > 0$, $\lim_{N\to\infty} H(N) < 0$ and the continuity of $G(N)$ on N, there must exist as least a threshold N^* such that $H(N^*) = 0$. Second, we have

$$\begin{aligned} H'(N) &= \{2\alpha(1-\sigma) - \sigma + \log(1-\sigma) + [2(1-\sigma)\alpha - \sigma]N\log(1-\sigma)\}(1-\sigma)^{N-2} \\ &\propto 2\alpha(1-\sigma) - \sigma + \log(1-\sigma) + [2(1-\sigma)\alpha - \sigma]N\log(1-\sigma), \end{aligned}$$

which decreases with N and must be negative when N is sufficiently high. This means that if $H'(2) < 0$, then $H'(N) < 0$ for all $N \geq 2$; if $H'(2) > 0$, then $H(N)$ can first increase and then decrease with N. Put together, the uniqueness of N_p^* can be established.

(ii) If $c > \underline{c}_p$, the platform does not acquire data when $K = 1$, which means that $\Delta\Pi(N)\big|_{K=1} < 0$. When $N \leq N^*$, we have $\frac{d\Delta\Pi(1)}{dK} > \frac{d\Delta\Pi(2)}{dK} > \cdots > \frac{d\Delta\Pi(N)}{dK} \geq 0$, which implies that for $n \in \{1, \ldots, N\}$, $\Delta\Pi(n)$ increases with K. As K increases, n^* will increase from 0 to n if $\lim_{K\to\infty} \Delta\Pi(n) \geq 0$. Similarly, When $N > N^*$, we have $\frac{d\Delta\Pi(1)}{dK} > \frac{d\Delta\Pi(2)}{dK} > \cdots > \frac{d\Delta\Pi(\tilde{n}-1)}{dK} \geq 0$, which implies that for $n \in \{1, \ldots, \tilde{n}-1\}$, $\Delta\Pi(n)$ increases with K. As K increases, n^* will increase from 0 to n if $\lim_{K\to\infty} \Delta\Pi(n) \geq 0$. As a result, if $\lim_{K\to\infty} \Delta\Pi(1) \geq 0 \Leftrightarrow c \leq \bar{c}_p = \frac{\sigma v^2}{4t}[2\alpha(K-1)N - 2\alpha K - 1] + \alpha\sigma v$. As K rises, n^* increases from 0 to a positive integer. Otherwise, if $c \geq \bar{c}_p$, $n^* = 0$ for any $1 \leq K < t/v$. □

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4):218–256, 2022.

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

S Nageeb Ali, Greg Lewis, and Shoshana Vasserman. Voluntary disclosure and personalized pricing. *The Review of Economic Studies*, 90(2):538–571, 2023.

Mark Armstrong. Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691, 2006.

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

Dirk Bergemann and Alessandro Bonatti. Selling cookies. *American Economic Journal: Microeconomics*, 7(3):259–294, 2015.

Dirk Bergemann and Alessandro Bonatti. Data, competition, and digital platforms. *American Economic Review*, 114(8):2553–2595, 2024.

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

Dirk Bergemann, Alessandro Bonatti, and Tan Gan. The economics of social data. *The RAND Journal of Economics*, 53(2):263–296, 2022.

David Bounie, Antoine Dubus, and Patrick Waelbroeck. Selling strategic information in digital competitive markets. *The RAND Journal of Economics*, 52(2):283–313, 2021.

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

Alexandre De Corniere and Romain De Nijs. Online advertising and privacy. *The RAND Journal of Economics*, 47(1):48–72, 2016.

Anthony Dukes and Lin Liu. Online shopping intermediaries: The strategic design of search environments. *Management Science*, 62(4):1064–1077, 2016.

Itay P Fainmesser, Andrea Galeotti, and Ruslan Momot. Digital privacy. *Management Science*, 69(6):3157–3173, 2023.

Avi Goldfarb and Verina F Que. The economics of digital privacy. *Annual Review of Economics*, 15:267–286, 2023.

Harish Guda and Upender Subramanian. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*, 65(5):1995–2014, 2019.

Andrei Hagiu and Bruno Jullien. Why do intermediaries divert search? *The RAND Journal of Economics*, 42(2):337–362, 2011.

Andrei Hagiu, Tat-How Teh, and Julian Wright. Should platforms be allowed to sell on their own marketplaces? *The RAND Journal of Economics*, 53(2):297–327, 2022.

Sinem Hidir and Nikhil Vellodi. Privacy, personalization, and price discrimination. *Journal of the European Economic Association*, 19(2):1342–1363, 2021.

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

Shota Ichihashi. The economics of data externalities. *Journal of Economic Theory*, 196: 105316, 2021b.

Shota Ichihashi. Dynamic privacy choices. *American Economic Journal: Microeconomics*, 15(2):1–40, 2023.

T Tony Ke and K Sudhir. Privacy rights and data security: Gdpr and personal data markets. *Management Science*, 69(8):4389–4412, 2023.

T Tony Ke and Yuting Zhu. Cheap talk on freelance platforms. *Management Science*, 67 (9):5901–5920, 2021.

T Tony Ke, Song Lin, and Michelle Y Lu. Information design of online platforms. *HKUST Business School Research Paper*, 070(2022), 2022.

T Tony Ke, Monic Sun, and Baojun Jiang. Peer-to-peer markets with bilateral ratings. *Marketing Science*, 2024.

Fei Long and Wilfred Amaldoss. Self-preferencing in e-commerce marketplaces: The role of sponsored advertising and private labels. *Marketing Science*, 2024.

Fei Long and Yunchuan Liu. Platform manipulation in online retail marketplace with sponsored advertising. *Marketing Science*, 43(2):317–345, 2024.

Jeanine Miklós-Thal, Avi Goldfarb, Avery Haviv, and Catherine Tucker. Frontiers: Digital hermits. *Marketing Science*, 2024.

Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2007.

Lynne Pepall, Dan Richards, and George Norman. *Industrial organization: Contemporary theory and empirical applications*. John Wiley & Sons, 2014.

Kun Qian and Sanjay Jain. Digital content creation: An analysis of the impact of recommendation systems. *Management Science*, 2024.

Jean-Charles Rochet and Jean Tirole. Two-sided markets: A progress report. *The RAND Journal of Economics*, 37(3):645–667, 2006.

Gleb Romanyuk and Alex Smolin. Cream skimming and information design in matching markets. *American Economic Journal: Microeconomics*, 11(2):250–276, 2019.

Bruce Schneier. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015.

Zijun Shi, Kannan Srinivasan, and Kaifu Zhang. Design of platform reputation systems: Optimal information disclosure. *Marketing Science*, 42(3):500–520, 2023.

Yue Wu, Kaifu Zhang, and V Padmanabhan. Matchmaker competition and technology provision. *Journal of Marketing Research*, 55(3):396–413, 2018.

Kai Hao Yang. Selling consumer data for profit: Optimal market-segmentation design and its consequences. *American Economic Review*, 112(4):1364–1393, 2022.

Zemin Zhong. Platform search design: The roles of precision and price. *Marketing Science*, 42(2):293–313, 2023.

Bo Zhou and Tianxin Zou. Competing for recommendations: The strategic impact of personalized product recommendations in online marketplaces. *Marketing Science*, 42(2):360–376, 2023.