연세대학교 상경대학
경제연구소
Economic Research Institute
Yonsei University

서울시 서대문구 연세로 50
50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
TEL: (+82-2) 2123-4065 FAX: (+82-2) 364-9149
E-mail: yeri4065@yonsei.ac.kr
http://yeri.yonsei.ac.kr/new

Monopoly Data Sales over Information-Sharing Networks

Jihwan Do
Yonsei University

Lining Han
Wuhan University

Xiaoxi Li
Wuhan University

April 2025
2025RWP-249

Economic Research Institute

Yonsei University

Monopoly Data Sales over Information-Sharing Networks*

Jihwan Do †1, Lining Han ‡2, and Xiaoxi Li §3

[1]School of Economics, Yonsei University
[2,3]Economics and Management School, Wuhan University

April, 2025

Abstract

This paper studies the monopoly data seller's problem when users are connected through an information-sharing network. When users' prior information is sufficiently noisy, the seller's optimal strategy targets a maximum independent set – the largest subset of users with no direct links. In this regime, data precision falls as the network becomes denser, yet we show – using the Caro–Wei bound, a classical result in graph theory – that it remains strictly above the socially efficient level in most networks. Further, any core–periphery network is Pareto-efficient, and any Pareto-efficient network exhibits a quasi-core-periphery structure. When users can coordinate network formation, the resulting equilibrium network also takes this form. Finally, we quantify the value of network information by comparing the seller profit to that with a misbelief.

Keywords: Information markets, information-sharing network, monopolistic pricing, maximum independence set

---

*Thanks to seminar participants at UHM, FDU, DKU for useful comments. Financial Supports from the Ministry of Education of China [Grant No.19YJC790033] and NSFC [Grant No.72273101, No.72473109] are greatly appreciated. All errors are our own.

†Email address: econ.jihwando@yonsei.ac.kr
‡Email address: hanlining@whu.edu.cn
§Email address: xxleewhu@gmail.com

1 Introduction

The rapid commoditization of data has transformed industries – from finance and healthcare to retail analytics and digital marketing – as firms invest heavily in collection, processing, and AI-driven analysis. Major platforms like WRDS, Bloomberg, and Datastream now generate substantial revenues by selling raw feeds, structured datasets, and customized analytics to academic and corporate clients. Yet the study of monopolistic pricing problems in this growing market calls for new theoretical examinations that take the following critical factor into consideration: buyers are embedded in social and professional networks that allow nearly costless information diffusion. For example, academic researchers post updates on personal sites or at conferences, and employees often share licensed access with colleagues, eroding the exclusivity that data sellers strive to monetize.[1]

The networked diffusion of data products presents a significant challenge for a monopolistic seller: How should one optimally target and price an information product when potential buyers can freely share signals with their peers?[2] As with many software products, data products incur negligible marginal costs, incentivizing the seller to maximize distribution. However, when information diffuses through a network, users may obtain substantial value from peer sharing rather than direct purchase. Traditional pricing theory suggests that a monopolist should segment the market by targeting high-value customers with differential pricing. Yet in a networked environment, consumers' willingness to pay is endogenously shaped by their network position and the purchasing behavior of others, complicating the design of optimal pricing strategies.

Furthermore, the efficiency loss associated with monopolistic pricing – a well-documented concern in traditional commodity markets – also arises in data markets. The presence of peer-to-peer information sharing introduces new dynamics that traditional welfare analysis may fail to capture. These network effects necessitate a more nuanced, quantitative assessment of welfare that accounts for both the diffusion process and the demand shaped by network interactions.

To study these questions, we develop a stylized monopoly-pricing model of data products over a network in which:

- Data quality is represented by a costly choice of precision in the seller's signal.
- Price discrimination is feasible – that is, users connected differently in the network may face different prices – while the quality of the data product remains homogeneous. [3]

[1]Data sellers routinely deploy legal and IP controls, underscoring the prevalence of informal sharing.

[2]While data privacy remains a sensitive and critical issue in these business models, it is not addressed in this paper. Instead, we assume that the data being collected, processed, and sold is free from disputes—either obtained through mutual agreements or publicly available but costly to collect.

[3]This occurs when an informational product for sale cannot be further divided or customized (e.g., a consumer

- Users' payoffs depend on actions under uncertainty about a common underlying state; additional signals improve payoffs via reduced error.
- Information sharing occurs along an undirected information-sharing network: if a user buys a signal, it is freely observed by her neighbors in the network.

In this framework, our first main result characterizes the monopolist's optimal targeting rule when users' common prior is sufficiently noisy. We show that the seller should sell exclusively to a maximum independent set of the network, that is, the largest set of users among whom there is no direct information-sharing link. Intuitively, by avoiding pairs of linked buyers, the seller prevents free-riding and maintains high marginal valuations for its signal. The elegance of this finding is that once one knows the network's independence number m, all of the key aspects of the optimal contract, such as prices, optimal precision, and profits, are determined in closed form as simple functions of m, the prior precision, and the seller's cost parameter. Though differential pricing is a usual method for market segmentation and is also allowed in our model, the optimal contract can be characterized by a unique price level for all users independent of their network locations. The group of target users who choose to purchase the data product is then screen out by their locations instead of differential prices.

While both the monopoly-optimal and the socially efficient data precision levels decrease as the network becomes denser (and m decreases), we show that there is in general a substantial efficiency gap between this outcome and the social planner's benchmark. Our second main contribution leverages a classical graph-theoretic bound of the Caro–Wei theorem, which asserts that for any network,

$$m \geq \sum_{i \in N} 1/(n_i + 1),$$

where n_i denotes user i's degree. From the planner's perspective, data is provided to all users, so the socially efficient precision equates total marginal benefit (weighted by degrees) to marginal cost and is asymptotically proportional to $\sum_{i \in N} 1/(n_i + 1)$. The gap

$$m - \sum_{i \in N} 1/(n_i + 1)$$

thus quantifies the extent to which monopoly provision remains too precise (from a welfare standpoint, thereafter the welfare loss) except when the network is a disjoint union of cliques, the only scenario in which the Caro–Wei bound is tight. This finding illuminates a general principle: by targeting a disconnected user set, the seller deliberately curtails information spillovers, preserving high marginal values and extracting greater surplus, but at the cost of over-precision relative to the social optimum.

trend report from a market research firm), or in markets such as newspapers or academic journals, where different users pay different prices for identical content. In this sense, we view information products with varying qualities as distinct commodities and focus on one particular product.

As exhibited in the main results, fundamental conflicts arise between the monopoly-optimal contract and socially efficient provision. We further look for Pareto improvements (efficiency) by exogenously regulating the network connections or exploring the users' endogenous formation of the network itself. We show that the set of Pareto-efficient networks coincides with those exhibiting a quasi-core–periphery structure. In these networks, a well-connected "core" of free-riders is linked to a "periphery" of purchasers, but core nodes need not be fully interconnected among themselves. Importantly, if users endogenously form the information-sharing network to maximize their payoffs, while anticipating the seller's strategy, then the resulting equilibrium network once again exhibits the quasi–core–periphery structure.

Finally, we quantify the value of network information to the monopolist by comparing profits under perfect knowledge of the network (actually the only necessary network statistic is the independence number m) with those obtained under two natural misbeliefs, assuming instead that the network is completely empty or fully connected. Our analysis shows that wrong network information can lead to substantial profit losses (unbounded under the misbeliefs as network size grows), which implies that data sellers would have a high willingness to pay for users' network information, especially for their independence number m.

The organization of the paper is as follows. Section 2 discusses the related literature. Section 3 presents the monopoly-pricing model and users' decision problem. In Section 4 we solve the problem and present our main results, i.e. characterizing the optimal monopoly contract and comparing it with the social planner's solution, leveraging the Caro–Wei bound to quantify the efficiency gap. In Section 5 we discuss the Pareto-efficient and network-formation equilibrium network structures, establishing the quasi-core–periphery characterization; and assess the value of network information. Section 6 concludes with directions for future research.

2 Related Literature

This article contributes to three strands of literature: network economics, markets for information, and information economics.

First of all, it advances the literature on network economics, particularly studies on optimal pricing and targeting strategies within network structures. Bloch (2016) provides a comprehensive survey of targeting and seeding problems in networks. Notably, Candogan et al. (2012) and Bloch and Quérou (2013) are closely related to our work, as they examine optimal pricing strategies for monopolistic sellers operating in networked environments. However, unlike their models, which focus on traditional goods in social networks, we investigate the pricing of information products – goods that differ from traditional ones in several important respects.

First, the unit cost of production reflects "quality" rather than "quantity," as there is no replication cost when selling an information product to additional users. Second, information products are non-rivalrous: a user's data can be shared with neighbors at no cost and without degradation in quality. This contrasts with traditional goods, where, even if a user benefits from neighbors' consumption due to network effects, it cannot fully substitute for their own consumption. Finally, a user's willingness to pay for an information product decreases as more neighbors consume it, implying that purchasing behavior exhibits strategic substitutability.

Fainmesser and Galeotti (2016) studies the value of information under network effects, assuming that the seller knows only the average characteristics of a random network but lacks local information about individual users. Under such conditions, the seller is restricted to a uniform pricing strategy. Chen et al. (2018), by contrast, examine competitive (rather than monopolistic) pricing in social networks. Other studies on network games involving strategic substitutes include Bramoullé and Kranton (2007), Bramoullé et al. (2014), and Allouch (2015).

Our results also relate to the literature on persuasion within networks (Egorov and Sonin, 2020; Candogan, 2022). For example, Egorov and Sonin (2020) show that optimal propaganda efforts may target peripheral rather than central users – a finding that parallels our result in which the optimal selling strategy corresponds to a maximum independent set of users in the social network.

Next, this paper contributes to the literature on markets for information, with a focus on data pricing and collection. Bergemann and Bonatti (2019) offer a comprehensive survey of research on information markets and present a unified modeling framework based on a stylized quadratic-Gaussian payoff structure for costly information acquisition and its role in decision-making. Acemoglu et al. (2022), Choi et al. (2019), and Bergemann et al. (2022) investigate issues related to online platforms and data brokers that collect consumer information and resell it to downstream firms, which then tailor products to consumer preferences. Our model abstracts from the product market and instead focuses solely on the information market, where information is treated as the final product. This information need not be limited to consumer data – it may include industrial trends, macroeconomic indicators, or other forms of data that provide actionable insights for decision-making.

While prior studies emphasize correlated demand and information leakage among consumers, leading to network effects, our approach highlights a different dimension by modeling information sharing among neighboring users within a social network. In models where online platforms act as data intermediaries (Acemoglu et al., 2022; Choi et al., 2019; Bergemann et al., 2022), concerns typically center around excessive data collection or inefficient privacy leakage used for price discrimination. We obtain a parallel result: the monopolistic seller overproduces information to improve data quality, driven by the incentive to target a maximum subset of users and thus limit downstream information sharing. Ichihashi (2021) examine competition among data intermediaries, while Jones and Tonetti (2020) incorporate data intermediaries into a growth model. Related to our work, Admati and Pfleiderer (1986, 1990) study the sale of information by a monopolist to potential traders in a speculative market

– focusing on price discrimination in the absence of a network structure, where agents form a continuum.

Finally, this paper contributes to the literature on information economics, particularly in the areas of information acquisition and sharing. Decision-making under uncertainty about an unknown state of the world lies at the core of research on costly information acquisition and sharing. Myatt and Wallace (2012, 2018) examine information acquisition in coordination games and oligopolistic price-setting problems. More recent studies explore information acquisition through networks in a variety of contexts, including network formation (Galeotti and Goyal, 2010; Herskovic and Ramos, 2020), network games (Myatt and Wallace, 2019; Leister, 2020), asset pricing (Ozsoylev and Walden, 2011; Han and Yang, 2013), and laboratory experiments (Goyal et al., 2017; Halim et al., 2019). In related work, Bloch et al. (2018) analyze the diffusion of rumors in social networks, while Gal-Or (1985), Li (2002), and Liu et al. (2021) study information sharing among retailers in supply chains.

The main distinction of our model from prior studies is that consumers do not acquire information independently. Rather, they purchase access to data collected by a monopolistic seller. Our analysis centers on the seller's joint decision-making over data quality and pricing strategies, taking into account the presence of a social network through which users may share information.

3 Model

We consider a simple model of a monopolistic seller's pricing problem involving a finite set of data users, $N = \{1, 2, ..., n\}$. Each user i has a decision-making problem whose payoff is determined by his action $a_i \in \mathbb{R}$ and unknown fundamental state $\theta \in \mathbb{R}$ as follows:

$$u_i (a_i, \theta) = -(a_i - \theta)^2 ,$$

where θ is unknown (random variable) following a normal distribution $N\left(0, \frac{1}{z_0}\right)$ with $z_0 > 0$.[4] Users share a common prior over θ. With this prior only, each decision maker's optimal action is to take the conditional expectation of the fundamental, which is 0. To improve expected payoffs, users may seek additional information, and a data market is assumed to meet this demand. The monopolistic data seller is capable of collecting and processing information concerning the fundamental state θ, which comes at a cost. The information product varies in its quality or data volume and is characterized by a signaling mechanism of θ. The seller can make a free copy of the data set and sell it to user i at the price p_i. With the product being purchased, user i receives a private signal about θ in the following form,

$$s_i = \theta + \epsilon_i,$$

[4]This modeling approach using the Gaussian environment is commonly used in the literature (e.g., Myatt and Wallace (2019); Herskovic and Ramos (2020); Acemoglu et al. (2022)).

where $\epsilon_i \sim N\left(0, \frac{1}{z}\right)$ is independent of θ and ϵ_j for all $j \neq i$. The information product's quality is actually measured by the signaling mechanism's precision z.

We use $C_i = (z, p_i)$ to denote the contract offered by the data seller to user i. Users are homogeneous, so they have the same learning ability and shall generate the signals with the same precision if they receive the same data.

There are two crucial aspects of our model. First, users share information through connections within a social network. Specifically, users are embedded in an undirected social network G, where $g_{ij} = g_{ji} = 1$ if i and j are connected, and $g_{ii} = 1$ for all i. Users i and j can observe each other's signals if and only if they are connected. [5] The set of the neighbors of i and its cardinality are respectively denoted by N_i and n_i, where $i \notin N_i$. Also, we will often denote $N \setminus \{i\}$ as $-i$ for brevity.

Second, the seller may offer different prices to different users (i.e., price discrimination is feasible), but must provide a uniform data quality (i.e., quality discrimination is not feasible). This happens when an informational product for sale cannot be further divided or refined (e.g., a consumer trend report from a market research firm), or more reasonably describes certain markets (e.g., newspapers and academic journals).

Since the uniform z measures the amount of data collected by the seller, we refer to that as the seller collecting data with precision z.

The timeline of this sequential decision problem is summarized as follows:

1. The seller simultaneously offers $C = (C_1, C_2, \ldots, C_n)$, which are publicly observed.
2. Each user i simultaneously chooses whether to accept or reject C_i.
   - If user i accepts C_i, user i and all of i's neighbors $\{j : j \in N_i\}$ will observe s_i.
   - If user i rejects C_i, s_i does not exist (not observable by any user).
3. Each user i chooses a_i.

Though the data quality is assumed to be uniform across users on the social network, the seller is allowed to set different prices for different users on the social network. It turns out to be the data seller's objective to target a subset of users who actually end up purchasing the information product. For those users who are not targeted, the seller could charge a price beyond their willingness to pay. For notational convenience, we denote $p_i = \infty$ for this case, and any $p_i < \infty$ is appropriately set such that user i shall choose to accept the offer. We are now able to introduce the following notation for the target set of users: for any set of contract offers $C = (z, p_i)_{i \in N}$,

$$M(C) \equiv \{i \in N | p_i < \infty\}.$$

[5] Since signal s_i's shock ϵ_i and s_j's are independent and decision makers' (users') decision-making problems are also independent, it is equivalent in our model for a signal to be shared and the data set that generates the signal to be shared with user j's neighbors.

For any $j \notin M(C)$, i.e. $p_j = \infty$, we let $s_j = \emptyset$. With this help, the seller's profit under $C = (z, p_i)_{i \in N}$ is given by

$$\Pi(C) = \sum_{i \in M(C)} p_i - \gamma z,$$

where γz is the total cost incurred for the seller to collect and process data for some $\gamma > 0$. On the other hand, given the offered contract $C = (z, p_i)_{i \in N}$ to users, each user i first determines if to accept this offer, i.e. buying the signaling mechanism at price p_i; and then, upon receiving the signal s_i of her/his own, and other signals that are free-rides from neighbors who bought, user i maximizes the expected payoff by choosing an action a_i to best meet the conditional expectation of the fundamental state θ. We write the induced expected payoff at the optimum as

$$\pi_i(C) = \max_{a_i} E\Big[E\left[-\left(a_i - \theta\right)^2 | s_i, s_j : j \in M(C) \cap N_i\right]\Big] - p_i,$$

if $C_i = (z, p_i)$ is accepted by user i, and otherwise,

$$\pi_i(C) = \max_{a_i} E\Big[E\left[-\left(a_i - \theta\right)^2 | s_j : j \in M(C) \cap N_i\right]\Big].$$

4 Analysis

4.1 Consumer Behaviors

We begin by analyzing a user's optimal decision-making problem given the offered contract $C = (z, p_j)_{j \in N}$.

Lemma 1. The optimal action of user i is given by

$$a_i^*(C) = \sum_{j: p_j < \infty} \left(\frac{z g_{ij}}{z_0 + z \sum_{j: p_j < \infty} g_{ij}} \right) s_j.$$

Given this, user i's expected payoff when accepting C_i is

$$\pi_i(C) = \max_{a_i} E\Big[E\left[-\left(a_i - \theta\right)^2 | s_i, s_j : j \in M(C) \cap N_i\right]\Big] - p_i.$$

If user i rejects C_i, the signal s_i is no longer available. However, user i can still observe the purchased signals of neighbors. Therefore, the expected payoff conditional on received signals, given the rejection of C_i, is

$$\pi_i(C) = \max_{a_i} E\Big[E\left[-\left(a_i - \theta\right)^2 | s_i, s_j : j \in M(C) \cap N_i\right]\Big]$$

The following lemma characterizes the maximum price that user i is willing to pay for access to the data and to receive an extra signal.

Lemma 2. Given $C = (z, p_j)_{j \in N}$, user i is willing to accept C_i if and only if

$$p_i \leq \frac{1}{z_0 + z\sum_{j \neq i: p_j < \infty} g_{ij}} - \frac{1}{z_0 + z\sum_{j: p_j < \infty} g_{ij}} = \frac{z}{(z_0 + z\sum_{j \neq i: p_j < \infty} g_{ij})(z_0 + z\sum_{j: p_j < \infty} g_{ij})}.$$

When there is only a single user, the optimal contract can be derived directly from the above lemma.

Example 1. In the case where $n = 1$, the seller's profit is given by $\Pi(C) = \frac{1}{z_0} - \frac{1}{z_0+z} - \gamma z$. Applying the first-order condition yields $z + z_0 = \frac{1}{\sqrt{\gamma}}$. Thus, in the optimal contract, the data price is $p_1 = \frac{1}{z_0} - \frac{1}{z_0+z} = \frac{1}{z_0} - \sqrt{\gamma}$ $\left(\text{assuming } z_0 < \frac{1}{\gamma}\right)$, whereas the optimal signal precision is $z = \frac{1}{\sqrt{\gamma}} - z_0$. $\diamond$

As illustrated in this example, a user has no incentive to purchase access to data and acquire an additional signal if the common prior is sufficiently accurate $\left(\text{i.e. } z_0 < \frac{1}{\gamma}\right)$. In this case, the contract offered by the seller becomes trivial. This observation holds generally, and we henceforth restrict attention to settings in which the common prior is sufficiently noisy to ensure that the optimal data quality is strictly positive.

4.2 Monopoly Analysis

4.2.1 Optimal Monopoly Contract

Recall G denotes an undirected network and N the set of users. Given γ the marginal cost of collecting data, and z_0 the precision of the common prior regarding the fundamental state, the data seller offers a contract $C = (z, p_i)_{i \in N}$ to users in the social network G in order to maximize the payoff $\Pi(C)$. As previously discussed, the seller's problem involves determining the quality of the information product's quality (i.e. data precision z), and selecting a subset of users, $M(C)$, who ultimately choose to purchase access to the dataset.

In what follows, we write $M_i \equiv M(C) \cap N_i$ to denote the subset of user i's neighbors who purchase the information product from the seller, $m = |M(C)|$ and $m_i = |M_i|$. Under these notations, designing an optimal contract C is equivalent to choosing (i) the target set $M(C)$, (ii) the precision of database z, and (iii) the prices $(p_i)_{i \in M(C)}$ for the target users, in a way that jointly maximizes the seller's profit conditional on user acceptance.

From Lemma 2, if the contract C is optimal for the seller and accepted by any user $i \in M(C)$, the price charged to user i is given by:

$$p(m_i, z) \equiv \frac{1}{z_0 + m_i z} - \frac{1}{z_0 + (m_i + 1)z}. \tag{1}$$

From equation (1), it is straightforward to observe that the price charged to user i decreases as m_i, the cardinality of neighbors in the target set, increases. Intuitively, as a user observes an increasing number of signals from her neighbors, the marginal benefit of purchasing access and generating her own signal diminishes. Consequently, the seller faces a trade-off between targeting a larger set of users with a lower price per data access and focusing on a smaller group of users at a higher price.

Before continuing, we introduce several notions of independent sets, which will play crucial roles in our analysis.

Definition 1. A set of nodes X is an *independent set* of the network G, if no two nodes in X are connected, i.e. $\forall i, j \in X$ with $i \neq j$, $g_{ij} = 0$. An independent set of maximum cardinality is called a *maximum independent set*.

The cardinality of the maximum independent set is referred to as the independence number of the network G.

An independent set of a network is a subset of nodes such that no pair of nodes within the subset is connected. The following result characterizes a condition under which the seller would not offer the information product to any two connected users in the optimal contract.

Proposition 1. Suppose $M(C)$ is not an independent set of G. Then, C is not optimal for the seller in G if $z > \min_{i \in M(C)} \frac{z_0}{1+m_i}$.

Proposition 1 provides a necessary condition for the contract C targeting a non-independent set to be optimal, by establishing an upper bound on the precision of the signal. If the precision exceeds this bound, the seller can increase their payoff by removing a user from the target set $M(C)$. When the data precision is sufficiently high, the marginal benefit of purchasing access to the data declines rapidly as more neighbors buy the information, making it unprofitable for the seller to sell to two connected users.

In general, the seller must simultaneously determine the precision of the data, the target users, and the prices. The main result is presented in the following theorem.

Theorem 1. Suppose $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Then, the optimal contract in G is characterized as follows:

(i) $M(C)$ is a maximum independent set of G.

(ii) $z = \sqrt{\frac{m}{\gamma}} - z_0$ and $p_i = \frac{1}{z_0} - \sqrt{\frac{\gamma}{m}}$ for all $i \in M(C)$, where m is the independence number of the network G.

Theorem 1 demonstrates that the optimal contract targets disconnected users when the common prior of the users is sufficiently noisy. Also, given γ, the marginal cost of data collection, and z_0, the precision of the common prior, the data precision depends only on m, the independence number of the network G. Both the optimal data precision z and the charged price p_i increase as the network's independence number increases, or as the precision of the common prior and the marginal cost of data collection decrease. Figure 1 provides examples of target users $M(C)$ in various networks.

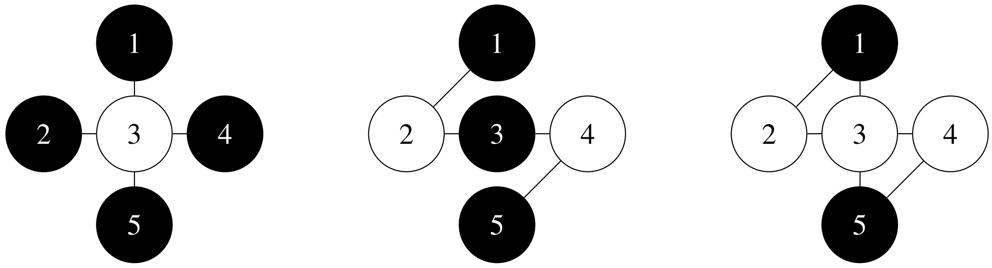


Figure 1: Examples of $M(C)$ (shaded circles represent targeted users).

Remark 1. One implication of Theorem 1 is that as the network G becomes denser – i.e., as more links are added without removing existing ones – the independence number m generally decreases, leading to a lower monopoly-optimal data precision.

Remark 2. The bound on z_0 in Theorem 1, $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$, can be replaced by a uniform one $z_0 < \frac{1}{4\sqrt{\gamma}}$ such that the result holds for a network G of any size n. This follows from the fact that $\frac{n+1}{2n+1}$ is decreasing to $\frac{1}{2}$ as n becomes larger and larger.

Remark 3. Antitrust regulations may impose stringent limits on price discrimination. In our monopoly-pricing framework, we can actually show that, uniform pricing can be sustained in some equilibrium, so price discrimination is not necessary. When the monopolist offers the optimal contract $(z, p) = \left(\sqrt{1/\gamma} - z_0), 1/z_0 - \sqrt{\gamma/m}\right)$ as described in Theorem 1, those target users are exactly indifferent between accepting and rejecting – so long as non-target users do not buy. Conversely, if all target users purchase, non-target users find that free-riding delivers a strictly higher payoff and thus choose not to purchase. Hence, the mechanism endogenously screens the target group by their locations rather than by charging different prices.

Remark 4. Theorem 1 also implies that, to attain maximal profit, the monopolist need not know the full network topology – knowing only the independence number m is sufficient to determine all aspects of the optimal contract. This is a sharp result which has strong policy implications for the monopoly's optimal pricing problem: choosing an efficient approximation algorithm to compute m (the problem itself is NP hard) might dominate the strategy of looking for the detailed network structure, especially when the later is over costly or may violate some privacy-protection policy. The result also verifies that developing efficient algorithm for computing networks' independence number has very concrete commercial values.

Note that a maximum independent set of a complete network contains only one node, whereas in a star network, it consists of the peripheral users. Theorem 1 leads to the following characterization of the seller's optimal target set in these special network structures.

Corollary 1. Suppose $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Then, in the optimal contract for G,

(i) If G is a complete network, the seller sells to only one user.

(ii) If G is a star network with user i as the center, the seller sells to all users except i.

4.2.2 Welfare under Monopoly

Given the characterization of Theorem 1, the seller's profit under the optimal contract is given by

$$\Pi\left(z_0, G\right) = \frac{m}{z_0} + \gamma z_0 - 2\sqrt{\gamma m}$$

where m is the cardinality of a maximum independent set of G.

Proposition 2. Assume $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$. Then, the seller's profit $\Pi(z_0, G)$ is strictly increasing in m and strictly decreasing in z_0.

It is worth emphasizing that the seller's profit depends solely on the cardinality of the maximum independent set: if two networks G and G' have the same maximum independent set size, then the seller's profits are identical. Furthermore, for a given network G, the profit is invariant to the choice of target set $M(C)$ as long as it constitutes a maximum independent set. This stands in stark contrast to user surplus, which can vary significantly depending on the network structure and the target set $M(C)$, even when the network G is fixed. To illustrate this, note that each user $i \in M(C)$ obtains a payoff of. To see this, observe first that each user $i \in M(C)$ obtains a payoff of

$$-\left(\frac{1}{z_0 + z}\right) - p_i = -\frac{1}{z_0},$$

whereas each free-rider $i \in N \setminus M(C)$ obtains a payoff of

$$-\frac{1}{z_0 + m_i z} = -\frac{1}{z_0 + m_i\left(\sqrt{\frac{m}{\gamma}} - z_0\right)}.$$

Thus, we augment the analysis by computing the consumer surplus, summarized in the following result:

Proposition 3. Assume $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$. Given C as the optimal contract and $M(C)$ as the associated maximum independent set for network G. The consumer surplus under $M(C)$ is

$$w\left(z_0, G, M(C)\right) \equiv -\sum_{i \in M(C)} \frac{1}{z_0} - \sum_{i \in N \setminus M(C)} \frac{1}{z_0 + m_i\left(\sqrt{\frac{m}{\gamma}} - z_0\right)}. \tag{2}$$

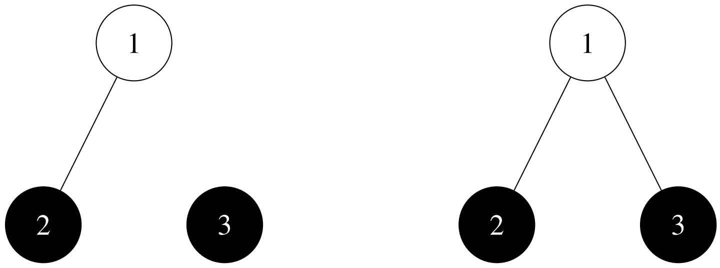


Figure 2: Same $M(C)$ with different consumer surplus: left: G; right: G'.

To examine how the network structure influences consumer surplus, consider the example in Figure 2, where users 2 and 3 are served by the seller, while user 1 free-rides.

Example 2. [6] Consider the set of users $N = \{1, 2, 3\}$ and two networks G and G' depicted in Figure 2. The only difference between them is that user 1 is connected only to user 2 in G, but to both user 2 and user 3 in G'. Formally, the only active links are $g_{12} = g'_{12} = g'_{13} = 1$. In both cases, the seller's optimal contract targets $M(C) = \{2, 3\}$, which constitutes the maximum independent set of both G and G'. Consequently, the contracts in both settings are identical: the precision of data collection is $z = \sqrt{\frac{2}{\gamma}} - z_0$, and the price charged is $p_2 = p_3 = \frac{1}{z_0} - \sqrt{\frac{\gamma}{2}}$. Thus, the seller's profits are equal in both networks, i.e., $\Pi(z_0, G) = \Pi(z_0, G')$.

Moreover, assuming $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$ (with $n = 3$), we can evaluate the consumer surplus from equation (2) in Proposition 3:

$$w(z_0, G, M(C)) = -\frac{2}{z_0} - \frac{1}{\sqrt{\frac{2}{\gamma}}} < w(z_0, G', M(C)) = -\frac{2}{z_0} - \frac{1}{2\sqrt{\frac{2}{\gamma}} - z_0}.$$

Therefore, the consumer surplus is higher under G' than under G, despite both networks having the same maximum independent set.

The next example, illustrated in Figure 3, demonstrates that consumer surplus depends on the selection of the target set: different choices of $M(C)$'s within a fixed network can yield different levels of consumer surplus.

Example 3. Consider the set of users $N = \{1, 2, 3, 4\}$ and the network G shown in Figure 3. It is straightforward to verify that the maximum independent sets are $\{1, 3\}$, $\{1, 4\}$, and $\{2, 4\}$. In the first case, users 1 and 3 are served (i.e. $M(C) = \{1, 3\}$), whereas in the second case, users 1 and 4 are served (i.e. $M(C') = \{1, 4\}$). Since both $M(C)$ and $M(C')$ are maximum independent sets of G, the seller achieves the same profit, that is maximal, $\Pi(z_0, G)$ in either case – under contracts C and C', respectively. The precision of the data is $z = z' = \sqrt{\frac{2}{\gamma}} - z_0$,

[6] We use $x \succ^{f.o.s.d.} y$ to denote "x first-order stochastically dominates y", and $x \succ^{s.o.s.d.} y$ to denote "x second-order stochastically dominates y".

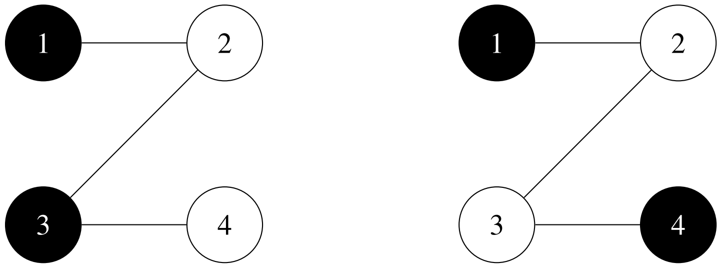


Figure 3: Different target sets with different consumer surplus: left: $M(C) = \{1,3\}$; right: $M(C') = \{1,4\}$. $F\big(\cdot, k(C)\big) = (1,1) \succ^{f.o.s.d.} (1,2) = F\big(\cdot, k(C')\big)$ and $w\big(z_0, G, M(C)\big) > w\big(z_0, G, M(C')\big)$.

and the prices are $p_1 = p_3 = p_1' = p_4' = \frac{1}{z_0} - \sqrt{\frac{\gamma}{2}}$. Assuming $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$ (here $n = 4$), we can compute the consumer surplus using equation (2) in Proposition 3:

$$w\big(z_0, G, M(C)\big) = -\frac{2}{z_0} - \frac{1}{2\sqrt{\frac{2}{\gamma}} - z_0} - \frac{1}{\sqrt{\frac{2}{\gamma}}} > w\big(z_0, G, M(C')\big) = -\frac{2}{z_0} - \frac{2}{\sqrt{\frac{2}{\gamma}}}.$$

Hence, the consumer surplus is higher under C than under C', which can be attributed to the number of free-riding opportunities. Under contract C', user 2 and user 3 each free-rides one signal – from user 1 and user 4, respectively. In contrast, under contract C, user 4 free-rides the signal of user 3, and user 2 benefits from signals generated by both users 1 and 3.

Consider C an optimal contract, with the target set $|M(C)|$ being maximum independent. Recall that $m = |M(C)|$ is the independence number of the network G, and $m_i = |M_i| = |M(C) \cap N_i|$ represents the number of signals that user i can free-ride under C. Given that $M(C)$ is the target, there are a total of $n - m$ free-riders in the network. Without loss of generality, we relabel the nodes of network G such that free-riders are indexed by $1, 2, ..., n-m$ ordered according to the number of accessible signals: $m_1 \geq m_2 \geq \cdots \geq m_{n-m}$.

We further introduce the vector of m_i values:

$$k(C) \equiv (m_1, m_2, \cdots, m_{n-m}),$$

and define the cumulative distribution function (CDF):

$$F(x|k(C)) = \frac{|\{j|m_j \leq x, 1 \leq j \leq n-m\}|}{n-m}.$$

With these notations, we present the following result, which provides sufficient conditions for comparing the welfare outcomes of two optimal contracts within the same social network.

Proposition 4. Let $M(C)$ and $M(C')$ denote the target sets of two optimal contracts C and C', respectively. Then, $w(z_0, G, M(C)) \geq w(z_0, G, M(C'))$ if either of the following holds:

(i) $F(\cdot|k(C))$ first-order stochastically dominates $F(\cdot|k(C'))$.

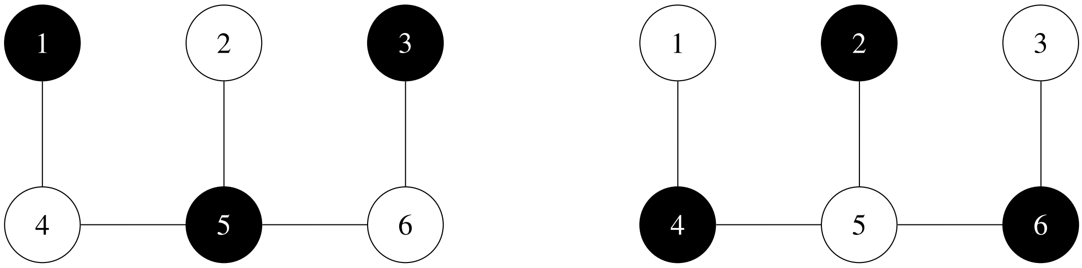


Figure 4: Different target sets with different consumer surplus: left: $M(C) = \{1, 3, 5\}$; right: $M(C') = \{2, 4, 6\}$. $F\big(\cdot, k(C)\big) = (2, 2, 1) \succ^{s.o.s.d.} (3, 1, 1) = F\big(\cdot, k(C')\big)$ and $w\big(z_0, G, M(C)\big) > w\big(z_0, G, M(C')\big)$.

(ii) $F\left(\cdot|k(C)\right)$ second-order stochastically dominates $F\left(\cdot|k(C')\right)$.

Note that condition (i) in Proposition 4 is equivalent to $k(C) \geq k(C')$, meaning that each coordinate of $k(C)$ is greater than or equal to that of $k(C')$. In contrast, the second condition (ii) corresponds to the case where $k(C)$ is more symmetric than $k(C')$. These results follow from the observation that a free-rider i's payoff is increasing and concave in m_i, the number of signals (i.e. neighbors) from which i free-rides. The pattern observed in Figure 3 directly illustrates Proposition 4 (i). Figure 4 presents an example corresponding to Proposition 4 (ii), where $F\left(\cdot|k(C)\right)$ second-order stochastically dominates $F\left(\cdot|k(C')\right)$. One can also use the expression in Proposition 3 to directly verify that $w\big(z_0, G, M(C)\big) > w\big(z_0, G, M(C')\big)$.

4.3 Social Planner's Problem

Next, we consider the socially efficient contract that maximizes the following social welfare function:

$$SW(C) = \Pi(C) + \sum_{i \in N} \pi_i(C).$$

First, note that the total cost of providing data is independent of the number of users served. This suggests that the socially efficient contract should serve all users, i.e., $M(C) = N$. This can be confirmed by verifying the following inequality:

$$\begin{aligned}
SW(C) & \sum_{i \in M(C)} p_i - \gamma z + \sum_{i \in M(C)} \left(-\frac{1}{z_0 + (m_i + 1)z} - p_i \right) + \sum_{i \in N \setminus M(C)} \left(-\frac{1}{z_0 + m_i z} \right) \\
&= \sum_{i \in M(C)} \left(-\frac{1}{z_0 + (m_i + 1)z} \right) + \sum_{i \in N \setminus M(C)} \left(-\frac{1}{z_0 + m_i z} \right) - \gamma z \\
&\leq \sum_{i \in N} \left(-\frac{1}{z_0 + (n_i + 1)z} \right) - \gamma z.
\end{aligned}$$

Since $M(C) = N$, the social welfare is

$$\sum_{i \in N} \left(-\frac{1}{z_0 + (n_i + 1)z} \right) - \gamma z.$$

Thus, the precision z^* in the socially efficient contract is the value z^* that solves:

$$\sum_{i \in N} \left(\frac{n_i + 1}{(z_0 + (n_i + 1)z^*)^2} \right) = \gamma. \tag{3}$$

Further, one can show that z^* is decreasing in n_i. Indeed, by implicit function theorem,

$$\frac{\partial z^*}{\partial n_i} = \frac{\frac{z_0 - (n_i+1)z^*}{(z_0 + (n_i+1)z^*)^3}}{\sum_{i \in N} 2(n_i + 1)^2 (z_0 + (n_i + 1)z^*)^3},$$

which is negative when $z_0 \to 0$, since z^* is bounded in this process.[7]

The following proposition summarizes the characterization of the efficient contract chosen by the social planner.

Proposition 5. The socially efficient contract $C^* = (z^*, p_i)_{i \in N}$ is characterized as follows:

(i) $p_i^* < \infty$ for all $i \in N$, so the entire set of users is served, i.e. $M(C^*) = N$.

(ii) The optimal precision z^* solves equation (3).

(iii) z^* is decreasing in each n_i for sufficiently small z_0.

Remark 5. Part (iii) of Proposition 5 indicates that under a sufficiently noisy prior, the socially efficient data precision z^* decreases as the network G becomes denser—i.e., as additional links are added without removing any existing ones, thereby increasing some n_i.

4.4 Monopoly vs. Planner: Implications for Data Precision

We now compare the monopoly and social planner regimes to evaluate whether, from a social standpoint, the monopolistic data seller provides information that is excessively precise or

[7]To see this, observe that

$$\gamma = \sum_{i \in N} \frac{n_i + 1}{(z_0 + (n_i + 1)z^*)^2} < \sum_{i \in N} \frac{1}{(n_i + 1)z^*} = \frac{1}{z^*} \sum_{i \in N} \frac{1}{n_i + 1} \Rightarrow z^* < \frac{1}{\gamma} \sum_{i \in N} \frac{1}{n_i + 1}$$

and

$$\gamma = \sum_{i \in N} \frac{n_i + 1}{(z_0 + (n_i + 1)z^*)^2} > \frac{n_1 + 1}{(z_0 + (n_1 + 1)z^*)^2} \Rightarrow z^* > \frac{1}{n_1 + 1} \left(\sqrt{\frac{n_1 + 1}{\gamma}} - z_0 \right) \overset{\text{as } z_0 \to 0}{\to} \frac{1}{\gamma \sqrt{n_1 + 1}} > 0.$$

Therefore, $z^* \in \left(\frac{1}{\gamma \sqrt{n_i + 1}}, \frac{1}{\gamma} \sum_{i \in N} \frac{1}{n_i + 1} \right)$.

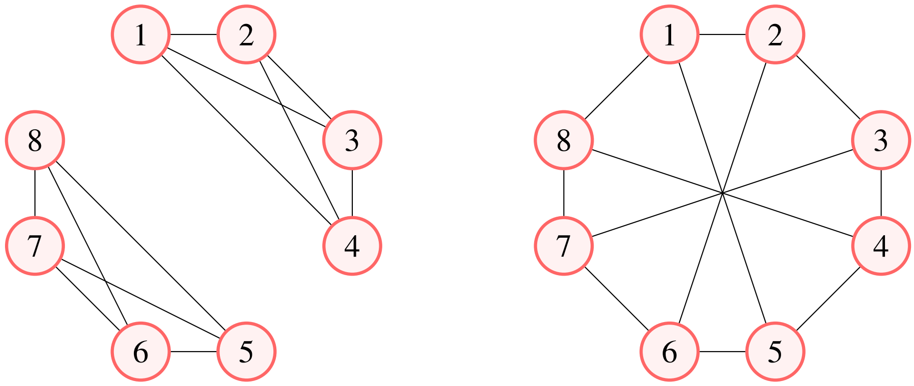


Figure 5: Different symmetric networks with $n = 8$ and $d = 3$: left: G_1; right: G_2.

insufficiently precise. In general, the socially efficient precision z^* does not admit a closed-form solution. To gain further insight and develop intuition, we first examine specific network structures that allow for an explicit derivation of z^*. We then compare it with the monopoly-optimal precision z^o in the following example. Throughout the analysis, we denote $\lfloor x \rfloor$ as the integer part of a real number x.

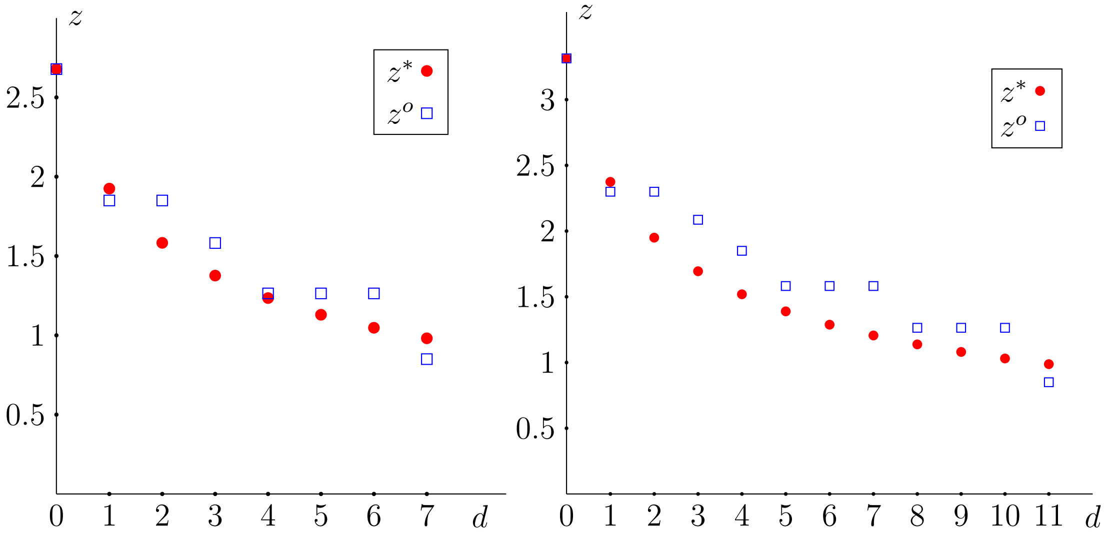
Example 4. Consider a network G in which every node has the same degree d, i.e., $n_i = d$ for all $i \in N$. Assume that $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$.

(i) If G is a cycle network, then $d = 2$ and the independence number is $m = \lfloor \frac{n}{2} \rfloor$. The socially efficient level of data precision is given by $z^* = \sqrt{\frac{n}{3\gamma}} - \frac{z_0}{3}$, while monopoly-optimal level of data quality is $z^o = \sqrt{\frac{m}{\gamma}} - z_0$ provided $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$. For values of z_0 such that $z_0 < \min\left\{\frac{3}{2\sqrt{\gamma}}\left(\sqrt{m} - \sqrt{\frac{n}{3}}\right), \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}\right\}$, we obtain $z^* < z^o$.

(ii) If G is a complete network with $n \geq 2$, then $d = n-1$ and $m = 1$. The socially efficient level of data precision is $z^* = \sqrt{\frac{1}{\gamma}} - \frac{z_0}{n}$, while monopoly-optimal level is $z^o = \sqrt{\frac{1}{\gamma}} - z_0$ for values of z_0 such that $z_0 < \frac{3}{2\sqrt{\gamma}}\left(\sqrt{m} - \sqrt{\frac{n}{3}}\right)$ provided $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$. In this case, for any $z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}$, we have $z^* > z^o$.

The example demonstrates that when the prior precision z_0 is sufficiently low, the monopoly-optimal data precision z^o may either exceed or fall short of the socially efficient level z^*. To further explore this relationship, we conduct a numerical analysis.

Numerical case We consider network sizes $n = 8$ or $n = 12$ —— even number for which a uniform degree $n_i = d$ across all $i \in N$ is feasible. We fix $z_0 = 0.15$ and $\gamma = 1$, ensuring the condition $z_0 < \min\{\frac{3}{2\sqrt{\gamma}}\left(\sqrt{m} - \sqrt{\frac{n}{3}}\right), \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1}\}$ holds. As in Example 4, we focus on symmetric networks where each node has the same degree d, and we plot the resulting values of z^* and z^o as d varies from 0 to $n-1$.



(a) $n = 8$, $z_0 = 0.15$, $\gamma = 1$ (b) $n = 12$, $z_0 = 0.15$, $\gamma = 1$

Figure 6: Numerical examples for comparison between z^* and z^o.

Note that for a given d, multiple network topologies may satisfy the uniform degree condition. For instance, when $n = 8$ and $d = 3$, more than one such network exists. Figure 5 provides two representative examples of networks satisfying these conditions:

(i) G_1: The network consists of two disconnected components, each of which forms a complete graph on 4 nodes.

(ii) G_2: The network is connected, and each node is linked to exactly 3 other nodes.

In this numerical study, we focus exclusively on networks of type G_2, excluding those of type G_1. The rationale is straightforward: in G_1-type networks, each disconnected component functions independently and can thus be analyzed separately. The numerical results are illustrated in Figure 6, and our key findings can be summarized as follows:

(i) When the network is empty (i.e., no connections among users, $d = 0$), we have $z^o = z^*$.

(ii) When the network is either a perfect matching (each user is connected to exactly one other user, $d = 1$) or a complete graph (each user is connected to all others, $d = n - 1$), it holds that $z^o < z^*$.

(iii) In all intermediate cases ($1 < d < n - 1$), the monopoly-optimal precision exceeds the socially efficient level, i.e., $z^o > z^*$.

As user connectivity increases, both the socially efficient level of data precision z^* and the seller-optimal level z^o decrease. This decline reflects the fact that each user's marginal valuation for additional precision diminishes as more information becomes available through network spillovers. Nevertheless, for most values of d, the data precision provided by the monopoly seller remains strictly higher than the socially efficient level. Our second main result formalizes this observation by characterizing the inefficiency of data precision under monopoly.

Theorem 2. Suppose that $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Let z^* be the socially efficient data precision and let z^o denote the precision chosen under the monopolist's optimal contract. Then for any network G the following holds:

(i) If G is a disjoint union of cliques, then we have $z^* \geq z^o$, with equality if and only if G has no connections (i.e. $n_i = 0$ for every $i \in N$).

(ii) If G is not a disjoint union of cliques, then for sufficiently small prior precision z_0, we have $z^* < z^o$.

This result implies that when the prior is sufficiently noisy, the monopolist provides excessively precise data to individual buyers from a social welfare perspective – except in the special case where the network consists of a disjoint union of cliques. Intuitively, this inefficiency arises because, unlike a social planner who internalizes the benefits of information sharing across the entire network, the monopolist optimally targets a disconnected subset of users. By doing so, the seller curtails information spillovers, thereby sustaining high marginal valuations for additional signals among the targeted users. This enables the monopolist to extract greater marginal revenue by offering higher data precision.

To make the result more precise, we briefly outline the core idea of the proof. Recall that the monopolistic seller targets a maximum independent set of the network, selling data to m users. Each buyer is charged a price equal to $\frac{1}{z_0} - \frac{1}{z+z_0}$. The seller's marginal profit with respect to data precision z is therefore given by

$$\frac{\partial \Pi}{\partial z} = \frac{m}{(z+z_0)^2} - \gamma \approx \frac{m}{z^2} - \gamma,$$

when the prior becomes highly noisy (i.e., as $z_0 \to 0$).

In contrast, a social planner provides data access to all users, with each user effectively receiving $n_i + 1$ signals in total. The utility derived from these signals is given by $-\frac{1}{(n_i+1)z+z_0}$, so the marginal contribution of increasing data precision to social welfare is

$$\frac{\partial SW}{\partial z} = \sum_{i \in N} \frac{n_i + 1}{\left[(n_i+1)z + z_0\right]^2} - \gamma \approx \sum_{i \in N} \frac{1}{(n_i+1)z^2} - \gamma,$$

here the approximation holds in the limit as $z_0 \to 0$. Under a noisy prior, the comparison between the socially efficient z^* and the monopolist's optimal choice of z^o thereafter asymptotically reduces to comparing the independence number m and the degree-weighted expression $\{n_i\}$, $\sum_{i \in N} \frac{1}{n_i+1}$.

Our analysis leverages a classical result from graph theory known as the Caro-Wei bound, which provides a lower bound on the independence number of a graph by the same degree-weighted expression as above. Specifically, it states that, for any network,

$$m \geq \sum_{i \in N} \frac{1}{n_i + 1}$$

where m denotes the size of a maximum independent set and n_i is the degree of node i. Equality holds if and only if the network is a disjoint union of cliques.

Case 1: When the network G is a disjoint union of cliques, we have equality in the Caro-Wei bound, i.e.,

$$m = \sum_{i \in N} \frac{1}{n_i + 1}.$$

In this case, the comparison between the monopolist's and the planner's incentives does not rely on the approximation as $z_0 \to 0$; instead, the common prior z_0 becomes the critical determinant. Our analysis shows that

$$\left.\frac{\partial SW}{\partial z}\right|_{z=z^o} \geq 0 = \left.\frac{\partial SW}{\partial z}\right|_{z=z^*},$$

for any $z_0 > 0$ with $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Hence, we deduce that $z^o \leq z^*$, with equality holding only in the trivial case where G is an empty graph.

Case 2: When G is not a disjoint union of cliques, the Caro-Wei bound is not tight:

$$m > \sum_{i \in N} \frac{1}{n_i + 1}.$$

In the limit as $z_0 \to 0$, we then obtain

$$\frac{\partial \Pi}{\partial z} > \frac{\partial SW}{\partial z}, \quad \forall z > 0.$$

In particular,

$$\left.\frac{\partial \Pi}{\partial z}\right|_{z=z^*} > 0 = \left.\frac{\partial SW}{\partial z}\right|_{z=z^*}.$$

This inequality implies that the monopolist has a strictly positive marginal incentive to further increase data precision at the socially optimal level z^*. The monopolist's strategy of targeting disconnected users preserves high marginal valuations, and this incentive dominates the planner's goal of improving data quality across the entire network. As a result, when the prior is sufficiently noisy, the monopolist provides overly precise data to each buyer, i.e., $z^o > z^*$.

Before concluding this section, we offer two remarks that shed further light on the implications of our analysis.

Remark 6. The result in Theorem 2 is closely tied to the Caro-Wei bound on the independence number in graph theory. The gap between the socially efficient level of data precision (asymptotically proportional to $\sqrt{\sum_{i\in N}\frac{1}{n_i+1}}$) and the monopolist-optimal level (asymptotically proportional to $\sqrt{m}$) can be decomposed into two components: one arising from the structural gap $\sqrt{\sum_{i\in N}\frac{1}{n_i+1}}-\sqrt{m}$, and the other from the common prior noise z_0. When the network is a disjoint union of cliques, the Caro-Wei bound is tight, so the difference in precision levels is governed solely by z_0. In contrast, when the network is not of this form, the structural gap dominates due to the looseness of the bound. To our knowledge, this connection between the Caro-Wei bound and socially efficient information provision in network settings is novel, and may offer fertile ground for future research at the interface of graph theory and economic networks.

Remark 7. Our findings underscore that the efficiency of information provision under monopoly, specifically, whether data is excessively or insufficiently precise from a welfare standpoint, can be highly sensitive to the underlying network of information-sharing links. This sensitivity is starkly illustrated by comparing a complete network G to a perturbed network $G-g_{ij}$, obtained by removing a single edge. Although these two networks are nearly indistinguishable in large populations, for any fixed number of users the consequences are dramatic: in the complete network G, the monopolist provides inefficiently imprecise data (i.e., $z^o < z^*$), whereas in the slightly perturbed network $G - g_{ij}$, the monopolist provides inefficiently precise data (i.e., $z^o > z^*$). This discontinuity highlights the intricate dependence of optimal contract design on the fine structure of the network.

5 Discussion

In the first subsection of this part, we investigate the *Pareto-efficient network structures* under the assumption that the seller adopts an optimal selling strategy. This analysis enables us to examine how modifying network connections can enhance users' welfare without reducing the seller's profit. We introduce a class of network structures—referred to as *quasi-core-periphery*—which fully characterizes the set of Pareto-efficient networks.

Next, we consider the case in which the social network forms endogenously among users prior to the pricing game. We show that the seller-optimal equilibrium network in this setting also corresponds to a quasi-core-periphery structure.

Finally, in the last subsection, we evaluate the value of network information from the seller's perspective. Specifically, we quantify the gain in monopolistic profit when the seller has perfect knowledge of the network structure, compared to when she operates under a misbelief – treating the network as either empty or complete.

5.1 Pareto Efficient Network

We fix an arbitrary network G and consider C to be a seller-optimal contract for G. Following our earlier notation, let $\pi_i(C)$ denote user i's expected payoff, taking into account their optimal decisions: whether to accept the contract C_i, and, upon acceptance, the optimal action $a_i^*(C)$ given the available signals. In this subsection, we write $\pi_i(C, G)$ instead of $\pi_i(C)$ to emphasize the dependence of user i's payoff on the network structure G. Since C is optimal, the seller's profit $\Pi(C)$ represents the maximum expected payoff achievable in the pricing game on network G. For consistency, we denote this as $\Pi(C, G)$ to highlight its dependence on G.

We are interested in identifying Pareto improvements in the joint payoff vector of the seller and users that can be achieved by modifying the network structure of G. Importantly, we assume that the seller always adjusts the offered contract optimally in response to any new network structure G'. A network G is said to be *Pareto-efficient* if no such welfare-improving modification is possible. Formally:

Definition 2. A network G is *Pareto-efficient* if, for any seller-optimal contract C under G, there does not exist another network G' (with modified connections) and a seller-optimal contract C' under G' such that the following conditions hold:

(i) $\Pi(C', G') \geq \Pi(C, G)$,

(ii) $\pi_i(C', G') \geq \pi_i(C, G)$ for all $i \in N$, and strict inequality holds for at least one $i \in N$.

Remark 8. It is worth emphasizing that we do not consider Pareto improvements that increase the seller's profit without affecting users' payoffs. In fact, such a scenario is not possible. Consider a network G, an optimal contract C under G, and let $M(C)$ denote the maximum independent set targeted by the seller by C. As shown in Proposition 2, when the prior is sufficiently noisy (i.e., z_0 is small), the seller's profit $\Pi(C, G)$ depends solely on the independence number $m = |M(C)|$ and not on the specific network structure.

Moreover, under the optimal contract design, no user in $M(C)$ is free-riding on others; each is individually targeted and pays for access to data. Therefore, the only way to increase the seller's profit is to enlarge $M(C)$, i.e., to convert some free-riders into paying users. However, this implies that at least one user who previously received a utility of at least $-\frac{1}{z_0+z}$ as a free-rider now receives $-\frac{1}{z_0}$ as a purchaser. Since this reduces that user's payoff, the outcome is not a Pareto improvement.

The direction of Pareto improvement now shifts toward enhancing the expected payoffs of users – particularly those who free-ride under the optimal contract. A natural and straightforward strategy is to add links from a free-rider to all purchasers, thereby increasing that free-rider's expected payoff without diminishing the utility or profit of any other participant.

Through this process, we arrive at a Pareto-improved network with an associated optimal contract characterized as follows: purchasers (i.e., periphery nodes) are mutually disconnected,

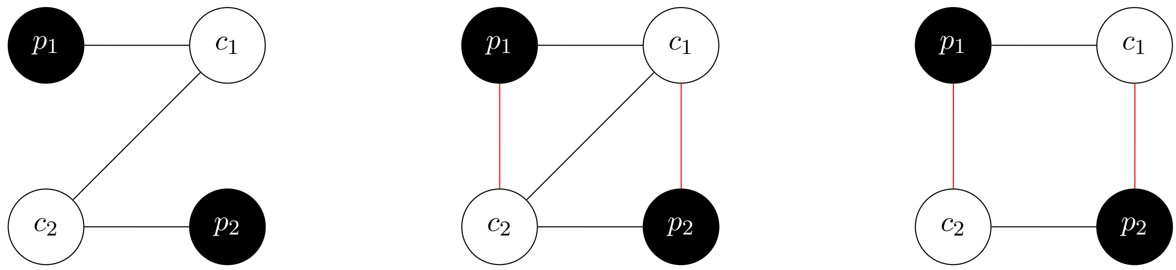


Figure 7: Left: G_1, not Pareto-efficient; middle: G_2, Pareto-efficient; right: G_3, Pareto-efficient. $G_1 \to G_2$: Pareto improving. In all networks, $\{p_1, p_2\}$ is the target set.

while free-riders (quasi-core nodes) are connected to all purchasers. Figure 7 illustrates this Pareto improvement process using a sequence of example networks.

Example 5 (Pareto Improvement)**.** Consider the initial network G_1, where the maximum independent set under the optimal contract is $M(C) = \{p_1, p_2\}$. Nodes c_1 and c_2 are free-riders who are not connected to all purchasers. To improve their payoffs without affecting others, we add the links $\{c_1, p_2\}$ and $\{c_2, p_1\}$, resulting in the modified network G_2. The new network G_2 is Pareto-efficient: $\{p_1, p_2\}$ are the purchasers (periphery nodes), and $\{c_1, c_2\}$ are the free-riders (quasi-core nodes), each now connected to both purchasers.

By contrast, network G_3, obtained by removing the link $\{c_1, c_2\}$ (a connection between two free-riders), does not arise from this Pareto-improvement procedure. Since neither c_1 nor c_2 purchases any signal, severing the link between them does not alter any user's expected payoff and thus does not constitute a Pareto improvement.

We now introduce the formal definition of networks with a *quasi-core-periphery* structure. This concept plays a central role in our analysis and allows us to fully characterize the set of Pareto-efficient networks.

Definition 3. A network G is said to have a *quasi-core-periphery structure* if the set of nodes N can be partitioned into two disjoint subsets: N_c (quasi-core nodes) and N_p (periphery nodes), such that the following conditions are satisfied:

(i) $N_c = N \setminus N_p$, and the independence number of the subgraph induced by N_c is at most $|N_p|$;

(ii) Every quasi-core node is connected to every periphery node: $g_{ij} = 1$ for all $(i, j) \in N_c \times N_p$;

(iii) No two periphery nodes are connected: $g_{ij} = 0$ for all $i, j \in N_p$.

Remark 9. Our definition of a quasi-core-periphery structure differs from the classical core-periphery structure in the literature, where core nodes are typically assumed to be fully interconnected. In contrast, we do not require full connectivity among quasi-core nodes. Instead,

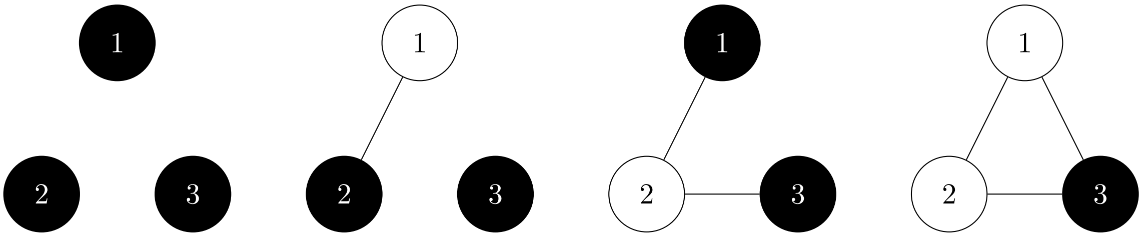


Figure 8: Left: G_1, $N_p = M(C) = \{1, 2, 3\}$, quasi-core-periphery; middle-left: G_2, $M(C) = \{2, 3\}$, not quasi-core-periphery; middle-right: G_3, $N_p = M(C) = \{1, 3\}$, quasi-core-periphery; right: G_4, $N_p = M(C) = \{3\}$, quasi-core-periphery.

we impose that the independence number of the subgraph induced by the quasi-core nodes does not exceed the number of periphery nodes.

This distinction is illustrated by network G_3 in Example 5. First, the connection between c_1 and c_2 is not necessary for G_3 to be Pareto-efficient. Second, since c_1 and c_2 are disconnected, the independence number of the subgraph induced by $N_c = \{c_1, c_2\}$ is 2, which is no larger than the number of periphery nodes $N_p = \{p_1, p_2\}$. However, if we were to include an additional disconnected node c_3 in N_c, forming $N_c = \{c_1, c_2, c_3\}$, the independence number of the induced subgraph would exceed the size of $N_p = \{p_1, p_2\}$. In that case, N_p could no longer serve as a maximum independent set in an optimal contract.

Example 6. Figure 8 lists all configurations of 3-user social networks. Among these, we observe that only G_2 (middle-left) does not conform to a quasi-core-periphery structure. Specifically, when N_p corresponds to the maximum independent set $M(C) = \{2, 3\}$[8], user 1 in $\{N_c\}$ is not connected to user 3. As noted in Proposition 6, G_2 is not Pareto-efficient. Adding the link between users 1 and 3 to G_2 results in a Pareto improvement for both the seller and all users.

In contrast, in G_4 (right), $N_p = M(C) = \{3\}$, and $N_c = \{1, 2\}$. Since users 1 and 2 are connected, the independence number of N_c is 1, which does not exceed $|N_p|$. Therefore, G_4 satisfies the quasi-core-periphery structure and is Pareto-efficient.

Following the analysis in this subsection, the networks defined above with a quasi-core-periphery structure fully characterize the set of Pareto-efficient networks:

Proposition 6. A network is Pareto-efficient *if and only if* it is a network of a quasi-core-periphery structure.

Note that a Pareto-efficient network maximizes the number of connections between free-riders and purchasers of data. The network where free-riders are connected to all purchasers

[8] Alternatively, the maximum independent set could be $M(C) = \{1, 3\}$, in which case a symmetric argument applies.

provides the highest consumer surplus, while the connections between free-riders do not affect the seller's profit or consumer surplus.[9] Thus, it is incentive-compatible for the seller to add connections between free-riders, while keeping the set of buyers as the maximum independent set.

5.2 Network Formation

Throughout most of the paper, we examine the problem of maximizing revenue or social welfare, assuming the social network is given. However, we can relax this assumption and instead consider a network formation process among users that precedes the seller's pricing strategy. To this end, let us first introduce the following notion, which focuses solely on users' efficiency:

Definition 4. A network G is *users-Pareto-efficient* if, for any seller-optimal contract C of network G, there exists no other network G' with modified connections and a seller-optimal contract C' for G' such that the payoffs of the users satisfy: $\pi_i(C', G') \geq \pi_i(C, G)$ for all $i \in N$, and with strict inequality for some $i \in N$.

Unlike the Pareto-efficient network, the users-Pareto-efficient network disregards the seller's profit and instead captures the idea that users collectively choose the social network based solely on their own efficiency and payoffs. If a network is Pareto-efficient, it is necessarily users-Pareto-efficient. Moreover, as discussed in Remark 8, the converse also holds because, given that a contract is maximizing the seller's profit, it is impossible to increase the seller's profit without making at least one user worse off.

Given this, consider now a multi-stage game in which the users can coordinate on a social network to maximize their joint payoffs before the seller offers the pricing contract. Let $\mathcal{E}$ denote the set of subgame perfect Nash equilibria such that, in a pricing subgame with multiple optimal pricing strategies (if they exist), the seller chooses the one that maximizes consumer surplus. Then, any equilibrium network G in $\mathcal{E}$ must be users-Pareto-efficient, since otherwise, the group of users would improve their joint surplus by deviating to another network. Combining this with the result in Proposition 6, we obtain the following characterization of the optimal network structure from the users' point of view:

Proposition 7. In the multi-stage game in which the users can coordinate on a social network, an equilibrium network in $\mathcal{E}$ is of a quasi-core-periphery structure.

[9] A bipartite network is a quasi-core-periphery network but not a core-periphery network. To maximize the spillover effect of information, each pair of a free-rider and a purchaser should be connected. Moreover, the purchaser group, as the target set (periphery nodes), must have a larger cardinality.

5.3 Value of Network Information

In this final section, we examine the value of information regarding the network structure for the seller. As discussed in Remark 4 following Theorem 1, the sufficient statistic of the network structure to design the monopoly-optimal contract is the independence number m. Here when referring to the information about network structure, we mean the independence number m.

To do so, we first calculate the seller's optimal profit when she/he has correct information of the network structure, which we denote as Π_p. Next, we explore the consequences of not having this information and examine how much loss the seller incurs under misbeliefs about the network. We consider two scenarios: (i) the seller sets prices optimally with the misbelief that the network is empty (i.e., no connections among nodes); (ii) the seller sets prices optimally with the misbelief that the network is complete (i.e., all nodes are fully connected). The profits in these two cases are denoted by Π_e and Π_c, respectively.[10]

When the seller lacks information about the network structure, she/he offers a uniform price to all users. With either belief (correct information or misbelief), it is not in equilibrium for two neighboring users to both purchase the signal. Thus, the price is always given by $p_i = \frac{1}{z_0} - \frac{1}{z_0+z}$, where z is the precision of data collected by the seller (to be determined based on the seller's information or belief), and users are willing to purchase the data when none of their neighbors buy it. Since there is only a linear cost γz for the information product with precision z, and no additional replication costs, the optimal precision z is determined by how many users will eventually purchase the data in equilibrium. This is set to equate the marginal revenue $\frac{m}{(z_0+z)^2}$ with the marginal cost γ.[11]

We thus obtain the following lemma:

Lemma 3. When $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$, the optimal information precision $z = \sqrt{\frac{n}{\gamma}} - z_0$. Moreover,

(i) To calculate Π_p, m is the maximum independence number.

(ii) To calculate Π_e, $m = n$ is the misbelieved maximum independence number.

(iii) To calculate Π_c, $m = 1$ is the misbelieved maximum independence number.

Illustrated in Figure 9 is an example of a 4-user cycle. The following lemma characterizes the seller's profit in each case.

Lemma 4. Assume $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Then, we have

[10] In Candogan et al. (2012), the value of network information, by comparing the profit of a monopolistic seller with correct information of the network structure to that assuming an empty network, is also studied. However, they consider users' strategic complementarities on the network.

[11] Note that multiple equilibria may arise at the users' data-purchasing stage. Among these, we focus on the equilibrium where the users who purchase the data form a maximum independent set (i.e., the maximum number of users who choose to buy when they are indifferent).

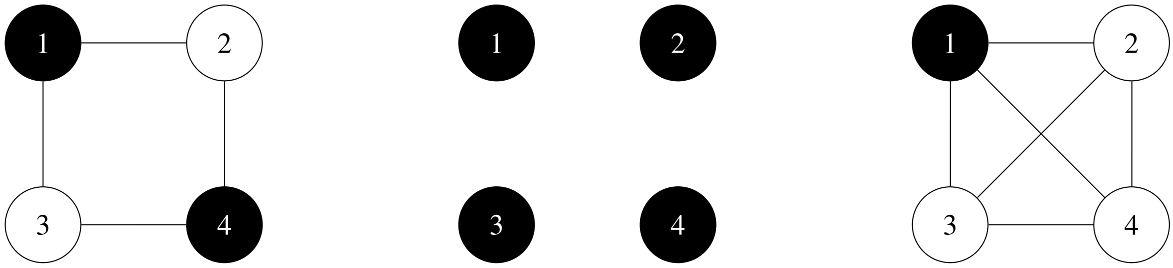


Figure 9: Left: G_1, perfect knowledge, $M(C_1) = \{1, 4\}$; middle: G_2, misbelief of empty network, $M(C_2) = \{1, 2, 3, 4\}$; right: G_3, misbelief of complete network, $M(C_3) = \{1\}$.

(i) $\Pi_e = \frac{m}{z_0} + \gamma z_0 - \left(\frac{m}{\sqrt{n}} + \sqrt{n}\right)\sqrt{\gamma}$,

(ii) $\Pi_c = \frac{m}{z_0} + \gamma z_0 - (m+1)\sqrt{\gamma}$,

(iii) $\Pi_p = \frac{m}{z_0} + \gamma z_0 - 2\sqrt{m\gamma}$.

We measure the value of information about the network structure by examining the differences $\Pi_p - \Pi_e$ and $\Pi_p - \Pi_c$ in the two cases. These differences in the seller's profit are caused by incorrect beliefs and the fact that, under correct information, the divergence in the seller's profits is measured by incomplete information. A higher difference indicates a higher value of the seller's network information. When the network structure and game parameters are fixed, these values are uniquely determined by the profits given in Lemma 4, which are expressed as follows:

$$\mathcal{V}_c(m) \equiv \Pi_p - \Pi_c = \sqrt{\gamma}\left(\sqrt{m} - 1\right)^2,$$

and

$$\mathcal{V}_e(m) \equiv \Pi_p - \Pi_e = \frac{\sqrt{\gamma}\left(\sqrt{n} - \sqrt{m}\right)^2}{\sqrt{n}}.$$

Note that these values are always nonnegative, meaning that information is valuable to the seller in both cases. Furthermore, $\mathcal{V}_c$ is increasing in m, while $\mathcal{V}_e$ is decreasing in m. This reflects the fact that the value of information about network structure becomes higher when the actual network and the one under the misbelief diverge more.

Proposition 8. Assume $z_0 \leq \frac{1}{4\sqrt{\gamma}}$. Then, for all $m \leq n$:

(i) $0 \leq \mathcal{V}_e(m) \leq \mathcal{V}_e(1) = \frac{\sqrt{\gamma}\left(\sqrt{n}-1\right)^2}{\sqrt{n}}$.

(ii) $\frac{\mathcal{V}_e(m)}{\Pi_p} \leq R_e(n) \equiv \frac{z_0\sqrt{\gamma}(\sqrt{n}-1)^2}{\sqrt{n}(1-z_0\sqrt{\gamma})^2}$, where $R_e(n) \to \infty$ as $n \to \infty$.

(iii) $0 \leq \mathcal{V}_c(m) \leq \mathcal{V}_c(n) = \sqrt{\gamma}\left(\sqrt{n} - 1\right)^2$.

(iv) $\frac{\mathcal{V}_c(m)}{\Pi_p} \leq R_c(n) \equiv \frac{z_0\sqrt{\gamma}(\sqrt{n}-1)^2}{(\sqrt{n}-z_0\sqrt{\gamma})^2}$, where $R_c(n) \leq \frac{1}{4}$ for all n.

From the proposition, we observe that both upper bounds on the value of information in (i) and (iii) increase in n and γ, meaning that the maximum value of network information can be higher in environments where collecting data is more costly, or there are more potential buyers in the market. From (iv), we also see that the ratio $\frac{\mathcal{V}_c(m)}{\Pi_p}$ is bounded above by $\frac{1}{4}$, suggesting that the percentage of profit loss caused by the seller's incorrect belief that the network is complete does not exceed $\frac{1}{4}$. On the other hand, such a bound is not present for $\frac{\mathcal{V}_e(m)}{\Pi_p}$ as shown in (ii), since R_e diverges to infinity as n increases. This is intuitive because collecting data with high precision in an empty network is not profitable when the actual network is highly connected and the independence number in the true network is small.

6 Conclusion

In this paper, we have explored a stylized model of a monopoly data seller within an information-sharing network among users, where the seller cannot discriminate against users based on data quality. We demonstrated that when the users' prior information is sufficiently noisy, the optimal selling strategy is characterized by a maximum independent set, which includes the largest group of users with no information-sharing links between them. Additionally, we showed that although the precision of the seller's data decreases as information-sharing links among users increase, it remains higher than the socially optimal level of precision.

While our analysis provides insights into data pricing within social networks and its implications for social welfare, several limitations remain. First, we abstracted from potential coordination motives by assuming that users' payoffs depend solely on the true state of the world, not on the actions of others. Although we believe that our results hold under the assumption of weak coordination motives, it is important to examine the robustness of our findings in such scenarios. Second, we have not fully addressed the case of incomplete information about the social network, particularly the seller's incentives to gather information or to screen for the network structure. Finally, our model does not consider the effects of competition among data sellers. These important topics are left for future research.

References

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

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

Admati, Anat R, and Paul Pfleiderer. 1990. "Direct and indirect sale of information." *Econometrica*, 58(4): 901–928.

Allouch, Nizar. 2015. "On the private provision of public goods on networks." *Journal of Economic Theory*, 157: 527–552.

Alon, Noga, and Joel H Spencer. 2016. *The Probabilistic Method.* John Wiley & Sons.

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

Bergemann, Dirk, and Alessandro Bonatti. 2019. "Markets for information: An introduction." *Annual Review of Economics*, 11: 85–107.

Bloch, Francis. 2016. "Targeting and pricing in social networks." *The Oxford Handbook of the Economics of Networks*.

Bloch, Francis, and Nicolas Quérou. 2013. "Pricing in social networks." *Games and Economic Behavior*, 80: 243–261.

Bloch, Francis, Gabrielle Demange, and Rachel Kranton. 2018. "Rumors and social networks." *International Economic Review*, 59(2): 421–448.

Bramoullé, Yann, and Rachel Kranton. 2007. "Public goods in networks." *Journal of Economic Theory*, 135(1): 478–494.

Bramoullé, Yann, Rachel Kranton, and Martin D'amours. 2014. "Strategic interaction and networks." *American Economic Review*, 104(3): 898–930.

Candogan, Ozan. 2022. "Persuasion in networks: public signals and cores." *Operations Research*, 70(4): 2264–2298.

Candogan, Ozan, Kostas Bimpikis, and Asuman Ozdaglar. 2012. "Optimal pricing in networks with externalities." *Operations Research*, 60(4): 883–905.

Caro, Yair. 1979. "New results on the independence number." Technical Report, Tel-Aviv University.

Chen, Ying-Ju, Yves Zenou, and Junjie Zhou. 2018. "Competitive pricing strategies in social networks." *The RAND Journal of Economics*, 49(3): 672–705.

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

Egorov, Georgy, and Konstantin Sonin. 2020. "Persuasion on networks." National Bureau of Economic Research.

Fainmesser, Itay P, and Andrea Galeotti. 2016. "Pricing network effects." *The Review of Economic Studies*, 83(1): 165–198.

Galeotti, Andrea, and Sanjeev Goyal. 2010. "The law of the few." *American Economic Review*, 100(4): 1468–92.

Gal-Or, Esther. 1985. "Information sharing in oligopoly." *Econometrica: Journal of the Econometric Society*, 329–343.

Goyal, Sanjeev, Stephanie Rosenkranz, Utz Weitzel, and Vincent Buskens. 2017. "Information acquisition and exchange in social networks." *The Economic Journal*, 127(606): 2302–2331.

Halim, Edward, Yohanes E Riyanto, and Nilanjan Roy. 2019. "Costly information acquisition, social networks, and asset prices: Experimental evidence." *The Journal of Finance*, 74(4): 1975–2010.

Han, Bing, and Liyan Yang. 2013. "Social networks, information acquisition, and asset prices." *Management Science*, 59(6): 1444–1457.

Herskovic, Bernard, and Joao Ramos. 2020. "Acquiring information through peers." *American Economic Review*, 110(7): 2128–52.

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

Jones, Charles I, and Christopher Tonetti. 2020. "Nonrivalry and the economics of data." *American Economic Review*, 110(9): 2819–2858.

Leister, C Matthew. 2020. "Information acquisition and welfare in network games." *Games and Economic Behavior*, 122: 453–475.

Li, Lode. 2002. "Information sharing in a supply chain with horizontal competition." *Management Science*, 48(9): 1196–1212.

Liu, Zekun, Dennis J Zhang, and Fuqiang Zhang. 2021. "Information sharing on retail platforms." *Manufacturing & Service Operations Management*, 23(3): 606–619.

Myatt, David P, and Chris Wallace. 2012. "Endogenous information acquisition in coordination games." *The Review of Economic Studies*, 79(1): 340–374.

Myatt, David P, and Chris Wallace. 2018. "Information use and acquisition in price-setting oligopolies." *The Economic Journal*, 128(609): 845–886.

Myatt, David P, and Chris Wallace. 2019. "Information acquisition and use by networked players." *Journal of Economic Theory*, 182: 360–401.

Ozsoylev, Han N, and Johan Walden. 2011. "Asset pricing in large information networks." *Journal of Economic Theory*, 146(6): 2252–2280.

Wei, VK. 1981. "A lower bound on the stability number of a simple graph." Bell Laboratories Tech Memorandum, No. 81-11217-9.

Appendix: Proofs

Proof of Proposition 1. Let $i \in M(C)$ be a user with the maximal value of m_i. The intuition is that if user i is served, many other users in $M(C)$ can free-ride on i's information. The seller may prefer to avoid serving such a user in order to extract higher payments from users in M_i (i.e., those who are neighbors of i in the network). To show this, we consider removing i from $M(C)$ and compare the resulting cost and benefit. We will demonstrate that the benefit strictly outweighs the cost, thereby improving the seller's revenue. This implies that $M(C)$ cannot be optimal if it is not an independent set.

Cost. The cost is the lost revenue from user i, given by:

$$p_i = \frac{1}{z_0 + m_i z} - \frac{1}{z_0 + (m_i + 1)z}.$$

Benefit. The benefit arises from the increase in prices for users in M_i. Let p_j and p'_j denote the price for user $j \in M_i$ before and after the removal of i, respectively. Then the total gain in revenue is:

$$\begin{aligned}\sum_{j \in M_i} (p'_j - p_j) &= \sum_{j \in M_i} \left[\left(\frac{1}{z_0 + (m_j - 1)z} - \frac{1}{z_0 + m_j z} \right) - \left(\frac{1}{z_0 + m_j z} - \frac{1}{z_0 + (m_j + 1)z} \right) \right] \\ &= \sum_{j \in M_i} \left(\frac{1}{z_0 + (m_j - 1)z} + \frac{1}{z_0 + (m_j + 1)z} - \frac{2}{z_0 + m_j z} \right).\end{aligned}$$

Define the difference term:

$$\Delta(m_j) := \frac{1}{z_0 + (m_j - 1)z} + \frac{1}{z_0 + (m_j + 1)z} - \frac{2}{z_0 + m_j z}.$$

Then the following properties hold:

(i) $\Delta(m_j) > 0$ since the function $f(x) = \frac{1}{x}$ is convex on $(0, \infty)$.

(ii) $\Delta'(m_j) < 0$ since the derivative is given by

$$\Delta'(m_j) = \frac{2z}{(z_0 + m_j z)^2} - \left(\frac{z}{(z_0 + (m_j - 1)z)^2} + \frac{z}{(z_0 + (m_j + 1)z)^2} \right),$$

which is negative due to the convexity of $f(x) = \frac{1}{x^2}$.

This implies that $\Delta(m_j)$ is strictly decreasing in m_j. Since $m_j \in [1, m_i]$ by the maximality of m_i, we obtain for all $j \in M_i$,

$$p'_j - p_j = \Delta(m_j) \geq \Delta(m_i) = \frac{1}{z_0 + (m_i - 1)z} + \frac{1}{z_0 + (m_i + 1)z} - \frac{2}{z_0 + m_i z}.$$

Thus, the total benefit satisfies:

$$\sum_{j \in M_i} (p'_j - p_j) \geq m_i \left(\frac{1}{z_0 + (m_i - 1)z} + \frac{1}{z_0 + (m_i + 1)z} - \frac{2}{z_0 + m_i z} \right).$$

Comparison. It remains to show that this benefit exceeds the cost:

$$m_i \left(\frac{1}{z_0 + (m_i - 1)z} + \frac{1}{z_0 + (m_i + 1)z} - \frac{2}{z_0 + m_i z} \right) > \frac{1}{z_0 + m_i z} - \frac{1}{z_0 + (m_i + 1)z}.$$

This is equivalent to:

$$\frac{m_i}{z_0 + (m_i - 1)z} + \frac{m_i + 1}{z_0 + (m_i + 1)z} > \frac{2m_i + 1}{z_0 + m_i z}.$$

A direct computation confirms that this inequality holds if and only if

$$\frac{z((m_i + 1)z - z_0)}{(2m_i + 1)(z_0 + (m_i - 1)z)(z_0 + m_i z)(z_0 + (m_i + 1)z)} > 0,$$

which is equivalent to:

$$z > \frac{z_0}{m_i + 1}. \quad \diamond$$

Proof of Theorem 1. If the optimal set $M(C)$ is a maximum independent set, the optimal contract values (z, p_i) are straightforward to characterize.[12]

In the following, we show that any optimal contract $C = (z, p_i)_{i \in N}$ must target an independent set $M(C)$ of the network G, and then argue that a maximum independent set maximizes the seller's profit.[13]

[12]Since the profit is given by $\Pi(C) = m\left(\frac{1}{z_0} - \frac{1}{z_0 + z}\right) - \gamma z$, the first-order condition yields the optimal $z = \sqrt{\frac{m}{\gamma}} - z_0$ and corresponding price $p_i = \frac{1}{z_0} - \sqrt{\frac{\gamma}{m}}$ for all $i \in M(C)$. The second-order condition is satisfied since the second derivative with respect to z is $-\frac{2m}{(z_0+z)^3} < 0$.

[13]Let $m' \leq m$ be the size of a targeted independent set. Then the optimal profit is

$$\Pi(C) = m' \left(\frac{1}{z_0} - \sqrt{\frac{\gamma}{m'}} \right) - \gamma \left(\sqrt{\frac{m'}{\gamma}} - z_0 \right) = \frac{m'}{z_0} - 2\sqrt{m'\gamma} + \gamma z_0,$$

Suppose instead that a current $M(C)$ is not an independent set, and the contract $C = (z, p_i)_{i \in N}$ is optimal under this $M(C)$. Then there must exist at least two connected nodes $i, j \in M(C)$. Let $i^* \in M(C)$ be such that m_{i^*} is maximal. By Proposition 1, a sufficient condition for C not to be optimal is:

$$z < \frac{z_0}{1 + m_{i^*}}. \tag{4}$$

We will use this condition to derive a contradiction from the assumption that C is optimal. Specifically, we show that increasing z leads to strictly higher profit. Let us consider the marginal revenue net of cost from increasing z on seller's profit:

$$\Pi(C) = \sum_{i \in M(C)} \left(\frac{1}{z_0 + m_i z} - \frac{1}{z_0 + (m_i + 1)z} \right) - \gamma z.$$

Hence, it is sufficient to show that the marginal revenue per user,

$$MR_i(C) := \frac{m_i + 1}{(z_0 + (m_i + 1)z)^2} - \frac{m_i}{(z_0 + m_i z)^2},$$

satisfies

$$MR_i(C) > \frac{\gamma}{m} \text{ for all } i \in M(C).$$

Define

$$\beta(z) := \frac{z_0^2 - m_i(m_i + 1)z^2}{(z_0 + (m_i + 1)z)^2 (z_0 + m_i z)^2},$$

so that $MR_i(C) = \beta(z)$. We aim to show that $\beta(z) > \frac{\gamma}{m}$ for all $z < \frac{z_0}{1+m_i}$ (note that this bound holds for all $i \in M(C)$ since $m_i \leq m_{i^*}$).

Observe that $\beta(z)$ is strictly decreasing in z over $[0, \infty)$ because the numerator decreases and the denominator increases in z. Therefore, to verify $\beta(z) > \frac{\gamma}{m}$ for all such z, it suffices to evaluate at the boundary:

$$z = \frac{z_0}{1 + m_i}.$$

Substituting, we obtain:

$$\begin{aligned}
\beta\left(\frac{z_0}{1 + m_i}\right) &= \frac{z_0^2 \left(1 - \frac{m_i}{m_i+1}\right)}{(z_0 + z_0)^2 \left(z_0 + \frac{m_i z_0}{m_i+1}\right)^2} \\
&= \frac{z_0^2 \left(\frac{1}{m_i+1}\right)}{(2z_0)^2 \left(z_0 \left(1 + \frac{m_i}{m_i+1}\right)\right)^2} \\
&= \frac{1}{4(m_i + 1) z_0^2 \left(1 + \frac{m_i}{m_i+1}\right)^2}.
\end{aligned}$$

which is increasing in $\sqrt{m'}$ because the minimum of this quadratic expression occurs at $\sqrt{m'} = z_0\sqrt{\gamma} < 1$, while $\sqrt{m'} \geq 1$ by assumption.

To ensure $\beta\left(\frac{z_0}{1+m_i}\right) > \frac{\gamma}{m}$, we use the inequalities:

$$m_i + 1 \leq m, \quad \frac{m_i}{m_i+1} \leq \frac{n}{n+1}, \quad \text{and} \quad z_0 < \frac{1}{2\sqrt{\gamma}} \cdot \frac{n+1}{2n+1}.$$

Combining these gives:

$$\beta\left(\frac{z_0}{1+m_i}\right) > \frac{\gamma}{m}.$$

This contradicts the optimality of C under a non-independent $M(C)$. Hence, any optimal contract must target an independent set. Finally, among all such sets, a maximum independent set maximizes profit, completing the proof. $\diamond$

Proof of Proposition 2. From Theorem 1, the seller's profit under the optimal contract is given by:

$$\Pi(z_0, G) = \frac{m}{z_0} + \gamma z_0 - 2\sqrt{\gamma m},$$

where m denotes the independence number of the network G.

First, we analyze the marginal effect of the independence number m on the seller's profit. The partial derivative is:

$$\frac{\partial \Pi(z_0, G)}{\partial m} = \frac{1}{z_0} - \sqrt{\frac{\gamma}{m}}.$$

Under the assumption of a noisy common prior, we have

$$z_0 < \frac{1}{2\sqrt{\gamma}} \cdot \frac{n+1}{2n+1}.$$

This implies:

$$\frac{1}{z_0} > \frac{2(2n+1)}{n+1}\sqrt{\gamma},$$

and thus,

$$\frac{\partial \Pi(z_0, G)}{\partial m} > \left(\frac{2(2n+1)}{n+1} - \sqrt{\frac{1}{m}}\right)\sqrt{\gamma}.$$

Since $m \geq 1$ and $\frac{2(2n+1)}{n+1} \geq 2$, the right-hand side is strictly positive, and hence:

$$\frac{\partial \Pi(z_0, G)}{\partial m} > 0.$$

Next, we examine how profit varies with the prior precision z_0:

$$\frac{\partial \Pi(z_0, G)}{\partial z_0} = -\frac{m}{z_0^2} + \gamma.$$

Using the bound on z_0 again, we find:

$$z_0^2 < \frac{1}{4\gamma}\left(\frac{n+1}{2n+1}\right)^2,$$

so:
$$\frac{1}{z_0^2} > 4\gamma \left(\frac{2n+1}{n+1}\right)^2.$$
Therefore:
$$\frac{\partial \Pi(z_0, G)}{\partial z_0} < -4m\gamma \left(\frac{2n+1}{n+1}\right)^2 + \gamma.$$
Since $m \geq 1$ and $\left(\frac{2n+1}{n+1}\right)^2 > 1$, we conclude that:
$$\frac{\partial \Pi(z_0, G)}{\partial z_0} < -3\gamma < 0.$$

In summary, the seller's profit is strictly increasing in the independence number m of the network and strictly decreasing in the precision z_0 of the common prior. $\diamond$

Proof of Proposition 4. From equation (2), the consumer surplus is given by
$$w\left(z_0, G, M(C)\right) = -\sum_{i \in M(C)} \left(\frac{1}{z_0}\right) - \sum_{i \in N \setminus M(C)} \left(\frac{1}{z_0 + m_i z^o}\right),$$
where the seller-optimal precision is denoted by $z^o = \sqrt{\frac{m}{\gamma}} - z_0$.

For any two optimal contracts C and C', we have $|M(C)| = |M(C')| = m$, so the first terms in the consumer surplus expressions are identical:
$$-\sum_{i \in M(C)} \left(\frac{1}{z_0}\right) = -\sum_{i \in M(C')} \left(\frac{1}{z_0}\right) = -\frac{m}{z_0}.$$
The second components differ as follows:
$$-\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i z^o} \quad \text{vs.} \quad -\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i' z^o}.$$
Define the utility function of a free-riding consumer by
$$\tilde{u}(m_i) = -\frac{1}{z_0 + m_i z^o},$$
which is increasing and concave in m_i since
$$\frac{d\tilde{u}(m_i)}{dm_i} = \frac{z^o}{(z_0 + m_i z^o)^2} > 0, \quad \text{and} \quad \frac{d^2\tilde{u}(m_i)}{dm_i^2} = -\frac{2(z^o)^2}{(z_0 + m_i z^o)^3} < 0.$$
Part (i). Suppose $F_{k(C)}(\cdot)$ first-order stochastically dominates $F_{k(C')}(\cdot)$, which is equivalent to $k(C) \geq k(C')$. Then, for all $i = 1, \ldots, n-m$,
$$m_i \geq m_i' \quad \Rightarrow \quad \tilde{u}(m_i) \geq \tilde{u}(m_i').$$

This implies

$$-\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i z^o} \geq -\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i' z^o}.$$

Therefore, the consumer surplus under contract C is weakly greater than that under contract C', i.e.,

$$w(z_0, G, M(C)) \geq w(z_0, G, M(C')).$$

Part (ii). Suppose $F_{k(C)}(\cdot)$ second-order stochastically dominates $F_{k(C')}(\cdot)$. Then, for any nondecreasing concave function $U : \mathbb{R} \to \mathbb{R}$, we have

$$\int U(x)\, dF_{k(C)}(x) \geq \int U(x)\, dF_{k(C')}(x).$$

In particular, $\tilde{u}(m_i) = \tilde{u}(m_i) = -\frac{1}{z_0 + m_i z^o}$ is increasing and concave, and $F_{k(C)}$, $F_{k(C')}$ represent the empirical distributions of $\{m_i\}_{i \notin M(C)}$ and $\{m_i'\}_{i \notin M(C')}$, respectively, so we obtain:

$$\int \tilde{u}(x)\, dF_{k(C)}(x) = -\frac{1}{n-m} \sum_{i=1}^{n-m} \frac{1}{z_0 + m_i z^o},$$

and

$$\int \tilde{u}(x)\, dF_{k(C')}(x) = -\frac{1}{n-m} \sum_{i=1}^{n-m} \frac{1}{z_0 + m_i' z^o}.$$

Multiplying both sides by $n - m$ preserves the inequality:

$$-\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i z^o} \geq -\sum_{i=1}^{n-m} \frac{1}{z_0 + m_i' z^o}.$$

Hence,

$$w(z_0, G, M(C)) \geq w(z_0, G, M(C')). \quad \diamond$$

Proof of Theorem 2 Suppose that $z_0 < \frac{1}{2\sqrt{\gamma}} \frac{n+1}{2n+1}$. Under this condition, the monopoly-optimal data precision is given by

$$z^o = \sqrt{\frac{m}{\gamma}} - z_0,$$

where m denotes the size of a maximum independent set.

In contrast, the socially efficient level of data precision z^* maximizes the social welfare function:

$$SW(z) = \sum_{i \in N} -\left[\frac{1}{(n_i + 1)z + z_0}\right] - \gamma z.$$

The first-order condition for optimality yields:

$$\sum_{i \in N} \frac{n_i + 1}{[(n_i + 1)z^* + z_0]^2} = \gamma.$$

To prove the result, we invoke the Caro-Wei bound from graph theory (see, e.g., Caro (1979), Wei (1981), and for a proof of the probabilistic method Alon and Spencer (2016)), which states:

$$\sum_{i \in N} \frac{1}{n_i + 1} \leq m,$$

with equality if and only if the network G is a disjoint union of cliques.

Part (i). Suppose G is a disjoint union of cliques. Then by the Caro-Wei bound, $\sum_{i \in N} \frac{1}{n_i+1} = m$. Define the function

$$h(z) := \sum_{i \in N} \frac{n_i + 1}{[(n_i + 1)z + z_0]^2}.$$

Observe that $h(z)$ is strictly decreasing in z, and satisfies $h(z^*) = \gamma$. It suffices to show that $h(z^o) \geq \gamma$, with equality only when G is the empty graph.

Substituting z^o into $h(z)$, we get:

$$h(z^o) = \sum_{i \in N} \frac{n_i + 1}{\left[(n_i + 1)\left(\sqrt{\frac{m}{\gamma}} - z_0\right) + z_0\right]^2} = \sum_{i \in N} \frac{n_i + 1}{\left((n_i + 1)\sqrt{\frac{m}{\gamma}} - n_i z_0\right)^2}.$$

Using the inequality $(n_i + 1)\sqrt{\frac{m}{\gamma}} - n_i z_0 \leq (n_i + 1)\sqrt{\frac{m}{\gamma}}$, we obtain:

$$h(z^o) \geq \sum_{i \in N} \frac{n_i + 1}{\left((n_i + 1)\sqrt{\frac{m}{\gamma}}\right)^2} = \sum_{i \in N} \frac{\gamma}{(n_i + 1)m} = \gamma.$$

Equality holds if and only if $n_i = 0$ for all $i \in N$, i.e., when the network is empty. Therefore, $z^* \geq z^o$, with equality only when G is empty.

Part (ii). Suppose G is not a disjoint union of cliques. Then the Caro-Wei bound implies:

$$\sum_{i \in N} \frac{1}{n_i + 1} < m.$$

Taking the limit as $z_0 \to 0$, we find:

$$z^o \to z^o_{\text{lim}} := \sqrt{\frac{m}{\gamma}}, \quad z^* \to z^*_{\text{lim}} := \sqrt{\frac{1}{\gamma} \sum_{i \in N} \frac{1}{n_i + 1}}.$$

Since $\sum_{i \in N} \frac{1}{n_i+1} < m$, we conclude that $z^*_{\text{lim}} < z^o_{\text{lim}}$, so for sufficiently small z_0, it follows that $z^* < z^o$. $\diamond$

Proof of Lemma 4. We analyze the seller's profit under three scenarios: assuming an empty network, assuming a complete network, and having perfect information about the actual network structure.

(i) Seller assumes the network is empty. In this case, the seller believes there are no information-sharing links among users, so the independence number is n. The price charged to each user is:

$$p_i = \frac{1}{z_0} - \frac{1}{z_0 + z_e},$$

and the seller's profit is:

$$\pi = np_i - \gamma z_e.$$

Maximizing profit, the seller chooses

$$z_e = \sqrt{\frac{n}{\gamma}} - z_0.$$

However, the true independence number of the actual network G is $m \leq n$, so the realized profit (based on actual purchases from m independent users) is:

$$\Pi_e = mp_i - \gamma z_e = \frac{m}{z_0} + \gamma z_0 - \left(\frac{m}{\sqrt{n}} + \sqrt{n}\right)\sqrt{\gamma}.$$

(ii) Seller assumes the network is complete. Here, the seller believes all users share information perfectly, so the independence number is 1. The price and profit are:

$$p_i = \frac{1}{z_0} - \frac{1}{z_0 + z_c}, \quad \pi = p_i - \gamma z_c.$$

Maximizing profit gives:

$$z_c = \sqrt{\frac{1}{\gamma}} - z_0.$$

Under the actual network with independence number m, the realized profit is:

$$\Pi_c = mp_i - \gamma z_c = \frac{m}{z_0} + \gamma z_0 - (m+1)\sqrt{\gamma}.$$

(iii) Seller has perfect information about the network. From Theorem 1, the optimal profit when the seller knows the network structure is:

$$\Pi_p = \frac{m}{z_0} + \gamma z_0 - 2\sqrt{\gamma m}.$$

□

Proof of Proposition 8. Since $z_0 \leq \frac{1}{4\sqrt{\gamma}} < \frac{1}{2\sqrt{\gamma}} \cdot \frac{n+1}{2n+1}$ for any n, the expressions for profits under different information assumptions from Lemma 4 are valid. We recall the following:

$$\mathcal{V}_e(m) \equiv \Pi_p - \Pi_e = \frac{\sqrt{\gamma}(\sqrt{n} - \sqrt{m})^2}{\sqrt{n}},$$
$$\mathcal{V}_c(m) \equiv \Pi_p - \Pi_c = \sqrt{\gamma}(\sqrt{m} - 1)^2.$$

Both $\mathcal{V}_e(m)$ and $\mathcal{V}_c(m)$ are monotonic in m: $\mathcal{V}_e(m)$ is decreasing and $\mathcal{V}_c(m)$ is increasing. The bounds follow directly:

Part (i): Since $\mathcal{V}_e(m)$ is decreasing in m, we have $\mathcal{V}_e(m) \le \mathcal{V}_e(1) = \frac{\sqrt{\gamma}(\sqrt{n}-1)^2}{\sqrt{n}}$.

Part (iii): Since $\mathcal{V}_c(m)$ is increasing in m, we have $\mathcal{V}_c(m) \le \mathcal{V}_c(n) = \sqrt{\gamma}(\sqrt{n}-1)^2$.

We now prove parts (ii) and (iv), concerning the relative inefficiency ratios.

Part (ii): The ratio of realized profit under the empty network assumption to the optimal profit is:

$$\frac{\Pi_e}{\Pi_p} = 1 - \frac{z_0\sqrt{\gamma}(\sqrt{m}-\sqrt{n})^2}{\sqrt{n}(\sqrt{m}-z_0\sqrt{\gamma})^2}.$$

Note that $\sqrt{m} - z_0\sqrt{\gamma} > 0$ and is increasing in m, while $(\sqrt{m}-\sqrt{n})^2$ is decreasing in m. Hence, $\frac{\Pi_e}{\Pi_p}$ is increasing in m, and the difference

$$\frac{\mathcal{V}_e(m)}{\Pi_p} = 1 - \frac{\Pi_e}{\Pi_p}$$

is decreasing in m. Evaluating at $m = 1$, the upper bound is:

$$\frac{\mathcal{V}_e(m)}{\Pi_p} \le \frac{z_0\sqrt{\gamma}(\sqrt{n}-1)^2}{\sqrt{n}(1-z_0\sqrt{\gamma})^2}.$$

Part (iv): The ratio of realized profit under the complete network assumption to the optimal profit is:

$$\frac{\Pi_c}{\Pi_p} = 1 - \frac{z_0\sqrt{\gamma}(\sqrt{m}-1)^2}{(\sqrt{m}-z_0\sqrt{\gamma})^2}.$$

We rewrite:

$$\frac{\sqrt{m}-1}{\sqrt{m}-z_0\sqrt{\gamma}} = 1 - \frac{1-z_0\sqrt{\gamma}}{\sqrt{m}-z_0\sqrt{\gamma}},$$

which is increasing in m. Therefore, $\frac{\Pi_c}{\Pi_p}$ is decreasing in m, and the inefficiency ratio

$$\frac{\mathcal{V}_c(m)}{\Pi_p} = 1 - \frac{\Pi_c}{\Pi_p}$$

is increasing in m. Evaluating at $m = n$, we obtain:

$$\frac{\mathcal{V}_c(m)}{\Pi_p} \le \frac{z_0\sqrt{\gamma}(\sqrt{n}-1)^2}{(\sqrt{n}-z_0\sqrt{\gamma})^2}.$$

Furthermore, since $\frac{\sqrt{n}-1}{\sqrt{n}-z_0\sqrt{\gamma}}$ is increasing in n, we find that $\left(z_0 < \frac{1}{2\sqrt{\gamma}}\frac{n+1}{2n+1} \text{ so } z_0\sqrt{\gamma} < 2/3\right)$:

$$\lim_{n\to\infty} \frac{z_0\sqrt{\gamma}(\sqrt{n}-1)^2}{(\sqrt{n}-z_0\sqrt{\gamma})^2} = z_0\sqrt{\gamma}.$$

Given the assumption that $z_0\sqrt{\gamma} \le \frac{2}{3}$, it follows that:

$$\frac{\mathcal{V}_c(m)}{\Pi_p} \le \frac{2}{3}.$$

◇