Alternative data market regulation: navigating the fine line between finding alpha and insider trading. [*]

Lina Lukyantseva

December 4, 2020

Abstract

I study a model in which traders buy information from a monopolistic seller which is later used to trade in the financial market. My model is based on the single-period version of Kyle (1985). The data seller can sell the signal of different quality, and the traders are heterogeneous in their ability to interpret the purchased signal. The regulator's utility function depends on the resulting price informativeness and the welfare of the liquidity traders. I show that trading on information bought through data markets has similar effects on price informativeness and the welfare of the liquidity traders as insider trading. Liquidity traders suffer even more if a small number of traders use the information than if only one trader uses it. I study optimal data seller's behavior in the absence of regulation and under different regulation regimes. Requiring the data seller to sell the data through an auction with payments contingent on the traders' future profits and setting a lower bound on the quantity that needs to be sold ex ante increases both price informativeness and liquidity traders' welfare.

[*] I am grateful to Mohammad Akbarpour, Jeremy Bulow, Ishaan Nerurkar and Michael Ostrovsky for helpful discussions. I thank Anat Admati and Paul Pfleiderer for their insightful comments.

1 Introduction

Website scraping, credit/debit card transactions, app usage, geolocation, satellite imagery – these are just some examples of alternative data that investment managers spend millions of dollars on every year. The alternative data market has more that quadrupled in the last three years – from $400m in 2017 to $1700m in 2020 (alternativedata.org), and is expected to grow at a compound annual growth rate of 40% from 2020 to 2027 (grandviewresearch.com). Such a burgeoning new market poses many questions for regulators. One such question is: where is the line between finding alpha and trading on insider information? Let's consider two real scenarios:

1. A data analyst for Capital One downloaded and analyzed data on retail purchases made with Capital One credit cards. He used this information to predict revenues of retailers; then, he traded retailers' stocks in advance of the public release of quarterly sales announcements by these companies. (SEC v. Huang (2016))

2. " Speaking at a conference on alternative investments at the London School of Economics, Matthew Granade, Point72's chief market intelligence officer, bragged that they scrutinise 80m credit card transactions every day. Coupled with satellite images that can scan car parks and geolocation data from mobile phones to show how many people are visiting various stores, the investment group can get a real-time idea of how companies are doing, long before their results are released.

   One LSE student asked how all this data could help Point72 if everyone had access to the same information. The answer was exclusivity agreements, Mr Granade said: "The great thing about this area is you can arrange deals where you are the only ones who get it."" ("Hedge funds see a gold rush in data mining", Financial times (2017))

The two stories are very similar – in both of them someone obtained a unique access to the credit card transactions data and used it to trade in the financial markets. What would we expect the legal consequences to be in these two cases?

Most (if not all) regulations in the world would classify the first story as a classic example of an illegal insider trading, and indeed the data analyst was found liable. The second story is, however, more controversial. For instance, the American and European regulations would disagree on whether it represents an instance of illegal insider trading. Paragraph 28 of the Market Abuse Regulation (EU) states that "Research and estimates based on publicly available data, should not per se be regarded as inside information and the mere fact that a transaction is carried out on the basis of research or estimates should not therefore be deemed to constitute use of inside information. However, for example, where the publication or distribution of information is routinely expected by the market and where such publication or distribution contributes to the price-formation process of financial instruments, or the information provides views from a recognised market commentator or institution which may inform the prices of related financial instruments, the information may constitute inside information. *Market actors must therefore consider the extent to which the information is non-public* and the possible effect on financial instruments traded in advance of its publication or distribution, to establish whether they would be trading on the basis of inside information."

In contrast, U.S. Securities and Exchange Commission provides the following definition: " Illegal insider trading refers generally to buying or selling a security, in breach of a fiduciary duty or other relationship of trust and confidence, on the basis of material, nonpublic information about the security."

As one can see, the European regulation prohibits trading on any material, nonpublic information while the U.S. regulation allows trading on material, nonpublic information as long as that information was obtained legally (for example, bought from a bank which clients agreed that their information can be shared). Which of these two policies makes more sense? And is there a better alternative? In this paper I propose an answer to the question: "why regulate data markets?". I also highlight some of the trade-offs that the regulators face when choosing between different policies.

My set up consists of a data market, a financial market, and a production economy. One risky asset is being traded in the financial market as in the one-period version of Kyle (1985). The model assumes there are two types of traders. The first type maximizes their profit (e.g. hedge funds). The second type trades for other reasons: for example, as a way of spending time during the pandemic (liquidity traders) and whose demand is random. The data seller can provide a dataset containing some useful information about the future realization of the payoff of the asset. The data seller can choose the quality of the data that he sells. The data needs to be interpreted in order to make an inference about the future payoff. The hedge funds are heterogeneous in their interpretation skills – after observing the same dataset some of them make more precise inferences than others. The regulator sets the rules which the data seller has to follow when selling the dataset to the hedge funds. The hedge funds then use the data to trade in the financial market. The resulting market price can be characterized in terms of how much information it contains. I measure price informativeness as the covariance between the price and the payoff. In general, price informativeness is increasing in the number of hedge funds who get access to the data, in their competence (i.e. data interpretation skills), and in the initial quality of the data. The amount of information contained in the asset's market price impacts the efficiency of resource allocation in the economy and, therefore, the output (Wurgler (2000), David et al. (2016)). When prices are informative, the capital is more likely to be invested in successful projects rather than those that are going to fail. This results in a higher output. [1]

The regulator's utility is increasing in the output (and, therefore, price informativeness) and in the welfare of the liquidity traders. Why does the regulator care about liquidity traders' profit even though it is just a transfer from one market participant to another? One reason could be fairness (in expectation hedge funds receive their profits at the expense of the liquidity traders), another could be that the liquidity traders might leave the market if their losses become too high which would adversely impact the overall economy (as the amount of liquidity in the capital markets would be significantly reduced).

I first study the benchmark scenario of no regulation. In the absence of regulation it is optimal for the data seller to choose the highest quality of the data and sell it to an exclusive set of the most competent funds by making Take-It-or-Leave-It offers. In this case liquidity traders' welfare

[1] The assumption here is that the information gets reflected in the prices substantially earlier than it would had the insider trading not occurred. An example of this is when a dataset helps to predict companies' performance months in advance of its quarterly/yearly releases. It is hard to justify selling access to exclusive information seconds before other traders get it by an increase in price efficiency. This might be why there has been so far more regulator's action in the domain of selling data for high-frequency trading: for instance, in 2013 New York Attorney General Eric Schneiderman requested that Thomson Reuters stops selling exclusive access to their consumer sentiment data to a small group of clients 2 seconds before its other clients.

achieves its absolute minimum, i.e. several hedge funds trading on exclusive data can hurt the liquidity traders even more than just one insider trader. Though, the price informativeness in this case is higher than if there was one insider trader. Therefore, it is possible that the regulator's utility function has the following property: In the case of selling the data through data markets, the upside of increased price informativeness offsets the downside of decreased liquidity traders' welfare, but in the case of insider trading, this does not happen. However, I show that it is possible for the regulator to ex ante increase both the price informativeness and the welfare of the liquidity traders by choosing the right policy.

Is the E.U. regulation an improvement on the U.S. one? Not necessarily. The European regulation does not specify which mechanisms should be used when selling the data, however, it says that the data should be publicly available. Since it is legal to sell, for example, Bloomberg Terminal or Refinitiv Eikon, I interpret publicly available as non-discriminatory rather than free. I therefore model the E.U. regulation as requiring the data seller to set a fixed price and sell to everyone who is willing to purchase the data at this price. It is intuitive that the data seller can still limit the number of buyers by simply setting the price high enough. Even though it is possible that such a policy incentivizes the data seller to sell to more buyers, in many cases he will choose to sell to less buyers than if there was no regulation. Including an additional buyer is now even more costly since the amount of money charged to other buyers needs to be reduced, not only to the extent that their profits got hurt by an increased competition, but to the price level acceptable to the new buyer. Moreover, if in the absence of regulation it was optimal for the data seller to sell to only one hedge fund, he would always continue doing so under the European regulation.

A natural alternative suggestion for the regulator would be to set a lower bound on the number of the datasets sold, but still allow the data seller to choose who to sell to and at what prices. An issue with this policy is that it may be no longer optimal for the data seller to sell to the most competent funds. If there are enough extremely incompetent funds in the market, the data seller will sell to just enough of the incompetent funds to satisfy the regulator, and to the set of funds who he would have sold to in the absence of regulation. Neither price informativeness nor liquidity traders' welfare would change in any significant way, and so the policy would end up being ineffective.

The above issue can be fixed if the regulator does not only set a lower bound on the quantity, but also imposes a mechanism on the data seller (such as an auction). It is important to choose a right auction format. One issue that may arise is as follows: suppose data seller sells K datasets through a $K + 1$ price auction K, then, the most competent funds will indeed win in equilibrium. However, this creates a new challenge for the regulator: it can be optimal for the data seller to make data noisy.

Finally, I study an auction with contingent payments and a lower bound on the quantity in which the funds bid a share of their future profit rather than a dollar amount. This auction format is based on Hansen (1985) and DeMarzo et al. (2005). I show that when such an auction is used, a subset of the most competent funds receive the data in equilibrium, and the data seller chooses to provide the highest quality of the data. This policy ex ante increases both the price informativeness and the welfare of the liquidity traders.

1.1 Related work

Financial market is the core element of the model, and the paper is directly related to the literature on one-period version of Kyle (1985). In Kyle (1985) there is one informed trader who knows the realization of the asset's payoff exactly. Admati and Pfleiderer (1988) study a model with multiple informed traders who observe the payoff with some error such that the realization of the error is the same for everyone. Dridi and Germain (2009) study a model with multiple informed traders who observe the payoff with some error but the errors are independent among the informed traders and the variances of the errors are allowed to be different. A special case of my signal structure when the hedge funds' interpretation errors are equal to zero is described by Admati and Pfleiderer (1988). A special case of my signal structure when the data seller chooses the highest quality of the data is described by Dridi and Germain (2009). In Jain and Mirman (1999) the market makers observe not only the aggregate demand but also a separate signal about the payoff. Lambert et al. (2017) provide a general framework in which the strategic traders and the market makers observe multidimensional signals that are jointly normally distributed but allowed to have an arbitrary covariance matrix. Caballe and Krishnan (1994) and Pasquariello (2007) study a one-period model in which several assets are being traded but require the traders to have symmetric information. There is a separate body of literature on the dynamic version of Kyle (1985) and Kyle (1989) in which the traders submit demand and supply curves instead of market orders.

This paper relates to a general topic of markets for information (see Bergemann et al. (2019) for a survey) and, in particular, markets for information that is used for trading. The seminal papers are by Admati and Pfleiderer (1986, 1987, 1988). The closest paper to my setting in the absence of regulation is Admati and Pfleiderer (1988) in which a monopolist sells information that is later being used for trading in one-period Kyle (1985) setting. Our results overlap when the strategic traders are risk-neutral and receive the same signal. In this scenario, it is optimal for the data seller to provide the data of the highest quality and sell to only one buyer (or trade in the market himself). Outside of this scenario, Admati and Pfleiderer (1988) show that it is optimal to sell to more than one buyer if the traders are risk-averse. I show that it is optimal to sell to more than one buyer when the traders are risk-neutral but observe different signals (i.e. their errors are independent). Garcia and Sangiorgi (2011) generalize Admati and Pfleiderer (1988) by allowing the data seller to add personalized noise. Chen and Wilhelm Jr (2012) consider a setting in which multiple traders have information that they can use for trading themselves or sell to other traders. In Foucault and Lescourret (2003) and Eren and Ozsoylev (2006) informed traders can share their information with each other rather than sell it.

Finally, this paper is related to the literature on insider trading. Shin (1996) and DeMarzo et al. (1998) solve for the optimal investigation and punishment policy when the regulator's objective is to maximize the welfare of the liquidity traders. In Shin (1996) there are an insider who knows the realization of the payoff and a market professional who observes a noisy signal and can improve the precision of their signal at a cost. Even though the regulation is costless it is optimal to still allow some amount of insider trading since it reduces the incentives of the market professional to invest in the quality of their signal and, hence, reduces the expected loss for the liquidity traders. DeMarzo et al. (1998) also conclude that not every trade should be investigated but rather because investigation is costly. Their optimal policy entails investigations following large trading volumes or large price movements or both. Manne (1969), Carlton and Fischel (1983), Leland (1992) argue that insider trading contributes to more informative prices. Shin (1996) and Fishman and Hagerty (1992) show that it is possible that insider trading decreases price informativeness since it reduces the incentives of other traders to acquire information and trade. Ausubel (1990) argues that insider traders may be themselves interested in insider trading regulation in order to incentivize other market participants to invest more without fearing being taken advantage of.

2 The model

The set up consists of a data market, a financial market and a production economy. The financial market is a modification of the one-period version of the set up introduced in Kyle (1985). One risky asset is being traded among three types of players: N hedge funds, liquidity traders and market makers. The distribution of the asset's payoff is commonly known: $\tilde{v} \sim N(\overline{v}, \sigma^2)$. The players in the data market are the same N hedge funds, a data seller (he) and a regulator (she).

Before the trading happens the data seller can produce a dataset at a fixed cost $R > 0$ and sell it to a subset of the funds. I assume that there is no resell of information and that the data seller cannot trade on the information himself. The dataset contains an unbiased estimate of the payoff $\tilde{v} + \eta$, where $\eta \sim N(0, c_\eta \sigma^2)$. The data seller can choose the value of c_η: $c_\eta = 0$ corresponds to no additional noise; the greater c_η the noisier the information. In practice adding noise can mean not properly cleaning the data or purposefully providing only a part of the available data. η is an objective measure of the quality of the data, higher c_η makes the data less informative for everyone. If fund i purchases the dataset, it observes $\tilde{v} + \eta + \epsilon_i$, where $\epsilon_i \sim N(0, c_i \sigma^2)$ is the fund-specific interpretation error. ϵ_i is independent of all other variables in the model. The interpretation is that the funds that have better data science teams/complementary datasets will have lower c_i. I will use the following terminology: if for two funds i and j $c_i < c_j$, I will say that i is more competent than j. I will refer to c_η as data error and to c_i as analyst i error for brevity (even though it would be correct to say data error variance and analyst error variance instead).

In the financial market the trading happens in two steps: first, hedge funds and liquidity traders simultaneously place market orders – quantities of the asset they they want to buy/sell, then market makers observe the aggregate demand and compete for the opportunity to fulfill it, this drives the market price to be equal to the expectation of $\tilde{v}$ conditional on the observed aggregate demand. Liquidity traders' demand y_l is random, $y_l \sim N(0, \sigma_l^2)$. Hedge fund i maximizes its expected profit given I_i – the information it has about the realization of $\tilde{v}$. If fund i has access to the data, $I_i = \tilde{v} + \eta + \epsilon_i$, otherwise, it does not have any additional information besides the commonly known distribution of $\tilde{v}$. The hedge funds do not behave as price takers, they understand that the size of their orders influences the market price. Fund i, therefore, chooses its demand y_i by solving the following problem:

$$\max_{y_i} \mathbb{E}\left(y_i \cdot \left(\tilde{v} - \mathbb{E}\left(\tilde{v} \mid \sum_{j=1}^{N} y_j + y_l\right)\right) \mid I_i\right),$$

where $\mathbb{E}\left(\tilde{v} \mid \sum_{j=1}^{N} y_j + y_l\right) = P$ is the resulting market price.

The market price can be characterized in terms of its informativeness. I define price informativeness (PI) as $Cov(\tilde{v}, P)$. The output in the economy Y is some increasing function of PI.

The regulator's utility function is $u_{reg} = g(Y(PI), \pi_l)$, where π_l is the expected profit of the liquidity traders. I assume that g is increasing in both arguments.

The timing of the game is as follows:

Step 1: The regulator chooses a policy;

Step 2: The profile of the analyst errors $\{c_1, ..., c_N\}$ is realized (c_i is drawn i.i.d. from some distribution F on $[0, \infty)$) and becomes publicly known;

Step 3: The data seller decides whether to collect a dataset at a fixed commonly known cost $R > 0$.

If he does not, then the game proceeds to step 5.

If he does, he also chooses $c_\eta \geq 0$. Then the game proceeds to step 4;

Step 4: Some funds purchase the data from the data seller (in accordance with the policy);

Step 5: The play in the financial market happens. The profits and the utility of the regulator are realized.

2.1 Discussion of the assumptions

One assumption is that the data seller's cost is independent of the quality of the data. The rationale for this assumption is that the fixed costs of producing the data are usually much higher than the variable costs: in many cases the data is a byproduct of the main business (for example, record of financial transactions for the banks or medical records for the insurance companies) in which case the data seller chooses the quality of the data by deciding how much of the available data he is willing to share. Even if the data seller is actually producing the data, it is costly to write code for scraping the Internet or extracting information from satellite images but it is much less costly to then apply it to additional websites/GPS coordinates to increase the precision of the signal.

Two strongest assumptions are that the funds' competence levels are publicly known and that all the funds know who purchased the data before the trading happens. These assumptions are not particularly realistic but understanding the mechanics of the model in this setting is useful for the future research that will relax these assumptions.

Finally, all random variables in the model are assumed to be normally distributed which is standard in the literature on Kyle (1985).

3 Financial market

3.1 Funds' perspective

Definition 3.1. A profile of market orders of the hedge funds $y = \{y_1, ..., y_N\}$ and a market price P constitute an equilibrium if the following conditions hold:

(1) The market price is equal to the expected payoff of the asset conditional on observing the order flow: $P = \mathbb{E}(\tilde{v} \mid \sum_{i=1}^{N} y_i)$;

(2) The market order y_i of an informed fund i maximizes its expected profit conditional on observing its signal from the data and given the primitives of the model and the market orders of other players:

$$y_i \in \arg\max_z \mathbb{E}\left(z \cdot \left(\tilde{v} - \mathbb{E}\left(\tilde{v} \mid z + \sum_{j \neq 1} y_j + y_l\right)\right) \mid \tilde{v} + \eta + \epsilon_i\right);$$

(3) The market order y_i of an uninformed fund i maximizes its expected profit given the primitives of the model and the market orders of other players:

$$y_i \in \arg\max_z \mathbb{E}\left(z \cdot \left(\tilde{v} - \mathbb{E}\left(\tilde{v} \mid z + \sum_{j \neq 1} y_j + y_l\right)\right)\right).$$

Proposition 1. [2] *Given a set of informed hedge funds $\mathbb{I}$, the following profile of market orders $y = \{y_1, ..., y_N\}$ and market price P constitute an equilibrium:*

(1) $P = \overline{v} + \lambda(\sum_{i=1}^N y_i + y_l)$;

(2) If fund i is uninformed (i.e. $i \notin \mathbb{I}$), then $y_i = 0$;

(3) If fund i is informed (i.e. $i \in \mathbb{I}$), then $y_i = \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v})$;

where

$$\alpha_i = \frac{1}{(1 + c_\eta + 2c_i)\left(1 + \sum_{j \in \mathbb{I}} \frac{1}{1 + \frac{2c_j}{1+c_\eta}}\right)} \tag{3.1}$$

and

$$\lambda = \frac{\sigma}{\sigma_l}\sqrt{\sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)}.[3] \tag{3.2}$$

Proof. See Appendix A.1. □

Proposition 1 states that the optimal market order of an informed fund i depends on which other funds are informed. The following proposition provide some insight into the direction of that dependence:

Proposition 2. *If $i, j \in \mathbb{I}$ if $c_i < c_j$, then (1) $\alpha_i > \alpha_j$ and (2) as $c_\eta \to \infty$, $\frac{\alpha_i}{\alpha_j} \to 1$.*

Proof. (1) By 3.1 for any $i, j \in \mathbb{I}$

$$\alpha_j = \frac{1 + c_\eta + 2c_i}{1 + c_\eta + 2c_j}\alpha_i. \tag{3.3}$$

Clearly, if $c_i < c_j$, then $\alpha_i > \alpha_j$.

[2] Dridi and Germain (2009) provide the analogues of Propositions 1-6 in the special cade when $\eta = 0$.

[3] Since both α_i and λ are functions of $\mathbb{I}$ and c_η, Iwill generally use the notation $\alpha_i(\mathbb{I}, c_\eta)$ and $\lambda(\mathbb{I}, c_\eta)$. However, sometimes $\mathbb{I}$ and c_η might be fixed and clear from the context, in which case Iwill use α_i and λ to simplify notation. If c_η is fixed but the set of informed funds is changing, Iwill write $\alpha_i(\mathbb{I})$ for clarity.

(2)

$$\lim_{c_\eta \to \infty} \frac{\alpha_i}{\alpha_j} = \lim_{c_\eta \to \infty} \frac{\frac{1+2c_j}{c_\eta} + 1}{\frac{1+2c_i}{c_\eta} + 1} = 1.$$

□

Proposition 2 implies that higher precision funds have larger demand coefficients ($\alpha_i > \alpha_j$ implies that $\frac{\alpha_i}{\lambda} > \frac{\alpha_j}{\lambda}$). It does not necessarily mean that i's market order is larger than j's market order since in a particular realisation ϵ_i could be very small and ϵ_j could be very large. But since on average a high signal is more likely to be caused by a high asset's payoff for i than it is for j, i "trusts" its signal more. However, this difference becomes less pronounced with the amount of added noise. Intuitively, if the data is very noisy, it does not really matter who has better interpretation skills.

Given the equilibrium market orders and market price, Ican derive π_i – the expected profit of an informed fund i:

Proposition 3. $\pi_i = \frac{\sigma^2 \alpha_i^2}{\lambda}(1 + c_\eta + c_i)$ [4].

Proof.

$$\pi_i = \mathbb{E}\left(y_i \cdot (\tilde{v} - P)\right) = \mathbb{E}\left(y_i \cdot \left(\tilde{v} - \overline{v} - \lambda\left(\sum_{i=1}^{N} y_j + y_l\right)\right)\right) =$$

$$= \mathbb{E}\left(\frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v}) \cdot \left(\tilde{v} - \overline{v} - \lambda\left(\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v}) + y_l\right)\right)\right) =$$

$$= \frac{\alpha_i}{\lambda}\left(\mathbb{E}\left((\tilde{v} - \overline{v})^2\right) - \alpha_i \mathbb{E}\left((\tilde{v} + \eta + \epsilon_i - \overline{v})^2\right) - (\sum_{j \in \mathbb{I}, j \neq i} \alpha_j)\mathbb{E}\left(\tilde{v} + \eta - \overline{v}\right)\right) =$$

$$= \frac{\alpha_i}{\lambda}\left(\sigma^2 - \alpha_i(\sigma^2 + c_\eta \sigma^2 + c_i \sigma^2) - (\sum_{j \in \mathbb{I}, j \neq i} \alpha_j)(\sigma^2 + c_\eta \sigma^2)\right) =$$

$$= \frac{\sigma^2 \alpha_i}{\lambda}\left(1 - \alpha_i(1 + c_\eta + c_i) - (\sum_{j \in \mathbb{I}, j \neq i} \alpha_j)(1 + c_\eta)\right) = \frac{\sigma^2 \alpha_i^2}{\lambda}(1 + c_\eta + c_i),$$

[4] Since π_i is a function of $\mathbb{I}$ and c_η, Iwill generally use the notation $\pi_i(\mathbb{I}, c_\eta)$. However, sometimes $\mathbb{I}$ and c_η might be fixed and clear from the context, in which case Iwill use π_i to simplify notation.

where the last equality holds since $1 - (\sum_{j \in \mathbb{I}, j \neq i} \alpha_j)(1 + c_\eta) = 2(1 + c_\eta + c_i)\alpha_i$ by A.2. □

Proposition 4. *If $i, j \in \mathbb{I}$ and $c_i < c_j$, then $\pi_i > \pi_j$.*

Proof. By 3.1

$$\alpha_j = \frac{1 + c_\eta + 2c_i}{1 + c_\eta + 2c_j} \alpha_i,$$

then

$$\alpha_i^2(1+c_\eta+c_i) - \alpha_j^2(1+c_\eta+c_j) = \alpha_i^2 \left(1 + c_\eta + c_i - \frac{(1 + c_\eta + c_j)(1 + c_\eta + 2c_i)}{1 + c_\eta + 2c_j}\right) = \frac{\alpha_i^2(1 + c_\eta)(c_j - c_i)}{1 + c_\eta + 2c_j} > 0.$$

Hence, $\pi_i = \frac{\sigma^2 \alpha_i^2}{\lambda}(1 + c_\eta + c_i) > \frac{\sigma^2 \alpha_j^2}{\lambda}(1 + c_\eta + c_j) = \pi_j$. □

Proposition 4 compares the expected profits of two different informed funds and shows that higher precision funds will enjoy higher expected profits. The following propositions provide some complementary comparative statics results for a particular fund:

Proposition 5. *The expected profit of an informed fund i $\pi_i = \frac{\sigma^2 \alpha_i^2}{\lambda}(1 + c_\eta + c_i)$ is decreasing in c_i given the variances of the errors of other informed funds c_{-i}.*

Proof. Consider c_i and $c_i\prime > c_i$. Let $\alpha_i, \alpha_j, \lambda$ $(\alpha_i\prime, \alpha_j\prime, \lambda\prime)$ be the equilibrium parameters associated with the profile of the error variances c_i, c_{-i} $(c_i\prime, c_{-i})$.

Case 1: $\lambda\prime > \lambda$. Using 3.1,

$$\alpha_i^2(1 + c_\eta + c_i) = \frac{1 + c_\eta + c_i}{\left(2(1 + c_\eta + c_i) + (1 + c_\eta + 2c_i) \sum_{j \in \mathbb{I}, j \neq i} \frac{1}{1 + \frac{2c_j}{1 + c_\eta}}\right)^2}.$$

Let $s = \sum_{j \in \mathbb{I}, j \neq i} \frac{1}{1 + \frac{2c_j}{1 + c_\eta}}$, note that $s > 0$. Then

$$\frac{\partial}{\partial c_i} \alpha_i^2(1 + c_\eta + c_i) = \frac{-(1 + c_\eta)(3s + 2) - 2(s + 1)c_i}{(2(1 + c_\eta + c_i) + (1 + c_\eta + 2c_i)s)^3} < 0.$$

Therefore, $\alpha_i^2(1 + c_\eta + c_i)$ is decreasing in c_i which implies that $\alpha_i\prime^2(1 + c_\eta + c_i\prime) < \alpha_i^2(1 + c_\eta + c_i)$. And since $\lambda\prime > \lambda$, $\pi_i\prime = \frac{\sigma^2 \alpha_i\prime^2(1 + c_\eta + c_i\prime)}{\lambda\prime} < \frac{\sigma^2 \alpha_i^2(1 + c_\eta + c_i)}{\lambda} = \pi_i$.

Case 2: $\lambda' \leq \lambda$.

Step 1: $\alpha_j' > \alpha_j$ since by 3.1,

$$\frac{1}{(1+c_\eta+2c_j)\left(1+\sum_{k\in\mathbb{I},k\neq i}\frac{1}{1+\frac{2c_k}{1+c_\eta}}+\frac{1}{1+\frac{2c_i'}{1+c_\eta}}\right)} > \frac{1}{(1+c_\eta+2c_j)\left(1+\sum_{k\in\mathbb{I},k\neq i}\frac{1}{1+\frac{2c_k}{1+c_\eta}}+\frac{1}{1+\frac{2c_i}{1+c_\eta}}\right)}.$$

Step 2: Since $\lambda' \leq \lambda$ and $\alpha_j' > \alpha_j$, $\pi_j' = \frac{\sigma^2\alpha_j'^2(1+c_\eta+c_j)}{\lambda'} > \frac{\sigma^2\alpha_j^2(1+c_\eta+c_j)}{\lambda} = \pi_j$.

Step 3:

$$\sum_{i\in\mathbb{I}}\pi_i = \sum_{i\in\mathbb{I}}\frac{\sigma^2\alpha_i^2(1+c_\eta+c_i)}{\lambda} = \frac{\sigma^2\sum_{i\in\mathbb{I}}\alpha_i^2(1+c_\eta+c_i)}{\frac{\sigma}{\sigma_l}\sqrt{\sum_{i\in\mathbb{I}}\alpha_i^2(1+c_\eta+c_i)}} = \sigma\sigma_l\sqrt{\sum_{i\in\mathbb{I}}\alpha_i^2(1+c_\eta+c_i)} = \lambda\sigma_l^2. \tag{3.4}$$

Step 4: Since $\lambda' \leq \lambda$, the sum of all the expected profits weakly decreases when i becomes less precise. But the expected profit of everyone except i increases. Therefore, it must be that i's expected profit decreases, i.e. $\pi_i' < \pi_i$.

□

Proposition 6. *For any two informed funds i and j i's expected profit $\pi_i = \frac{\sigma^2\alpha_i^2}{\lambda}(1+c_\eta+c_i)$ is increasing in c_j.*

Proof. Consider c_j and $c_j' > c_j$. Let $\alpha_i, \alpha_j, \lambda$ $(\alpha_i', \alpha_j', \lambda')$ be the equilibrium parameters associated with the profile of the error variances c_j, c_{-j} (c_j', c_{-j}).

Case 1: $\lambda' \leq \lambda$. Note that $\alpha_i' > \alpha_i$ since

$$\frac{1}{(1+c_\eta+2c_i)\left(1+\sum_{k\in\mathbb{I},k\neq j}\frac{1}{1+\frac{2c_k}{1+c_\eta}}+\frac{1}{1+\frac{2c_j'}{1+c_\eta}}\right)} > \frac{1}{(1+c_\eta+2c_i)\left(1+\sum_{k\in\mathbb{I},k\neq j}\frac{1}{1+\frac{2c_k}{1+c_\eta}}+\frac{1}{1+\frac{2c_j}{1+c_\eta}}\right)}$$

Therefore, $\pi_i' = \frac{\sigma^2\alpha_i'^2}{\lambda'}(1+c_\eta+c_i) > \frac{\sigma^2\alpha_i^2}{\lambda}(1+c_\eta+c_i) = \pi_i$.

Case 2: $\lambda' > \lambda$. By 3.4 it means that the sum of the profits of the informed funds increased. By Proposition 5 j's profit decreased. Hence, the sum of the profits of all the informed funds without

j increased. For any $i \in \mathbb{I} \setminus \{j\}$

$$\frac{\partial \pi_i}{\partial c_j} = \frac{\sigma^2(1 + c_\eta + c_i)}{1 + c_\eta + 2c_i} \cdot \frac{\partial}{\partial c_j}\left(\frac{1}{\lambda\left(1 + \sum_{k\in\mathbb{I}} \frac{1}{1+\frac{2c_k}{1+c_\eta}}\right)}\right)$$

and

$$\pi_i\prime - \pi_i = \int_{c_j}^{c_j\prime} \frac{\partial \pi_i}{\partial c_m} dc_m = \frac{\sigma^2(1 + c_\eta + c_i)}{1 + c_\eta + 2c_i} \int_{c_j}^{c_j\prime} \frac{\partial}{\partial c_m}\left(\frac{1}{\lambda\left(1 + \sum_{k\in\mathbb{I}} \frac{1}{1+\frac{2c_k}{1+c_\eta}}\right)}\right) dc_m.$$

Note that for all $i \in \mathbb{I} \setminus \{j\}$ the term under the integral is the same, and $\frac{\sigma^2(1+c_\eta+c_i)}{1+c_\eta+2c_i} > 0$. Hence, the direction of the change of the profit is the same for all the informed funds without j. Since the sum of their profits increased, it needs to be the case that $\pi_i\prime > \pi_i$ for all $i \in \mathbb{I} \setminus \{j\}$. □

Proposition 7. *For any set of informed funds $\mathbb{I}$, any $i \in \mathbb{I}$ and any $j \in \mathbb{F} \setminus \mathbb{I}$ $\pi_i(\mathbb{I} \cup \{j\}, c_\eta) < \pi_i(\mathbb{I}, c_\eta)$.*

Proof. Suppose towards a contradiction that there exists a set of informed funds $\mathbb{I}$, some $i \in \mathbb{I}$ and some $j \in \mathbb{F} \setminus \mathbb{I}$ such that $\pi_i(\mathbb{I} \cup \{j\}, c_\eta) \geq \pi_i(\mathbb{I}, c_\eta)$. Let $\pi_i(\mathbb{I} \cup \{j\}, c_\eta) = \pi_i^0$. Then since by Proposition 6 $\pi_i(\mathbb{I} \cup \{j\}, c_\eta)$ is increasing in c_j, $\lim_{c_j\to\infty} \pi_i(\mathbb{I} \cup \{j\}, c_\eta) > \pi_i^0$.

However,

$$\lim_{c_j\to\infty} \alpha_j(\mathbb{I} \cup \{j\}, c_\eta) = \lim_{c_j\to\infty} \frac{1}{(1 + c_\eta + 2c_j)\left(1 + \sum_{k\in\mathbb{I}\cup\{j\}} \frac{1}{1+\frac{2c_k}{1+c_\eta}}\right)} = 0$$

and for all $m \in \mathbb{I}$

$$\lim_{c_j\to\infty} \alpha_m(\mathbb{I} \cup \{j\}, c_\eta) = \lim_{c_j\to\infty} \frac{1}{(1 + c_\eta + 2c_m)\left(1 + \sum_{k\in\mathbb{I}\cup\{j\}} \frac{1}{1+\frac{2c_k}{1+c_\eta}}\right)} = \alpha_m(\mathbb{I}, c_\eta)$$

Therefore,

$$\lim_{c_j\to\infty}\lambda(\mathbb{I}\cup\{j\},c_\eta)=\lambda(\mathbb{I},c_\eta)$$

and

$$\lim_{c_j\to\infty}\pi_i(\mathbb{I}\cup\{j\},c_\eta)=\sigma^2(1+c_\eta+c_i)\lim_{c_j\to\infty}\left(\frac{\alpha_i^2(\mathbb{I}\cup\{j\},c_\eta)}{\lambda(\mathbb{I}\cup\{j\},c_\eta)}\right)=\sigma^2(1+c_\eta+c_i)\left(\frac{\alpha_i^2(\mathbb{I},c_\eta)}{\lambda(\mathbb{I},c_\eta)}\right)=\pi_i^0,$$

contradiction. □

Corollary 7.1. *For any set of informed funds $\mathbb{I}$, any $i\in\mathbb{I}$ and any set $\mathbb{X}\subseteq\mathbb{F}\setminus\mathbb{I}$ $\pi_i(\mathbb{I}\cup\mathbb{X},c_\eta)<\pi_i(\mathbb{I},c_\eta)$.*

Proof. The result follows by adding the elements of $\mathbb{X}$ one by one and applying Proposition 7. □

3.2 Regulator's perspective

3.2.1 Price informativeness

Proposition 8. *$PI=\sigma^2\sum_{i\in\mathbb{I}}\alpha_i$.*

Proof.

$$PI=Cov(\tilde{v},P)=Cov\left(\tilde{v},\overline{v}+\lambda\left(\sum_{i\in\mathbb{I}}\frac{\alpha_i}{\lambda}(\tilde{v}+\eta+c_i-\overline{v})+y_l\right)\right)=\sigma^2\sum_{i\in\mathbb{I}}\alpha_i$$

since $\tilde{v}$ is independent of η,y_l,c_i for all $i\in\mathbb{I}$. □

Proposition 9. *For any set of informed funds $\mathbb{I}$, $0<\sum_{i\in\mathbb{I}}\alpha_i<1$.*

Proof. For any $i\in\mathbb{I}$

$$\alpha_i=\frac{1}{(1+c_\eta+2c_i)\left(1+\sum_{j\in\mathbb{I}}\frac{1}{1+\frac{2c_j}{1+c_\eta}}\right)}>0$$

Hence, $\sum_{i\in\mathbb{I}}\alpha_i>0$.

By A.2 $1=2(1+c_\eta+c_i)\alpha_i+(1+c_\eta)\sum_{j\in\mathbb{I},j\neq i}\alpha_j>(1+c_\eta)\sum_{i\in\mathbb{I}}\alpha_i$. Hence, $\sum_{i\in\mathbb{I}}\alpha_i<\frac{1}{1+c_\eta}\leq 1$. □

Proposition 8 shows that price informativeness depends only on the analyst errors of the informed funds and the data error (since they determine α) as well as the variance of the asset's payoff itself. Since σ^2 is constant in the model, I will simply use the notation $PI(\mathbb{I}, c_\eta)$.

How does price informativeness change when we change $\mathbb{I}$ or c_η?

Proposition 10. *Given a set of informed funds $\mathbb{I}$, PI is decreasing in c_η, and $\lim_{c_\eta \to \infty} PI = 0$.*

Proof. Fix $i \in \mathbb{I}$ such that $c_j \leq c_i$ for all $i \in \mathbb{I}$. By 3.3 and 3.1,

$$PI = \sigma^2 \sum_{i \in \mathbb{I}} \alpha_i = \frac{\sigma^2 \sum_{j \in \mathbb{I}} \frac{1+c_\eta+2c_i}{1+c_\eta+2c_j}}{(1+c_\eta+2c_i)\left(1+\sum_{j\in\mathbb{I}} \frac{1}{1+\frac{2c_j}{1+c_\eta}}\right)}.$$

$$\frac{\partial}{\partial c_\eta} \sum_{j \in \mathbb{I}} \frac{1+c_\eta+2c_i}{1+c_\eta+2c_j} = \sum_{j \in \mathbb{I}} \frac{2(c_j - c_i)}{(1+c_\eta+2c_j)^2} \leq 0$$

which means that the numerator is weakly decreasing in c_η. The denominator is strictly increasing in c_η, hence, PI is decreasing in c_η.

$$\lim_{c_\eta \to \infty} \sum_{j \in \mathbb{I}} \frac{1+c_\eta+2c_i}{1+c_\eta+2c_j} = \lim_{c_\eta \to \infty} \sum_{j \in \mathbb{I}} \frac{\frac{1+2c_i}{c_\eta}+1}{\frac{1+2c_j}{c_\eta}+1} = |\mathbb{I}|$$

$$\lim_{c_\eta \to \infty} (1+c_\eta+2c_i)\left(1+\sum_{j\in\mathbb{I}} \frac{1}{1+\frac{2c_j}{1+c_\eta}}\right) = \infty$$

Hence, $\lim_{c_\eta \to \infty} PI = 0$. □

Proposition 11. *For any given variance of added noise c_η, the following statements hold:*

(i) for any $\mathbb{X}, \mathbb{Y} \subseteq \mathbb{F}$ if $\mathbb{X} \subset \mathbb{Y}$, then $PI(\mathbb{X}, c_\eta) < PI(\mathbb{Y}, c_\eta)$;

(ii) for any $\mathbb{X} \subset \mathbb{F}$ and any $y, z \in \mathbb{F} \setminus \mathbb{X}$, $c_y < c_z \iff PI(\mathbb{X} \cup \{y\}, c_\eta) > PI(\mathbb{X} \cup \{z\}, c_\eta)$.

Proof. (i) Given c_η, fix $\mathbb{X} \subset \mathbb{F}$ and $i \in \mathbb{F} \setminus \mathbb{X}$. Let $\tilde{\mathbb{X}} = \mathbb{X} \cup \{i\}$. Appendix A.2 shows that

$$\sum_{j \in \tilde{\mathbb{X}}} \alpha_j(\tilde{\mathbb{X}}) = \sum_{j \in \mathbb{X}} \alpha_j(\mathbb{X}) + \frac{\left(\left(\sum_{j \in \mathbb{X}} \alpha_j(\mathbb{X})\right)(1+c_\eta) - 1\right)^2}{2(1+c_\eta+c_i) - \left(\sum_{j\in\mathbb{X}} \alpha_j(\mathbb{X})\right)(1+c_\eta)^2}. \tag{3.5}$$

Here Iuse the notation $\alpha_j(\tilde{\mathbb{X}})$ and $\alpha_j(\mathbb{X})$ to emphasize that for a fund j α_j is different when the set of informed funds is $\tilde{\mathbb{X}}$ from when the set of informed funds is $\mathbb{X}$.

From the proof of proposition 9 we know that $\sum_{j\in\mathbb{X}} \alpha_j < \frac{1}{1+c_\eta}$. This implies that $\left(\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)-1\right)^2 > 0$ and $2(1+c_\eta+c_i)-\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)^2 > 1+c_\eta+2c_i > 0$. Therefore, $\sum_{j\in\tilde{\mathbb{X}}} \alpha_j > \sum_{j\in\mathbb{X}} \alpha_j$ and so $PI(\tilde{\mathbb{X}}, c_\eta) > PI(\mathbb{X}, c_\eta)$.

Now, Ican build any $\mathbb{Y} \supset \mathbb{X}$ from $\mathbb{X}$ by adding the elements from $\mathbb{Y} \setminus \mathbb{X}$ one by one. Since price informativeness increases at each step, $PI(\mathbb{Y}, c_\eta) > PI(\mathbb{X}, c_\eta)$.

(ii)

$$c_y < c_z \iff$$

$$\sum_{j\in\mathbb{X}} \alpha_j + \frac{\left(\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)-1\right)^2}{2(1+c_\eta+c_y)-\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)^2} > \sum_{j\in\mathbb{X}} \alpha_j + \frac{\left(\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)-1\right)^2}{2(1+c_\eta+c_z)-\left(\sum_{j\in\mathbb{X}} \alpha_j\right)(1+c_\eta)^2} \iff$$

$$\sigma^2 \sum_{j\in\mathbb{X}\cup\{y\}} \alpha_j > \sigma^2 \sum_{j\in\mathbb{X}\cup\{z\}} \alpha_j \iff PI(\mathbb{X}\cup\{y\}, c_\eta) > PI(\mathbb{X}\cup\{z\}, c_\eta).$$

□

Roughly speaking, Propositions 10 and 11 together imply that the more funds get the data, the more competent they are, and the higher the quality of the data, the more informative the prices.

3.2.2 Liquidity traders' welfare

Proposition 12. $\pi_l = -\lambda\sigma_l^2$.

Proof.

$$\pi_l = \mathbb{E}\left(y_l(\tilde{v}-P)\right) = \mathbb{E}(y_l(\tilde{v}-\bar{v}-\lambda(\sum_{i\in\mathbb{I}} y_i + y_l))) = -\lambda\sigma_l^2$$

since y_l is independent of all other variables in the model. □

4 No regulation

4.1 Equilibrium analysis

In the absence of regulation, the Data Seller is able to commit to selling to an exclusive set of funds. In reality, such a commitment is enforced by a contractual obligation (i.e. the Data Seller and a buyer sign a contract that prevents the Data Seller to sell to anyone besides the restricted set of funds [5], and if he does, the buyer has the right to sue him in court). The model below is stylized, it does not include the possibility of legal action explicitly, instead the Data Seller commits to not sell to other funds by making the offers simultaneously and publicly:

Step 1: Data Seller chooses c_η which becomes commonly known;

Step 2: Data Seller simultaneously and publicly makes Take-It-or-Leave-It offers to some set $\mathbb{I}$ of the funds;

Step 3: Each fund that received an offer decides whether to accept it or not. Those funds who accepted the offer receive the dataset at their respective prices.

Let $(\mathbb{I}^*, c_\eta^*)$ be a solution to the problem

$$\max_{(\mathbb{I}, c_\eta)} \sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta) \text{ such that } \mathbb{I} \subseteq \mathbb{F},\ \ c_\eta \geq 0.$$

Proposition 13. *The following strategy profile constitutes a Nash equilibrium:*

Data Seller: at step 1, chooses $c_\eta = c_\eta^*$. *At step 2, makes an offer to all* $i \in \mathbb{I}^*(c_\eta)$ *at price* $p_i = \pi_i(\mathbb{I}^*(c_\eta), c_\eta)$, *where* $\mathbb{I}^*(c_\eta)$ *is a solution to the problem*

$$\max_{\mathbb{I}} \sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta) \textit{ such that } \mathbb{I} \subseteq \mathbb{F};$$

Fund i: if i receives an offer, i accepts if $p_i \leq \pi_i(\mathbb{K}, c_\eta)$, *where* $\mathbb{K}$ *is the set of all funds who received an offer, and rejects otherwise.*

[5] Theorem 2 implies that it is not necessary to specify funds' identities in the contract, it is sufficient to specify the number of the buyers. All the players will then be able to make the correct inference about the equilibrium set of buyers.

Proof. Given the strategies of the funds, for any set $\mathbb{K}$ and any c_η it is optimal for the Data Seller to set $p_i = \pi_i(\mathbb{K}, c_\eta)$ for all $i \in \mathbb{K}$. Then at step 2 the Data Seller solves the problem

$$\max_{\mathbb{I}} \sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta) \text{ such that } \mathbb{I} \subseteq \mathbb{F},$$

and $\mathbb{I}^*(c_\eta)$ is optimal. And at step 1 the Data Seller solves the problem

$$\max_{c_\eta} \sum_{i \in \mathbb{I}^*(c_\eta)} \pi_i(\mathbb{I}^*(c_\eta), c_\eta) \text{ such that } c_\eta \geq 0,$$

and c_η^* is optimal. Hence, there is no profitable deviation for the Data Seller.

Now consider a fund $i \in \mathbb{I}^*$. By playing the above strategy, it gets zero payoff. Any other strategy would lead to either accepting the offer or rejecting the offer, both resulting in zero payoff. Hence, there is no profitable deviation for the funds either. □

On the equilibrium path, the Data Seller chooses a data error c_η and a set $\mathbb{I}$ of funds to make offers to so as to maximize the sum of their expected profits from playing in the financial market. He then extracts these expected profits by making Take-It-or-Leave-It offers that the funds accept.

What can we say about optimal c_η and $\mathbb{I}$?

Proposition 14. [6] *For any set $\mathbb{I} \subseteq \mathbb{F}$ $\sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta)$ is maximized when $c_\eta = 0$.*

Proof. Note that

$$\sum_{i \in \mathbb{I}} \pi_i = \sum_{i \in \mathbb{I}} \frac{\sigma^2 \alpha_i^2 (1 + c_\eta + c_i)}{\lambda} = \frac{\sigma^2 \sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)}{\frac{\sigma}{\sigma_l} \sqrt{\sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)}} = \sigma \sigma_l \sqrt{\sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)} = \sigma_l^2 \lambda. \tag{4.1}$$

Therefore, maximizing $\sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta)$ is equivalent to maximizing $\sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)$.

[6] Admati and Pfleiderer (1988) show that it is optimal to not add any noise in the special case when there are no interpretation errors. Garcia and Sangiorgi (2011) show that it is optimal to not add any noise in the setting without interpretation errors but when the data seller can add personalized noise (this setting neither nests nor is nested by my model).

Fix any $k \in \mathbb{I}$. By 3.3,

$$\sum_{i\in\mathbb{I}} \alpha_i^2(1+c_\eta+c_i) = \alpha_k^2(1+c_\eta+c_k) + \alpha_k^2 \sum_{j\in\mathbb{I},j\neq k} \frac{(1+c_\eta+2c_k)^2(1+c_\eta+c_j)}{(1+c_\eta+2c_j)^2} =$$

$$\alpha_k^2 \sum_{j\in\mathbb{I}} \frac{(1+c_\eta+2c_k)^2(1+c_\eta+c_j)}{(1+c_\eta+2c_j)^2} = \alpha_k^2(1+c_\eta+2c_k)^2 \sum_{j\in\mathbb{I}} \frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2}$$

Substituting the expression for α_k from 3.1,

$$\alpha_k^2(1+c_\eta+2c_k)^2 \sum_{j\in\mathbb{I}} \frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2} = \frac{(1+c_\eta+2c_k)^2 \sum_{j\in\mathbb{I}} \frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2}}{(1+c_\eta+2c_k)^2 \left(1+\sum_{j\in\mathbb{I}} \frac{1}{1+\frac{2c_j}{1+c_\eta}}\right)^2} = \frac{\sum_{j\in\mathbb{I}} \frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2}}{\left(1+\sum_{j\in\mathbb{I}} \frac{1}{1+\frac{2c_j}{1+c_\eta}}\right)^2}. \tag{4.2}$$

For any $j \in \mathbb{I}$ $\frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2}$ is decreasing in c_η since

$$\frac{\partial}{\partial c_\eta} \frac{1+c_\eta+c_j}{(1+c_\eta+2c_j)^2} = \frac{-(1+c_\eta)}{(1+c_\eta+2c_j)^3} < 0,$$

hence, the numerator of 4.2 is decreasing in c_η. And the denominator of 4.2 is increasing in c_η (since for any $j \in \mathbb{I}$ $\frac{2c_j}{1+c_\eta}$ is decreasing in c_η and, thus, $\frac{1}{1+\frac{2c_j}{1+c_\eta}}$ is increasing in c_η). Therefore, 4.2 is decresing in c_η, and $\sum_{i\in\mathbb{I}} \pi_i(\mathbb{I}, c_\eta)$ is maximized when $c_\eta = 0$. □

Theorem 1. *In the absence of regulation Data Seller always chooses $c_\eta = 0$.*

Proof. The result follows immediately from Proposition 14 since in the absence of regulation, in equilibrium, when the Data Seller proposes to a set $\mathbb{I}$ of funds, his payoff is equal to $\sum_{i\in\mathbb{I}} \pi_i(\mathbb{I}, c_\eta)$. □

To simplify further notation let $\pi_{DS}^{\text{no reg}}(\mathbb{I}) = \sum_{i\in\mathbb{I}} \pi_i(\mathbb{I}, 0)$.

Proposition 15. *For any set of funds $\mathbb{I} \subset \mathbb{F}$ and any $i, j \in \mathbb{F} \setminus \mathbb{I}$ such that $c_i < c_j$*

if $\pi_{DS}^{no\ reg}(\mathbb{I} \cup \{j\}) \geq \pi_{DS}^{no\ reg}(\mathbb{I})$, then $\pi_{DS}^{no\ reg}(\mathbb{I} \cup \{i\}) > \pi_{DS}^{no\ reg}(\mathbb{I} \cup \{j\})$.

Proof. See Appendix A.3. □

Theorem 2. *In the absence of regulation the Data Seller makes offers to a subset of the most competent funds.*

Proof. Suppose, towards a contradiction, that a set of funds $\mathbb{I} \subset \mathbb{F}$ is optimal and there exist $j \in \mathbb{I}$ and $i \in \mathbb{F} \setminus \mathbb{I}$ such that $c_i < c_j$. Since $\mathbb{I}$ is optimal, $\pi_{DS}^{\text{no reg}}(\mathbb{I}) \geq \pi_{DS}^{\text{no reg}}(\mathbb{I} \setminus \{j\})$. But then by Proposition 15 $\pi_{DS}^{\text{no reg}}(\mathbb{I} \setminus \{j\} \cup \{i\}) > \pi_{DS}^{\text{no reg}}(\mathbb{I} \setminus \{j\} \cup \{j\}) = \pi_{DS}^{\text{no reg}}(\mathbb{I})$ which means that $\mathbb{I}$ is not optimal – contradiction. □

How large is the optimal subset of the funds? In general, the answer depends on the realization of the profile of analyst errors. It is, however, easy to characterize in the special case when all the funds are equally competent.

Example 4.1. *Let $c_i = c$ for all $i \in \{1, ..., N\}$. Then in the absence of regulation it is optimal for the data seller to make offers to $1 + 2c$[7] (or one of the two closest integers if c is not an integer) funds.*

This result is straightforward: let the data seller make n offers, then using 3.1 and substituting $c_\eta = 0$ we get $\alpha_i = \frac{1}{n+1+2c}$. Then using 3.2 $\lambda = \frac{\sigma}{\sigma_l} \cdot \frac{\sqrt{n(1+c)}}{n+1+2c}$. By 3.4 the data seller's profit is maximized when lambda is maximized which happens at $n = 1 + 2c$.

It is optimal to sell to more buyers when funds are less precise since it is harder for the market makers to make a correct inference about $\tilde{v}$ after observing the noisy aggregate demand and so adding an additional buyer creates less competition effect on the existing buyers.

4.2 Implications for the regulator

Proposition 16. *In the absence of regulation the equilibrium outcome is the worst possible outcome from the liquidity traders' perspective.*

Proof.

$$\pi_{DS}^{\text{no reg}}(\mathbb{I}, c_\eta) = \sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta) = \lambda(\mathbb{I}, c_\eta) \cdot \sigma_l^2 = -\pi_l,$$

where the second equality is by 3.4 and the third equality is by Proposition 12. Hence, by maximizing his own profit the data seller automatically minimizes the welfare of the liquidity traders. □

[7] Admati and Pfleiderer (1988) show a special case of this result: when there are no interpretation errors (i.e. $c = 0$), it is optimal to have one informed trader.

Another way to get the above result is to notice that the financial market is a closed system that consists of the hedge funds, the market makers, and the liquidity traders. The uninformed hedge funds do not trade so their profit is zero, and the market makers do not receive any profit in equilibrium. Therefore, in expectation any profit of the informed hedge funds is necessarily a loss for the liquidity traders.

The implication of Proposition 16 is that any regulator's policy would weakly improve the welfare of the liquidity traders. Combining this with the results from Propositions 10 and 11 we can conclude that if a policy incentivized the data seller to sell to a superset of the current set of buyers (when there is no regulation) while keeping the quality of the data high, price informativeness would increase (strictly), and the welfare of the liquidity traders would increase (weakly) as well.

When studying different policies in the following sections I will focus on how they affect price informativeness.

5 Fixed price policy

The first policy I consider is the European regulation which I interpret as requiring the data seller to set a fixed price and then sell to anyone who is willing to buy at that price.

The timing of the game is as follows:

At step 1, the Data Seller chooses a level of noise c_η and a price p.

At step 2, the hedge funds simultaneously decide whether to purchase the dataset.

At step 3, the set of buyers becomes publicly known, and the play in the financial market happens.

Definition 5.1. Given p and c_η, step-2-equilibrium is a profile of funds' decisions about purchasing the dataset such that if $\mathbb{X}$ is the set of funds who buy and $i \in \mathbb{X}$, then $\pi_i(\mathbb{X}, c_\eta) \geq p$ and if $i \notin \mathbb{X}$, then $\pi_i(\mathbb{X} \cup \{i\}, c_\eta) < p$. Let $B(p, c_\eta)$ denote the set of funds-buyers in equilibrium.

Definition 5.2. Most-competent-step-2-equilibrium is a step-2-equilibrium in which $B(p, c_\eta)$ is a subset of the most competent funds.

Proposition 17. *Any combination of p and c_η induces at least one most-competent-step-2-equilibrium.*

Proof. Let's enumerate the funds from 1 to N so that $c_1 \leq ... \leq c_N$, and for all $i \in \{1, ..., N\}$ let's compare $\pi_i(\{1, ..., i\}, c_\eta)$ and p. Three cases are possible.

Case 1: $\pi_i(\{1, ..., i\}, c_\eta) < p$ for all $i \in \{1, ..., N\}$. Then there exists a most-competent-step-2-equilibrium in which no one buys the dataset. Indeed, if some fund $i \in \{1, ..., N\}$ was to deviate and buy, then $\pi_i(\{i\}, c_\eta) \leq \pi_1(\{1\}, c_\eta) < p$, where the first inequality is by Proposition 5 and the second inequality is by assumption.

Case 2: $\pi_i(\{1, ..., i\}, c_\eta) \geq p$ for all $i \in \{1, ..., N\}$.Then there exists a most-competent-step-2-equilibrium in which all the funds buy the dataset. Indeed, for any $i \in \{1, ..., N\}$ $\pi_i(\{1, ..., N\}, c_\eta) \geq \pi_N(\{1, ..., N\}, c_\eta) \geq p$, where the first inequality is by Proposition 4 and the second inequality is by assumption.

Case 3: For some but not all $i \in \{1, ..., N\}$ $\pi_i(\{1, ..., i\}, c_\eta) < p$. Let j be the smallest of such i. Note that $j > 1$ since otherwise for all $k > j$ $\pi_k(\{1, ..., k\}, c_\eta) \leq \pi_j(\{1, ..., k\}, c_\eta) < \pi_j(\{1, ..., j\}, c_\eta) < p$, where the first inequality is by Proposition 4, the second inequality is by Corollary 7.1, and the third inequality is by assumption, but then we would be in a case 1 type scenario instead of type 3.

There exists a most-competent-step-2-equilibrium in which funds $\{1, ..., j-1\}$ buy the dataset and funds $\{j, ..., N\}$ do not. Indeed, for all $k < j-1$ $\pi_k(\{1, ..., j-1\}, c_\eta) \geq \pi_{j-1}(\{1, ..., j-1\}, c_\eta) \geq p$, where the first inequality is by Proposition 4 and the second inequality is by assumption, and for all $k > j$ $\pi_k(\{1, ..., j-1\} \cup \{k\}, c_\eta) \leq \pi_j(\{1, ..., j-1\} \cup \{j\}, c_\eta) < p$, where the first inequality is by Proposition 5 and the second inequality is by assumption. □

To simplify further analysis I will focus on most-competent-step-2-equilibria.

At step 1 the data seller chooses p and c_η so as to maximize his expected profit $\mathbb{E}\left(\pi_{DS}^{FP}\right) = p \cdot \mathbb{E}\left(|B(p, c_\eta)|\right)$. The expectation is needed since it is possible that a combination of p and c_η induces multiple most-competent-step-2-equilibria.

The following two examples show that the optimal set of buyers can both increase and decrease under the fixed price policy compared to the benchmark of no regulation.

Example 5.1. *Consider a data market with two funds: $c_1 = 1, c_2 = 15$, and the cost of providing the dataset is zero.*

In the absence of regulation by Theorem 1 it is optimal to choose $c_\eta = 0$, and by Theorem 2 it is optimal to sell either to fund 1 alone or to both funds.

$\pi_1(\{1\}, 0) \approx 0.3536\sigma\sigma_l$

$\pi_1(\{1,2\}, 0) + \pi_2(\{1,2\}, 0) \approx 0.3330\sigma\sigma_l + 0.0249\sigma\sigma_l = 0.3579\sigma\sigma_l$

Hence, the data seller will choose to sell to both funds.

Under the fixed price policy there are three possible most-competent-step-2-equilibria – when no one buys, when fund 1 buys, and when both funds buy. Of course, an equilibrium in which no one buys cannot be optimal. Suppose that optimal p and c_η induce an equilibrium in which only fund 1 buys. Then $p = \pi_1(\{1\}, c_\eta)$ since a lower price would result in a lower profit, and a higher price cannot result in fund 1 buying in equilibrium. By Proposition 14 $\pi_1(\{1\}, c_\eta)$ is maximized when $c_\eta = 0$, in which case $\pi_{DS}^{FP} = 0.3536\sigma\sigma_l$.

If optimal p and c_η induce an equilibrium in which both funds buy, then $p = \pi_2(\{1,2\}, c_\eta)$ since a lower price would result in a lower profit, and a higher price cannot result in fund 2 buying in equilibrium. $\pi_2(\{1,2\}, c_\eta)$ is maximized when $c_\eta \approx 17.7$ [8], in which case $\pi_{DS}^{FP} \approx 2 \cdot 0.025\sigma\sigma_l = 0.05\sigma\sigma_l$.

Hence, under the fixed price policy the data seller will sell only to fund 1.

Example 5.2. *Consider a data market with three funds: $c_1 = 0.7, c_2 = c_3 = 0.77$, and the cost of providing the dataset is zero.*

Following the logic of the previous example, in the absence of regulation the data seller chooses between the following three options:

$\pi_1(\{1\}, 0) \approx 0.3835\sigma\sigma_l$

$\pi_1(\{1,2\}, 0) + \pi_2(\{1,2\}, 0) \approx 0.21603\sigma\sigma_l + 0.20082\sigma\sigma_l = 0.41685\sigma\sigma_l$

$\pi_1(\{1,2,3\}, 0) + \pi_2(\{1,2,3\}, 0) + \pi_3(\{1,2,3\}, 0) \approx 0.14577\sigma\sigma_l + 0.1355\sigma\sigma_l + 0.1355\sigma\sigma_l = 0.41677\sigma\sigma_l$

Hence, the data seller will choose to sell to funds 1 and 2.

Under the fixed price policy the data seller chooses between $\pi_1(\{1\}, c_\eta)$, $2 \cdot \pi_2(\{1,2\}, c_\eta)$, and $3 \cdot \pi_3(\{1,2,3\}, c_\eta)$. Again, $\pi_1(\{1\}, c_\eta)$ is maximized when $c_\eta = 0$ in which case $\pi_1(\{1\}, 0) \approx 0.3835\sigma\sigma_l$.

[8] optimized numerically

$2 \cdot \pi_2(\{1,2\}, c_\eta)$ is maximized when $c_\eta = 0$[9] *in which case $2 \cdot \pi_2(\{1,2\}, 0) \approx 2 \cdot 0.20082\sigma\sigma_l = 0.40164\sigma\sigma_l$.*

$3 \cdot \pi_3(\{1,2,3\}, c_\eta)$ is maximized when $c_\eta = 0$[10] *in which case $3 \cdot \pi_3(\{1,2,3\}, 0) \approx 3 \cdot 0.1355\sigma\sigma_l = 0.4065\sigma\sigma_l$.*

Hence, under fixed price policy the data seller will sell to all three funds.

Proposition 18. *If in the absence of regulation it is optimal for the data seller to sell to only one fund, then under the fixed price policy it is still optimal and feasible to sell to only one fund.*

Proof. Since in the absence of regulation it was optimal for the data seller to sell to only one fund, by Theorem 2 it was fund 1 and by Theorem 1 it was optimal to choose $c_\eta = 0$, i.e. the data seller's profit was $\pi_1(\{1\}, 0)$.

Let $B(p, c_\eta)$ be the equilibrium set of buyers at step 2 and let $|B(p, c_\eta)| = n$. Note that $p \leq \pi_n(B(p, c_\eta), c_\eta)$ since fund n buys in equilibrium. Then $\pi_{DS}^{FP} = n \cdot p \leq n \cdot \pi_n(B(p, c_\eta), c_\eta) \leq \sum_{i \in B(p, c_\eta)} \pi_i(B(p, c_\eta), c_\eta) \leq \pi_1(\{1\}, 0)$, where the second inequality is by Proposition 4, and the last inequality is by assumption. Therefore, $\pi_1(\{c_1\}, 0)$ is an upper bound on the data seller's profit under the fixed price policy.

It is feasible to achieve this upper bound by choosing $c_\eta = 0$ and setting $p = \pi_1(\{1\}, 0)$. Then fund 1 buys in equilibrium while the other funds do not since for any $k > 1$ $\pi_k(\{1, k\}, 0) < \pi_k(\{k\}, 0) \leq \pi_1(\{1\}, 0) = p$, where the first inequality is by Proposition 7 and the second inequality is by Proposition 5. □

The above theorem shows that a fixed price policy is not effective when a hedge fund gets a unique access to the information. A natural next step for the regulator is then to consider directly requiring the data seller to sell to at least some number of buyers.

6 Lower bound on the quantity

Suppose the regulator sets a lower bound on the number of datasets that need to be sold $\underline{n}$ but after that the data seller decides who to sell to and at what prices (by making simultaneous public

[9] optimized numerically

[10] optimized numerically

Take-It-or-Leave-It offers as when there is no regulation). Our hope is that it would still be optimal to sell to the most precise funds and choose the highest quality of the data.

Definition 6.1. A policy is analyst efficient if for any realization of analyst errors $\{c_1, ..., c_N\}$ any equilibrium set of buyers is a subset of the most precise funds.

Definition 6.2. A policy is data efficient if for any realization of analyst errors $\{c_1, ..., c_N\}$ if the data seller provides the dataset, he chooses the highest quality of the data.

Theorem 3. *Lower bound on the number of buyers is a data efficient policy.*

Proof. The equilibrium strategies are analogous to Proposition 13 with the difference that $\mathbb{I}^*, c_\eta^*$ is now a solution to

$$\max_{(\mathbb{I}, c_\eta)} \sum_{i \in \mathbb{I}} \pi_i(\mathbb{I}, c_\eta) \text{ such that } \mathbb{I} \subseteq \mathbb{F},\ |\mathbb{I}| \geq \underline{n},\ c_\eta \geq 0.$$

On the equilibrium path the data seller still extracts the profits of the informed funds and, therefore, by Proposition 14 it is optimal to choose $c_\eta = 0$. □

Theorem 4. *Lower bound on the number of buyers is not an analyst efficient policy.*

Proof. Consider a data market with four funds: $c_1 = c_2 = 0.05$, $c_3 = c_4 = 1$, and the cost of providing the dataset is zero. The table below provides the optimal set of buyers, the data seller's profit and the resulting price informativeness for different lower bounds.

$\underline{n}$	optimal set of informed funds	π_{DS}	PI
0 or 1	{1}	$0.49\sigma\sigma_l$	$0.48\sigma^2$
2	{1,2}	$0.47\sigma\sigma_l$	$0.65\sigma^2$
3	{1,3,4}	$0.44\sigma\sigma_l$	$0.61\sigma^2$
4	{1,2,3,4}	$0.42\sigma\sigma_l$	$0.71\sigma^2$

When $\underline{n} = 3$ the optimal set of buyers is $\{1, 3, 4\}$ which is not a subset of the most precise funds. □

An interesting consequence of Theorem 4 is that a stricter policy does not necessarily imply higher price informativeness. When $\underline{n} = 2$ the data seller sells to funds 1 and 2. When the policy is stricter and $\underline{n} = 3$, even though the data seller sells to more funds (1,3, and 4), two of them are less precise and, as a result, the price informativeness is lower than it was with $\underline{n} = 2$.

Intuitively, by adding the incompetent funds the data seller creates less competition to the funds he actually wants to sell to and, hence, can charge them more than if he sold to more competent new funds. In case the new funds are "extremely incompetent" the data seller's profit under the regulation will approach his profit in the absence of regulation:

Proposition 19. *Let $\mathbb{I}^*$ be the set of informed funds in the absence of regulation, $|\mathbb{I}^*| = k$. Let $\underline{n} > k$ be the lower bound on the quantity, and suppose that there exists a set $\mathbb{X}$ of size $\underline{n} - k$ of "extremely incompetent" funds (i.e. $c_i \to \infty$ for all $i \in \mathbb{X}$) that are not in the current set of buyers $\mathbb{I}^*$.*

Then

$$\pi_{DS}(\mathbb{I}^* \cup \mathbb{X}) \to \pi_{DS}(\mathbb{I}^*),$$

$$PI(\mathbb{I}^* \cup \mathbb{X}) \to PI(\mathbb{I}^*),$$

$$\pi_l(\mathbb{I}^* \cup \mathbb{X}) \to \pi_l(\mathbb{I}^*).$$

Proof. By 3.1 for all $i \in \mathbb{X}$ $\alpha_i(\mathbb{I}^* \cup \mathbb{X}) \to 0$, and for all $i \in \mathbb{I}^*$ $\alpha_i(\mathbb{I}^* \cup \mathbb{X}) \to \alpha_i(\mathbb{I}^*)$. Hence, by 3.2 $\lambda(\mathbb{I}^* \cup \mathbb{X}) \to \lambda(\mathbb{I}^*)$. This implies that $\pi_{DS}(\mathbb{I}^* \cup \mathbb{X}) \to \pi_{DS}(\mathbb{I}^*)$ since for all $i \in \mathbb{I}^*$

$$\pi_i(\mathbb{I}^* \cup \mathbb{X}) = \frac{\sigma^2 \alpha_i^2(\mathbb{I}^* \cup \mathbb{X})}{\lambda(\mathbb{I}^* \cup \mathbb{X})}(1 + c_\eta + c_i) \to \frac{\sigma^2 \alpha_i^2(\mathbb{I}^*)}{\lambda(\mathbb{I}^*)}(1 + c_\eta + c_i) = \pi_i(\mathbb{I}^*)$$

and for all $i \in \mathbb{X}$

$$\pi_i(\mathbb{I}^* \cup \mathbb{X}) = \frac{\sigma^2 \alpha_i^2(\mathbb{I}^* \cup \mathbb{X})}{\lambda(\mathbb{I}^* \cup \mathbb{X})}(1 + c_\eta + c_i) \to 0;$$

$$PI(\mathbb{I}^* \cup \mathbb{X}) = \sigma^2 \sum_{i \in \mathbb{I}^* \cup \mathbb{X}} \alpha_i \to \sigma^2 \sum_{i \in \mathbb{I}^*} \alpha_i = PI(\mathbb{I}^*);$$

$$\pi_l(\mathbb{I}^* \cup \mathbb{X}) = -\sigma^2 \lambda(\mathbb{I}^* \cup \mathbb{X}) \to -\sigma^2 \lambda(\mathbb{I}^*) = \pi_l(\mathbb{I}^*).$$

□

By adding the "extremely incompetent" funds the data seller satisfies the regulation, however, the market outcomes that are relevant to the regulator (PI and π_l) do not change in any meaningful way, and the policy end up being useless. In order to make sure that the policy is analyst efficient, the regulator needs to not only set a lower bound on the number of buyers but also impose a mechanism that should be used when selling the data. An auction can serve as such a mechanism.

7 Auction with a lower bound on the quantity

Suppose the regulator sets a lower bound on the number of datasets that need to be sold $\underline{n}$, and after that the data seller decides whether to provide the dataset. If he does, he chooses a data error c_η which becomes publicly known, and then he sells some number K of the datasets through a $K+1$ price auction [11] (with ties broken randomly) such that $K \geq \underline{n}$.

Proposition 20. *Let's enumerate the funds from 1 to N so that $c_1 \leq ... \leq c_K$. Given c_η and K, an equilibrium bidding strategy is*

if $i \leq K$, $b_i = \pi_i(\{1, ..., K\}, c_\eta)$,

if $i > K$, $b_i = \pi_i(\{1, ..., K-1, i\}, c_\eta)$.

Proof. First, note that if $c_i < c_j$, then $b_i > b_j$: if $i \leq K$ and $j \leq K$, then $b_i > b_j$ by Proposition 4, if $i > K$ and $j > K$, then $b_i > b_j$ by Proposition 5, and if $i \leq K$ and $j > K$, then $b_i \geq b_K \geq b_j$, where the first inequality is by Proposition 4 and the second is by Proposition 5. Since $c_i < c_j$ and $i \leq K < j$, at least one of $c_i < c_K$ and $c_K < c_j$ holds and, hence, at least one of the inequalities is strict. Therefore, K most competent funds receive the dataset if the proposed strategy is played.

Let's now verify that the proposed bids constitute an equilibrium. Case 1: $c_K < c_{K+1}$. If $c_i < c_{K+1}$, then i wins and gets a positive profit $\pi_i(\{1, ..., K\}, c_\eta) - b_{K+1} = b_i - b_{K+1} > 0$. The only consequential deviation would be to bid less than b_{K+1} and lose which is not profitable.

If $c_i \geq c_{K+1}$, then i loses. The only consequential deviations would be to bid b_K (and win with positive probability) or to bid more than b_K (and win for sure), both at price b_K. But $\pi_i(\{1, ..., K-1, i\}, c_\eta) - b_K = b_i - b_K < 0$ so this deviation is not profitable.

[11] Theorems 5 and 6 still hold if one considers pay-as-bid auction instead.

Case 2: $c_K = c_{K+1}$. The argument for $c_i < c_{K+1}$ and $c_i > c_{K+1}$ still holds. Consider type c_{K+1}. It gets zero payoff (either by losing or by winning at the price of its profit). If it deviates to bid less than b_K, it'll lose for sure and still get zero. If it deviates to bid more than b_K, it'll win for sure at the price of b_K, i.e. its profit, and therefore, still get zero. Hence, there is no profitable deviation. □

Theorem 5. *Auction with a lower bound on the quantity is an analyst efficient policy.*

Proof. See the first paragraph of the proof of Proposition 20. □

Theorem 6. *Auction with a lower bound on the quantity is not a data efficient policy.*

Proof. Consider a data market with three funds: $c_1 = c_2 = 0.05, c_3 = 2$, and the cost of providing the dataset is zero. Let $\underline{n} = 2$. The data seller will choose to sell $K = 2$ datasets since $K = 3$ would result in zero profit. In equilibrium funds 1 and 2 will get the dataset but the data seller's profit is determined by the bid of fund 3 $b_3 = \pi_3(\{1, 3\}, c_\eta)$ which is maximized when $c_\eta \approx 2.02$. Hence, it is optimal for the data seller to not choose the highest quality of the data, and the policy is not data efficient.

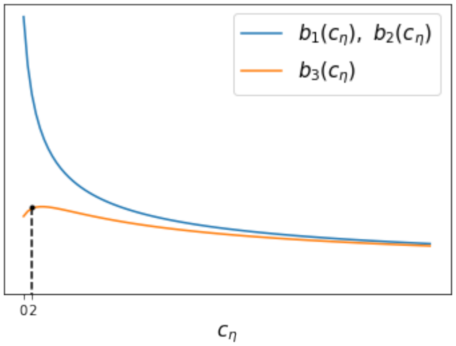

Figure 1: By selling a nosier version of the dataset the data seller makes it less valuable for the actual buyers but induces a higher bid that determines his profit.

□

The fact that fund 3 values a nosier version of the dataset more than the most informative one can seem counterintuitive at first. The reason, however, is that even though a nosier version of the data is less useful for fund 3, it is also less useful for fund 1. Since fund 3 is worse at interpreting the data than fund 1, adding some noise reduces fund 3's comparative disadvantage and increases its expected profit.

The takeaway from this section is that using an auction is a natural way of making sure that the funds that value data the most (i.e. the most precise ones) get it in equilibrium, however, not every auction format creates incentives for the data seller to provide the data of the highest quality.

8 Auction with contingent payments and a lower bound on the quantity

Suppose the regulator imposes the following auction format on the data seller:

At step 1 the regulator sets the lower bound on the quantity $\underline{n}$ and an entry fee f.

At step 2 the data seller chooses c_η which becomes publicly known, and a number $K \geq \underline{n}$ of datasets to sell.

At step 3 the hedge funds submit their bids as a share of their future profit.

At step 4 the funds with top K bids win (with ties broken randomly) and pay $K+1$ bid's share of their profits.

At step 5 the regulator collects the entry fees from the winners and returns the entry fees to the losers.

This auction format is based on Hansen (1985) and DeMarzo, Kremer and Skrzypacz (2005) who show that it has desirable properties from the revenue maximization perspective. It turns out that it also has desirable properties from the perspective of creating the right incentives for the data seller.

Proposition 21. *Let's enumerate the funds from 1 to N so that $c_1 \leq ... \leq c_K$. Given c_η and K, an equilibrium bidding strategy is*

if $i \leq K$, $b_i = 1 - \frac{f}{\pi_i(\{1,...,K\},c_\eta)}$,

if $i > K$, $b_i = 1 - \frac{f}{\pi_i(\{1,...,K-1,i\},c_\eta)}$.

Proof. First, note that if $c_i < c_j$, then $b_i > b_j$ (see the proof of Proposition 20). Therefore, K most competent funds receive the dataset if the proposed strategy is played.

Let's now verify that the proposed bids constitute an equilibrium. Case 1: $c_K < c_{K+1}$. If $i < K+1$, then i wins and gets a positive profit

$$(1-b_{K+1}) \cdot \pi_i(\{1,...,K\}, c_\eta) - f = \left(\frac{\pi_i(\{1,...,K\}, c_\eta)}{\pi_{K+1}(\{1,...,K-1,K+1\}, c_\eta)} - 1 \right) \cdot f > 0.$$

The only consequential deviation would be to bid less than b_{K+1} and lose which is not profitable.

If $i \geq K+1$, then i loses. The only consequential deviations would be to bid b_K (and win with positive probability) or to bid more than b_K (and win for sure), both at price b_K. But

$$(1-b_K) \cdot \pi_i(\{1,...,K-1,i\}, c_\eta) - f = \left(\frac{\pi_i(\{1,...,K-1,i\}, c_\eta)}{\pi_K(\{1,...,K\}, c_\eta)} - 1 \right) \cdot f < 0,$$

so such a deviation is not profitable.

Case 2: $c_K = c_{K+1}$. The argument for $c_i < c_{K+1}$ and $c_i > c_{K+1}$ still holds. Consider type c_{K+1}. It gets zero payoff – either by losing or by winning and getting

$$(1-b_K) \cdot \pi_{K+1}(\{1,...,K-1,K+1\}, c_\eta) - f = \left(\frac{\pi_{K+1}(\{1,...,K-1,K+1\}, c_\eta)}{\pi_K(\{1,...,K\}, c_\eta)} - 1 \right) \cdot f = 0.$$

If it deviates to bid less than b_K, it'll lose for sure and still get zero. If it deviates to bid more than b_K, it'll win for sure at the price of b_K share of its profit, and therefore, still get zero. Hence, there is no profitable deviation. □

Theorem 7. *Auction with contingent payments and a lower bound on the quantity is an analyst efficient policy.*

Proof. See the first paragraph of the proof of Proposition 21. □

Theorem 8. *Auction with contingent payments and a lower bound on the quantity is a data efficient policy.*

Proof. For any number of the datasets K that the data seller chooses to sell, the data seller's profit is $b_{K+1} \cdot \sum_{i=1}^{K} \pi_i(\{1, ..., K\}, c_\eta)$. By Proposition 14 $\sum_{i=1}^{K} \pi_i(\{1, ..., K\}, c_\eta)$ is maximized when $c_\eta = 0$. □

What is the optimal $\underline{n}$? It depends on the regulator's preferences: by setting $\underline{n}$ low the regulator would sometimes not get as much price informativeness/liquidity traders' welfare as possible, by setting $\underline{n}$ high the dataset would sometimes not be produced if the data seller's expected profit does not cover the cost.

9 Conclusion

In this paper I show that trading on data sold through data markets can have similar consequences as insider trading. The regulator might, therefore, consider intervening. When considering different regulation regimes one should be mindful about how they affect the equilibrium set of data buyers as well as data seller's incentives to provide high quality data. Requiring the data seller to sell the information through an auction with contingent payments and a lower bound on the quantity increases both liquidity traders' welfare and price informativeness in expectation. The model, however, does not account for the fact that in reality traders hold many different assets, and have access to multiple information sources. The model also does not consider scenarios where the seller is not a monopolist. Including these features into the model is a direction for future research.

A Appendix

A.1 Proof of Proposition 1:

Lemma 1. *If* $a \sim N(\mu_a, \sigma_a^2)$ *and* $b \sim N(\mu_b, \sigma_b^2)$*, then* $\mathbb{E}(a \mid b) = \mu_a + \frac{Cov(a,b)}{\sigma_b^2}(b - \mu_b)$*.*

Proof. Let $\lambda = \frac{Cov(a,b)}{\sigma_b^2}$, then $Cov(a - \lambda b, b) = Cov(a,b) - \lambda Var(b) = 0$. Since for any scalars x_1, x_2 such that x_1, x_2 are not both equal to zero,

$$x_1(a - \lambda b) + x_2 b \sim N\left(x_1\mu_a + (x_2 - x_1\lambda)\mu_b,\ x_1^2\sigma_a^2 + (x_2 - x_1\lambda)^2\sigma_b^2\right),$$

$a - \lambda b$ and b are jointly normally distributed. Since their covariance is zero, they are independent. Then $\mathbb{E}(a \mid b) = \mathbb{E}(a - \lambda b \mid b) + \lambda b = \mathbb{E}(a - \lambda b) + \lambda b = \mu_a + \frac{Cov(a,b)}{\sigma_b^2}(b - \mu_b)$. □

Step 1: If fund i is informed, then $y_i = \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v})$ maximizes i's expected profit:

$$\mathbb{E}\left(y_i\left(\tilde{v} - P\right) \mid \tilde{v} + \eta + \epsilon_i\right) = \mathbb{E}\left(y_i\left(\tilde{v} - \overline{v} - \lambda\left(y_i + \sum_{j\in\mathbb{I}, j\neq i}\frac{\alpha_j}{\lambda}(\tilde{v} + \eta + \epsilon_j - \overline{v}) + y_l\right)\right) \mid \tilde{v} + \eta + \epsilon_i\right)$$

Since for all j ϵ_j and y_l have zero mean and are independent of all other random variables we can rewrite the above expression as

$$y_i(\mathbb{E}(\tilde{v} \mid \tilde{v} + \eta + \epsilon_i) - \overline{v}) - \lambda y_i^2 - y_i(\sum_{j\in\mathbb{I}, j\neq i}\alpha_j)(\mathbb{E}(\tilde{v} + \eta \mid \tilde{v} + \eta + \epsilon_i) - \overline{v}) \qquad \text{(A.1)}$$

Note that $\tilde{v} + \eta + \epsilon_i \sim N(\overline{v},\ \sigma^2 + c_\eta\sigma^2 + c_i\sigma^2)$. Then by Lemma 1

$$\mathbb{E}(\tilde{v} \mid \tilde{v} + \eta + \epsilon_i) = \overline{v} + \frac{Cov(\tilde{v}, \tilde{v} + \eta + \epsilon_i)}{\sigma^2 + c_\eta\sigma^2 + c_i\sigma^2}(\tilde{v} + \eta + \epsilon_i - \overline{v}) = \overline{v} + \frac{\sigma^2}{\sigma^2 + c_\eta\sigma^2 + c_i\sigma^2}(\tilde{v} + \eta + \epsilon_i - \overline{v}) =$$

$$= \overline{v} + \frac{1}{1 + c_\eta + c_i}(\tilde{v} + \eta + \epsilon_i - \overline{v})$$

and

$$\mathbb{E}(\tilde{v}+\eta \mid \tilde{v}+\eta+\epsilon_i) = \overline{v}+\frac{Cov(\tilde{v}+\eta, \tilde{v}+\eta+\epsilon_i)}{\sigma^2+c_\eta\sigma^2+c_i\sigma^2}(\tilde{v}+\eta+\epsilon_i-\overline{v}) = \overline{v}+\frac{\sigma^2+c_\eta\sigma^2}{\sigma^2+c_\eta\sigma^2+c_i\sigma^2}(\tilde{v}+\eta+\epsilon_i-\overline{v}) =$$

$$= \overline{v} + \frac{1+c_\eta}{1+c_\eta+c_i}(\tilde{v}+\eta+\epsilon_i-\overline{v})$$

Substituting back to A.1 we get

$$-\lambda y_i^2 + y_i(\tilde{v}+\eta+\epsilon_i-\overline{v})\left(\frac{1}{1+c_\eta+c_i} - \frac{1+c_\eta}{1+c_\eta+c_i}\cdot\sum_{j\in\mathbb{I}, j\neq i}\alpha_j\right)$$

Since it is a concave quadratic function, the optimum is at

$$y_i = \frac{\left(\frac{1}{2(1+c_\eta+c_i)} - \frac{1+c_\eta}{2(1+c_\eta+c_i)}\cdot\sum_{j\in\mathbb{I}, j\neq i}\alpha_j\right)}{\lambda}\cdot(\tilde{v}+\eta+\epsilon_i-\overline{v})$$

It is left to check that

$$\alpha_i = \frac{1}{2(1+c_\eta+c_i)} - \frac{1+c_\eta}{2(1+c_\eta+c_i)}\cdot\sum_{j\in\mathbb{I}, j\neq i}\alpha_j \text{ for all } i\in\mathbb{I}, \text{ or}$$

$$2(1+c_\eta+c_i)\alpha_i + (1+c_\eta)\sum_{j\in\mathbb{I}, j\neq i}\alpha_j = 1 \quad \text{for all } i\in\mathbb{I} \tag{A.2}$$

By A.2, for any $i, j \in \mathbb{I}$ such that $j \neq i$

$$2(1+c_\eta+c_i)\alpha_i + (1+c_\eta)\alpha_j = 2(1+c_\eta+c_j)\alpha_j + (1+c_\eta)\alpha_i$$

and, hence,

$$\alpha_j = \frac{1+c_\eta+2c_i}{1+c_\eta+2c_j}\alpha_i. \tag{A.3}$$

By substituting into A.2, we get

$$2(1+c_\eta+c_i)\alpha_i + \alpha_i(1+c_\eta)\sum_{j\in\mathbb{I}, j\neq i}\frac{1+c_\eta+2c_i}{1+c_\eta+2c_j} = 1$$

$$\alpha_i \left(1 + c_\eta + 2c_i + (1 + c_\eta) \sum_{j \in \mathbb{I}} \frac{1 + c_\eta + 2c_i}{1 + c_\eta + 2c_j}\right) = 1$$

$$\alpha_i = \frac{1}{(1 + c_\eta + 2c_i)\left(1 + (1 + c_\eta)\sum_{j \in \mathbb{I}} \frac{1}{1 + c_\eta + 2c_j}\right)} = \frac{1}{(1 + c_\eta + 2c_i)\left(1 + \sum_{j \in \mathbb{I}} \frac{1}{1 + \frac{2c_j}{1 + c_\eta}}\right)}$$

Therefore, $y_i = \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v})$ is indeed optimal.

Step 2: If fund i is uninformed, then $y_i = 0$ maximizes i's expected profit:

$$\mathbb{E}\left(y_i\left(\tilde{v} - P\right)\right) = \mathbb{E}\left(y_i\left(\tilde{v} - \overline{v} - \lambda\left(y_i + \sum_{j \in \mathbb{I}} \frac{\alpha_j}{\lambda}(\tilde{v} + \eta + \epsilon_j - \overline{v}) + y_l\right)\right)\right)$$

Since i does not have any additional information, from i's perspective $\mathbb{E}(\tilde{v}) = \overline{v}$ and $\mathbb{E}(\tilde{v} + \eta + \epsilon_j) = \overline{v}$ for all j. Therefore, i maximizes $-\lambda y_i^2$ which means that it is optimal to demand $y_i = 0$.

Step 3: The market price is equal to the expected payoff of the asset conditional on observing the order flow:

Lemma 2. $\sum_{i \in \mathbb{I}} \alpha_i - \left(\sum_{i \in \mathbb{I}} \alpha_i\right)^2 (1 + c_\eta) - \sum_{i \in \mathbb{I}} \alpha_i^2 c_i = \sum_{i \in \mathbb{I}} \alpha_i^2 (1 + c_\eta + c_i)$.

Proof. Let $|\mathbb{I}| = n$ and let $\mathbb{\alpha} = (\alpha_i : i \in \mathbb{I})$. We can rewrite A.2 in matrix notation as

$$\mathbb{A}\mathbb{\alpha} = \mathbb{1}_n,$$

where $\mathbb{1}_n$ is a column vector of ones of length n,

$$\mathbb{A} = (1 + c_\eta)\mathbb{1}_n\mathbb{1}_n^\intercal + (1 + c_\eta)\mathbb{I}_n + 2\mathbb{C}, \tag{A.4}$$

where $\mathbb{I}_n$ is the identity matrix of size $n \times n$ and $\mathbb{C}$ is a diagonal matrix with entries c_j for $j \in \mathbb{I}$.

Since at step 1 we have established that A.2 always has a unique solution, we can write

$$\mathbb{\alpha} = \mathbb{A}^{-1}\mathbb{1}_n. \tag{A.5}$$

We now want to show that

$$\mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{1}_n \mathbb{1}_n^\intercal \mathbb{\alpha} - \mathbb{\alpha}^\intercal \mathbb{C} \mathbb{\alpha} = (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{\alpha} + \mathbb{\alpha}^\intercal \mathbb{C} \mathbb{\alpha}.$$

Rearranging the terms,

$$\mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{1}_n \mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{\alpha} = \mathbb{\alpha}^\intercal (2\mathbb{C}) \mathbb{\alpha}.$$

Substituting $2\mathbb{C} = \mathbb{A} - (1+c_\eta)\mathbb{1}_n \mathbb{1}_n^\intercal - (1+c_\eta)\mathbb{I}_n$ by A.4 into the right-hand side, we get

$$\mathbb{\alpha}^\intercal (2\mathbb{C}) \mathbb{\alpha} = \mathbb{\alpha}^\intercal (\mathbb{A} - (1+c_\eta)\mathbb{1}_n \mathbb{1}_n^\intercal - (1+c_\eta)\mathbb{I}_n)\mathbb{\alpha} = \mathbb{\alpha}^\intercal \mathbb{A} \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{1}_n \mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{\alpha}$$

By A.5 $\mathbb{\alpha}^\intercal = \mathbb{1}_n^\intercal \mathbb{A}^{-\intercal}$. Since $\mathbb{A}$ is symmetric, $\mathbb{A}^{-\intercal} = \mathbb{A}^{-1}$ and, hence, $\mathbb{\alpha}^\intercal \mathbb{A} \mathbb{\alpha} = \mathbb{1}_n^\intercal \mathbb{A}^{-1} \mathbb{A} \mathbb{\alpha} = \mathbb{1}_n^\intercal \mathbb{\alpha}$. Then,

$$\mathbb{\alpha}^\intercal (2\mathbb{C}) \mathbb{\alpha} = \mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{1}_n \mathbb{1}_n^\intercal \mathbb{\alpha} - (1+c_\eta)\mathbb{\alpha}^\intercal \mathbb{\alpha},$$

as desired. □

We are now ready to finish the proof.

$$\tilde{y} = \sum_{i=1}^{N} y_i + y_l = \sum_{i \in \mathbb{I}} y_i + y_l = \sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v}) + y_l$$

Note that $\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda}(\tilde{v} + \eta + \epsilon_i - \overline{v}) + y_l \sim N(0,\ Var(\tilde{y}))$, where $Var(\tilde{y}) = (\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda})^2 \sigma^2 + (\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda})^2 c_\eta \sigma^2 + \sum_{i \in \mathbb{I}} (\frac{\alpha_i}{\lambda})^2 c_i \sigma^2 + \sigma_l^2$.

Then by Lemma 1 $P = \mathbb{E}(\tilde{v} \mid \tilde{y}) = \overline{v} + \frac{Cov(\tilde{v}, \tilde{y})}{Var(\tilde{y})} \cdot \tilde{y}$.

$$\frac{Cov(\tilde{v}, \tilde{y})}{Var(\tilde{y})} = \frac{\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda} \sigma^2}{(\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda})^2 \sigma^2 + (\sum_{i \in \mathbb{I}} \frac{\alpha_i}{\lambda})^2 c_\eta \sigma^2 + \sum_{i \in \mathbb{I}} (\frac{\alpha_i}{\lambda})^2 c_i \sigma^2 + \sigma_l^2} = \lambda$$

To see that the last equality holds multiply the numerator and the denominator of the left part by λ^2, then divide both parts of the equality by λ and solve for λ^2:

$$\frac{(\sum_{i\in\mathbb{I}}\alpha_i)\sigma^2\lambda}{(\sum_{i\in\mathbb{I}}\alpha_i)^2\sigma^2+(\sum_{i\in\mathbb{I}}\alpha_i)^2c_\eta\sigma^2+(\sum_{i\in\mathbb{I}}\alpha_i^2c_i)\sigma^2+\lambda^2\sigma_l^2}=\lambda$$

$$\lambda^2=\frac{\sigma^2}{\sigma_l^2}\left(\sum_{i\in\mathbb{I}}\alpha_i-\left(\sum_{i\in\mathbb{I}}\alpha_i\right)^2(1+c_\eta)-\sum_{i\in\mathbb{I}}\alpha_i^2c_i\right)=\frac{\sigma^2}{\sigma_l^2}\sum_{i\in\mathbb{I}}\alpha_i^2(1+c_\eta+c_i)\text{ by Lemma 2}$$

Therefore, indeed $P=\mathbb{E}(\tilde{v}\mid\tilde{y})=\overline{v}+\lambda\tilde{y}$.

A.2 Proof of formula 3.5 in Proposition ??:

Let $|\mathbb{X}|=n$ and let $\mathbb{a}_n=(\alpha_i(\mathbb{X}):i\in\mathbb{X})$ and $\mathbb{a}_{n+1}=(\alpha_i(\tilde{\mathbb{X}}):i\in\tilde{\mathbb{X}})$.

By A.5

$$\mathbb{a}_n=\mathbb{A}^{-1}\mathbb{1}_n,$$

where $\mathbb{1}_n$ is a column vector of ones of length n,

$$\mathbb{A}=(1+c_\eta)\mathbb{1}_n\mathbb{1}_n^\intercal+(1+c_\eta)\mathbb{I}_n+2\mathbb{C},$$

where $\mathbb{I}_n$ is the identity matrix of size $n\times n$ and $\mathbb{C}$ is a diagonal matrix with entries c_j for $j\in\mathbb{X}$.

Analogously,

$$\mathbb{a}_{n+1}=\mathbb{A}_{n+1}^{-1}\mathbb{1}_{n+1},$$

where

$$\mathbb{A}_{n+1}=\left[\begin{array}{c:c}\mathbb{A}_n & (1+c_\eta)\mathbb{1}_n\\ \hdashline (1+c_\eta)\mathbb{1}_n^\intercal & 2(1+c_\eta+c_i)\end{array}\right]$$

We are interested in the relationship between $\sum_{j\in\mathbb{X}}\alpha_j(\mathbb{X})=\mathbb{1}_n^\intercal\mathbb{a}_n=\mathbb{1}_n^\intercal\mathbb{A}^{-1}\mathbb{1}_n$ and $\sum_{j\in\tilde{\mathbb{X}}}\alpha_j(\tilde{\mathbb{X}})=\mathbb{1}_{n+1}^\intercal\mathbb{a}_{n+1}=\mathbb{1}_{n+1}^\intercal\mathbb{A}_{n+1}^{-1}\mathbb{1}_{n+1}$.

Let $z = \mathbb{1}_n^\intercal \mathbb{A}^{-1} \mathbb{1}_n$ and let

$$\mathbb{A}_{n+1}^{-1} = \left[\begin{array}{c:c} B_{11} & B_{12} \\ \hdashline B_{21} & B_{22} \end{array}\right] \tag{A.6}$$

Using a formula for 2 × 2 block matrix inverse,

$$B_{11} = \mathbb{A}_n^{-1} + \mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n \left(2(1+c_\eta+c_i) - (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n\right)^{-1} (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1} =$$

$$\mathbb{A}_n^{-1} + (1+c_\eta)^2 \mathbb{A}_n^{-1}\mathbb{1}_n \left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} \mathbb{1}_n^\intercal \mathbb{A}_n^{-1}$$

$$B_{12} = -\mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n \left(2(1+c_\eta+c_i) - (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n\right)^{-1} =$$

$$-(1+c_\eta)\mathbb{A}_n^{-1}\mathbb{1}_n \left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1}$$

$$B_{21} = -\left(2(1+c_\eta+c_i) - (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n\right)^{-1} (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1} =$$

$$-(1+c_\eta)\left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} \mathbb{1}_n^\intercal \mathbb{A}_n^{-1}$$

$$B_{22} = \left(2(1+c_\eta+c_i) - (1+c_\eta)\mathbb{1}_n^\intercal \mathbb{A}_n^{-1}(1+c_\eta)\mathbb{1}_n\right)^{-1} = \left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1}$$

Note that B_{11} is $n \times n$, B_{12} is $n \times 1$, B_{21} is $1 \times n$, and B_{22} is 1×1. Since $\mathbb{1}_{n+1}^\intercal \mathbb{A}_{n+1}^{-1} \mathbb{1}_{n+1}$ is the sum of all entries of $\mathbb{A}_{n+1}^{-1}$, we can rewrite it as

$$\mathbb{1}_{n+1}^\intercal \mathbb{A}_{n+1}^{-1} \mathbb{1}_{n+1} = \mathbb{1}_{n+1}^\intercal \left[\begin{array}{c:c} B_{11} & B_{12} \\ \hdashline B_{21} & B_{22} \end{array}\right] \mathbb{1}_{n+1} = \mathbb{1}_n^\intercal B_{11} \mathbb{1}_n + \mathbb{1}_n^\intercal B_{12} + B_{21}\mathbb{1}_n + B_{22} =$$

$$= z + (1+c_\eta)^2 z^2 \left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} -$$

$$-2(1+c_\eta)z\left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} +$$

$$+\left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} =$$

$$= z + \left(2(1+c_\eta+c_i) - (1+c_\eta)^2 z\right)^{-1} ((1+c_\eta)z - 1)^2.$$

Hence,

$$\sum_{j\in\tilde{\mathbb{X}}}\alpha_j(\tilde{\mathbb{X}}) = \sum_{j\in\mathbb{X}}\alpha_j(\mathbb{X}) + \frac{\left(\left(\sum_{j\in\mathbb{X}}\alpha_j(\mathbb{X})\right)(1+c_\eta)-1\right)^2}{2(1+c_\eta+c_i)-\left(\sum_{j\in\mathbb{X}}\alpha_j(\mathbb{X})\right)(1+c_\eta)^2}.$$

A.3 Proof of Proposition 15:

Step 1: $\pi_{DS}^{\text{no reg}}(\mathbb{X}) = \sigma_l^2\lambda(\mathbb{X},0) = \sigma\sigma_l\sqrt{\sum_{i\in\mathbb{X}}\alpha_i(\mathbb{X},0) - (\sum_{i\in\mathbb{X}}\alpha_i(\mathbb{X},0))^2 - \sum_{i\in\mathbb{X}}(\alpha_i(\mathbb{X},0))^2c_i}$, where the first equality follows by 4.1 and the second equality follows by Lemma 2.

Let $\tilde{\mathbb{X}} = \mathbb{X}\cup\{j\}$. Then $\pi_{DS}^{\text{no reg}}(\mathbb{X}\cup\{j\}) \geq \pi_{DS}^{\text{no reg}}(\mathbb{X})$ is equivalent to

$$\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)-\left(\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)\right)^2-\sum_{k\in\tilde{\mathbb{X}}}(\alpha_k(\tilde{\mathbb{X}},0))^2c_i \geq \sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)-\left(\sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)\right)^2-\sum_{k\in\mathbb{X}}(\alpha_k(\mathbb{X},0))^2c_i. \tag{A.7}$$

Let $|\mathbb{X}| = n$ and let $\mathbb{a}_n = (\alpha_k(\mathbb{X}) : k\in\mathbb{X})$ and $\mathbb{a}_{n+1} = (\alpha_k(\tilde{\mathbb{X}}) : k\in\tilde{\mathbb{X}})$.

By A.5

$$\mathbb{a}_n = \mathbb{A}_n^{-1}\mathbb{1}_n,$$

where $\mathbb{1}_n$ is a column vector of ones of length n,

$$\mathbb{A}_n = \mathbb{1}_n\mathbb{1}_n^\intercal + \mathbb{I}_n + 2\mathbb{C}_n,$$

where $\mathbb{I}_n$ is the identity matrix of size $n\times n$ and $\mathbb{C}_n$ is a diagonal matrix with entries c_i for $k\in\mathbb{X}$.

Let

$$z = \sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X}) = \mathbb{1}_n^\intercal\mathbb{a}_n = \mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n$$

and

$$w = \sum_{k\in\mathbb{X}}(\alpha_k(\mathbb{X},0))^2c_i = \mathbb{a}_n^\intercal\mathbb{C}_n\mathbb{a}_n = \mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n.$$

Then

$$\sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)-\left(\sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)\right)^2-\sum_{k\in\mathbb{X}}(\alpha_k(\mathbb{X},0))^2c_i = z-z^2-w.$$

Using 3.5 and substituting $c_\eta = 0$, we get

$$\sum_{k \in \tilde{\mathbb{X}}} \alpha_k(\tilde{\mathbb{X}}, 0) = z + (2(1+c_j) - z)^{-1} (z-1)^2.$$

Then

$$\left(\sum_{k \in \tilde{\mathbb{X}}} \alpha_k(\tilde{\mathbb{X}}, 0) \right)^2 = z^2 + 2z\,(2(1+c_j) - z)^{-1} (z-1)^2 + (2(1+c_j) - z)^{-2} (z-1)^4$$

We would now like to express $\sum_{k \in \tilde{\mathbb{X}}} (\alpha_k(\tilde{\mathbb{X}}, 0))^2 c_i$ in terms of z, w, c_j. By A.5,

$$\mathbb{\alpha}_{n+1} = \mathbb{A}_{n+1}^{-1} \mathbb{1}_{n+1},$$

where

$$\mathbb{A}_{n+1} = \left[\begin{array}{c|c} \mathbb{A}_n & \mathbb{1}_n \\ \hline \mathbb{1}_n^\intercal & 2(1+c_j) \end{array} \right]$$

Let

$$\mathbb{C}_{n+1} = \left[\begin{array}{c|c} \mathbb{C}_n & \mathbb{0}_n \\ \hline \mathbb{0}_n^\intercal & c_j \end{array} \right]$$

where $\mathbb{0}_n$ is a column vector of zeros of length n. Then

$$\sum_{k \in \tilde{\mathbb{X}}} (\alpha_k(\tilde{\mathbb{X}}, 0))^2 c_i = \mathbb{\alpha}_{n+1}^\intercal \mathbb{C}_{n+1} \mathbb{\alpha}_{n+1} = \mathbb{1}_{n+1}^\intercal \mathbb{A}_{n+1}^{-1} \mathbb{C}_{n+1} \mathbb{A}_{n+1}^{-1} \mathbb{1}_{n+1}.$$

Using A.6 and substituting $c_\eta = 0$, we get

$$\mathbb{A}_{n+1}^{-1} = \left[\begin{array}{c|c} \mathbb{A}_n^{-1} + \mathbb{A}_n^{-1} \mathbb{1}_n \,(2(1+c_i) - z)^{-1}\, \mathbb{1}_n^\intercal \mathbb{A}_n^{-1} & -\mathbb{A}_n^{-1} \mathbb{1}_n \,(2(1+c_i) - z)^{-1} \\ \hline -\,(2(1+c_i) - z)^{-1}\, \mathbb{1}_n^\intercal \mathbb{A}_n^{-1} & (2(1+c_i) - z)^{-1} \end{array} \right]$$

$$\mathbb{A}_{n+1}^{-1} \mathbb{C}_{n+1} \mathbb{A}_{n+1}^{-1} = \left[\begin{array}{c|c} D_{11} & D_{12} \\ \hline D_{21} & D_{22} \end{array} \right]$$

where

$D_{11} = \mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1} + (2(1+c_j)-z)^{-1}\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1} + (2(1+c_j)-z)^{-1}\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1} + (2(1+c_j)-z)^{-2}\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}^\intercal\mathbb{A}_n^{-1} + (2(1+c_j)-z)^{-2}c_j\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}$

$D_{12} = -(2(1+c_j)-z)^{-1}\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}\mathbb{A}_n^{-1}\mathbb{1}_n c_j$

$D_{21} = -(2(1+c_j)-z)^{-1}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1} - (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1} - (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}c_j$

$D_{22} = (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-2}c_j$

Note that D_{11} is $n \times n$, D_{12} is $n \times 1$, D_{21} is $1 \times n$, and D_{22} is 1×1. Since $\mathbb{1}_{n+1}^\intercal\mathbb{A}_{n+1}^{-1}\mathbb{C}_{n+1}\mathbb{A}_{n+1}^{-1}\mathbb{1}_{n+1}$ is the sum of all entries of $\mathbb{A}_{n+1}^{-1}\mathbb{C}_{n+1}\mathbb{A}_{n+1}^{-1}$, we can rewrite it as

$$\mathbb{1}_{n+1}^\intercal\mathbb{A}_{n+1}^{-1}\mathbb{C}_{n+1}\mathbb{A}_{n+1}^{-1}\mathbb{1}_{n+1} = \mathbb{1}_{n+1}^\intercal\left[\begin{array}{c:c} D_{11} & D_{12} \\ \hdashline D_{21} & D_{22} \end{array}\right]\mathbb{1}_{n+1} = \mathbb{1}_n^\intercal D_{11}\mathbb{1}_n + \mathbb{1}_n^\intercal D_{12} + D_{21}\mathbb{1}_n + D_{22} =$$

$\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-1}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-1}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-2}c_j\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n -$

$(2(1+c_j)-z)^{-1}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n c_j -$

$(2(1+c_j)-z)^{-1}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n - (2(1+c_j)-z)^{-2}c_j\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{1}_n +$

$(2(1+c_j)-z)^{-2}\mathbb{1}_n^\intercal\mathbb{A}_n^{-1}\mathbb{C}_n\mathbb{A}_n^{-1}\mathbb{1}_n + (2(1+c_j)-z)^{-2}c_j =$

$w + (2(1+c_j)-z)^{-1}2w(z-1) + (2(1+c_j)-z)^{-2}(w+c_j)(z-1)^2.$

We can now express $\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0) - \left(\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)\right)^2 - \sum_{k\in\tilde{\mathbb{X}}}(\alpha_k(\tilde{\mathbb{X}},0))^2 c_i$ in terms of z, w, c_j:

$$\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0) - \left(\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)\right)^2 - \sum_{k\in\tilde{\mathbb{X}}}(\alpha_k(\tilde{\mathbb{X}},0))^2 c_i =$$

$$z + (2(1+c_j)-z)^{-1}\,(z-1)^2 - \Big(z^2 + 2z\,(2(1+c_j)+z)^{-1}\,(z-1)^2 +$$

$$(2(1+c_j)-z)^{-2}\,(z-1)^4\ \Big) - \Big(w + (2(1+c_j)-z)^{-1}2w(z-1) + (2(1+c_j)-z)^{-2}(w+c_j)(z-1)^2\Big) =$$

$$z - z^2 - w + \frac{(1-z)((4w+4z^2-5z+1)c_j - wz + 3w - z^3 + 4z^2 - 4z + 1)}{(2(1+c_j)-z)^2}$$

We can rewrite A.7 as

$$z - z^2 - w + \frac{(1-z)((4w+4z^2-5z+1)c_j - wz + 3w - z^3 + 4z^2 - 4z + 1)}{(2(1+c_j)-z)^2} \geq z - z^2 - w$$

or

$$(4w+4z^2-5z+1)c_j - wz + 3w - z^3 + 4z^2 - 4z + 1 \geq 0 \tag{A.8}$$

since $z < 1$ by Proposition 9.

Step 2: We will show that if A.8 holds, then

$\sum_{k\in\mathbb{X}\cup\{h\}} \alpha_i(\mathbb{X}\cup\{h\},0) - \left(\sum_{k\in\mathbb{X}\cup\{h\}} \alpha_i(\mathbb{X}\cup\{h\},0)\right)^2 - \sum_{k\in\mathbb{X}\cup\{h\}} (\alpha_i(\mathbb{X}\cup\{h\},0))^2 c_i$ is decreasing in c_h on $[0, c_j]$, i.e.

$$\frac{\partial}{\partial c_h}\left(z - z^2 - w + \frac{(1-z)((4w+4z^2-5z+1)c_h - wz + 3w - z^3 + 4z^2 - 4z + 1)}{(2(1+c_h)-z)^2}\right) < 0. \tag{A.9}$$

$$\frac{\partial}{\partial c_h}\left(z - z^2 - w + \frac{(1-z)((4w+4z^2-5z+1)c_h - wz + 3w - z^3 + 4z^2 - 4z + 1)}{(2(1+c_h)-z)^2}\right) =$$

$$\frac{(z-1)(2(4w+4z^2-5z+1)c_h + 4w + 3z^2 - 5z + 2)}{(2(1+c_h)-z)^3}$$

Since $z < 1$, $z - 1 < 0$ and $2(1+c_h) - z \geq 2 - z > 0$. Hence, showing A.9 is equivalent to showing that

$$2(4w+4z^2-5z+1)c_h + 4w + 3z^2 - 5z + 2 > 0. \tag{A.10}$$

We want to show that A.8 implies A.10.

Case 1: $4w+4z^2-5z+1 \geq 0$. Then $2(4w+4z^2-5z+1)c_h + 4w + 3z^2 - 5z + 2$ is either constant or increasing in c_h, and so it is sufficient to check that A.10 holds when $c_h = 0$. By assumption $4w \geq -4z^2 + 5z - 1$. Hence, $4w + 3z^2 - 5z + 2 \geq -z^2 + 1 > 0$ since $z < 1$ by Proposition 9.

Case 2: $4w+4z^2-5z+1 < 0$. Then A.8 implies that $c_j \leq \frac{wz-3w+z^3-4z^2+4z-1}{4w+4z^2-5z+1}$. We need to check A.10 for all $c_h \in [0, c_j]$ for all $c_j \in [0, \frac{wz-3w+z^3-4z^2+4z-1}{4w+4z^2-5z+1}]$ [12]. Since in this case $2(4w+4z^2-$

[12] $4w+4z^2-5z+1 < 0$ implies that $wz-3w+z^3-4z^2+4z-1 < \frac{1}{4}(z^2-1) < 0$ and, therefore, $\frac{wz-3w+z^3-4z^2+4z-1}{4w+4z^2-5z+1} > 0$.

$5z+1)c_h + 4w + 3z^2 - 5z + 2$ is a decreasing function of c_h, it is sufficient to check that A.10 holds at $c_h = \frac{wz-3w+z^3-4z^2+4z-1}{4w+4z^2-5z+1}$, i.e. that $2(wz-3w+z^3-4z^2+4z-1)+4w+3z^2-5z+2>0$.

$$2(wz-3w+z^3-4z^2+4z-1)+4w+3z^2-5z+2=$$

$$2(z-1)w+2z^3-5z^2+3z+1>2(z-1)(-z^2+\frac{5}{4}z-\frac{1}{4})+2z^3-5z^2+3z+1=$$

$$\frac{1}{2}(3-z^2)>0,$$

where the inequalities hold because $z<1$ and because $4w+4z^2-5z+1<0$ implies $w<-z^2+\frac{5}{4}z-\frac{1}{4}$.

Step 3: Remember that $\tilde{\mathbb{X}}=\mathbb{X}\cup\{j\}$. Let $\hat{\mathbb{X}}=\mathbb{X}\cup\{i\}$ for some i such that $c_i<c_j$. By step 1 since $\pi_{DS}^{\text{no reg}}(\mathbb{X}\cup\{j\})\geq\pi_{DS}^{\text{no reg}}(\mathbb{X})$,

$$\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)-\left(\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)\right)^2-\sum_{k\in\tilde{\mathbb{X}}}(\alpha_k(\tilde{\mathbb{X}},0))^2c_i\geq\sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)-\left(\sum_{k\in\mathbb{X}}\alpha_k(\mathbb{X},0)\right)^2-\sum_{k\in\mathbb{X}}(\alpha_k(\mathbb{X},0))^2c_i.$$

Then by step 2,

$$\sum_{k\in\hat{\mathbb{X}}}\alpha_k(\hat{\mathbb{X}},0)-\left(\sum_{k\in\hat{\mathbb{X}}}\alpha_k(\hat{\mathbb{X}},0)\right)^2-\sum_{k\in\hat{\mathbb{X}}}(\alpha_k(\hat{\mathbb{X}},0))^2c_i>\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)-\left(\sum_{k\in\tilde{\mathbb{X}}}\alpha_k(\tilde{\mathbb{X}},0)\right)^2-\sum_{k\in\tilde{\mathbb{X}}}(\alpha_k(\tilde{\mathbb{X}},0))^2c_i,$$

which means that $\pi_{DS}^{\text{no reg}}(\mathbb{X}\cup\{i\})>\pi_{DS}^{\text{no reg}}(\mathbb{X}\cup\{j\})$ again by step 1. This concludes the proof.

References

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

_ and _ , "Viable allocations of information in financial markets," *Journal of Economic Theory*, 1987, *43* (1), 76–115.

_ and _ , "Selling and trading on information in financial markets," *The American Economic Review*, 1988, *78* (2), 96–103.

Ausubel, Lawrence M, "Insider trading in a rational expectations economy," *The American Economic Review*, 1990, pp. 1022–1041.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan, "The economics of social data," 2019.

Caballe, Jordi and Murugappa Krishnan, "Imperfect competition in a multi-security market with risk neutrality," *Econometrica*, 1994, *62* (3), 695–704.

Carlton, Dennis W and Daniel R Fischel, "The regulation of insider trading," *Stanford Law Review*, 1983, pp. 857–895.

Chen, Zhaohui and William J Wilhelm Jr, "Sell-side information production in financial markets," *Journal of Financial and Quantitative Analysis*, 2012, pp. 763–794.

David, Joel M, Hugo A Hopenhayn, and Venky Venkateswaran, "Information, misallocation, and aggregate productivity," *The Quarterly Journal of Economics*, 2016, *131* (2), 943–1005.

DeMarzo, Peter M, Ilan Kremer, and Andrzej Skrzypacz, "Bidding with securities: Auctions and security design," *American economic review*, 2005, *95* (4), 936–959.

_ , Michael J Fishman, and Kathleen M Hagerty, "The optimal enforcement of insider trading regulations," *Journal of Political Economy*, 1998, *106* (3), 602–632.

Dridi, Ramdan and Laurent Germain, "Noise and competition in strategic oligopoly," *Journal of Financial Intermediation*, 2009, *18* (2), 311–327.

Eren, Nevzat and Han N Ozsoylev, "Communication dilemma in speculative markets," *Available at SSRN 905907*, 2006.

Fishman, Michael J and Kathleen M Hagerty, "Insider trading and the efficiency of stock prices," *The RAND Journal of Economics*, 1992, pp. 106–122.

Foucault, Thierry and Laurence Lescourret, "Information sharing, liquidity and transaction costs," *Finance*, 2003, *24*, 45–78.

Garcia, Diego and Francesco Sangiorgi, "Information sales and strategic trading," *The Review of Financial Studies*, 2011, *24* (9), 3069–3104.

Hansen, Robert G, "Auctions with contingent payments," *The American Economic Review*, 1985, *75* (4), 862–865.

Jain, Neelam and Leonard J Mirman, "Insider trading with correlated signals," *Economics Letters*, 1999, *65* (1), 105–113.

Kyle, Albert S, "Continuous auctions and insider trading," *Econometrica: Journal of the Econometric Society*, 1985, pp. 1315–1335.

_ , "Informed speculation with imperfect competition," *The Review of Economic Studies*, 1989, *56* (3), 317–355.

Lambert, Nicolas S, Michael Ostrovsky, and Mikhail Panov, "Online Appendix to "Strategic Trading in Informationally Complex Environments"," 2017.

Leland, Hayne E, "Insider trading: Should it be prohibited?," *Journal of Political Economy*, 1992, *100* (4), 859–887.

Manne, Henry G, "Insider trading and the law professors," *Vand. L. Rev.*, 1969, *23*, 547.

Pasquariello, Paolo, "Imperfect competition, information heterogeneity, and financial contagion," *The Review of Financial Studies*, 2007, *20* (2), 391–426.

Shin, Jhinyoung, "The optimal regulation of insider trading," *Journal of Financial Intermediation*, 1996, *5* (1), 49–73.

Wurgler, Jeffrey, "Financial markets and the allocation of capital," *Journal of financial economics*, 2000, *58* (1-2), 187–214.