Valuing Financial Data

Maryam Farboodi* Dhruv Singal†

Laura Veldkamp‡ Venky Venkateswaran§

March 2, 2022¶

Abstract

How should an investor value financial data? The answer is complicated because it depends on the characteristics of all investors. We develop a sufficient statistics approach that uses equilibrium asset return moments to summarize all relevant information about others' characteristics. It can value data that is public or private, about one or many assets, relevant for dividends or for sentiment. While different data types have different valuations, heterogeneous investors value the same data very differently, which suggests a low price elasticity for data demand. Heterogeneous investors' data valuations are also affected very differentially by market illiquidity.

*MIT Sloan, NBER and CEPR; farboodi@mit.edu

†Columbia Business School; dhruv.singal@gsb.columbia.edu

‡Columbia Business School, NBER, and CEPR; lv2405@columbia.edu

§NYU Stern School of Business and NBER; vvenkate@stern.nyu.edu

¶Thanks to Adrien Matray and Vincent Glode for valuable conversations and suggestions. Keywords: Data valuation, portfolio theory, information choice.

Investment management firms are gradually transforming themselves from users of small data and simple asset pricing models to users of big data and computer-generated statistical models. Amidst this transformation, investors' strategic focus is shifting from the choice of pricing model to the choice of data they acquire. A key question for modern financial firms is: How much should they be willing to pay for a stream of financial data? This paper devises and puts to use a methodology to estimate this dollar value, based the investor's own characteristics, but without needing to know the characteristics of others.

From information-based theories, we know many qualitative features of firms that make data valuable – large firms, growth stocks, firms with risky payoffs, assets that are sensitive to news, assets that others are uninformed about. After all, data is simply a stream of digitized information. But for an investor who is considering purchasing a data set, knowing the representative investor's theoretical value for the data is not very useful. An investor with a large portfolio values data more, while an investor who invests in a restricted set of assets values data less. An investor with lots of other data is less willing to pay for additional data, while an investor who trades more frequently might value data more or less. All these effects depend on the asset market equilibrium, which in turn depends on the characteristics of every other investor. Data value also depends on which other investors buy that same data. To make matters more complex, we also know that illiquidity or price impact of a trade make information less valuable (Kacperczyk, Nosal, and Sundaresan, 2021), but how this interacts with investor heterogeneity, quantitatively, is less understood.

Our simple procedure to estimate the value of any data series, to an investor with specific characteristics, reveals enormous dispersion in how different investors value the same data. Unlike financial assets, data assets are not equally valuable to all. The dispersion in private valuations for data matters for our understanding of data markets because it suggest a low price elasticity of aggregate data demand.

It is important to point out that our procedure leads to an estimate of private *value* to an investor, which could be different from a transaction price that one might observe when data

is sold. Knowing the private values of market participants allows one to trace out a demand curve for data. Some investors would have values greater than the equilibrium price, some less. This is like a shopper determining how much they value a sweater. Knowing that the sweater's market price is $50 does not make that the shopper's value – it might be the wrong color or size. Alternatively, the shopper might be willing to pay $100 for the sweater and still not buy it because they find a similar sweater for less. Understanding how customers (investors) value a product (a data set) is different from calculating a market clearing or equilibrium price. Valuations are important because they allow us to evaluate consumer surplus and welfare, teach us about demand elasticity, markups and market competition, and allow one to ask if observed transactions prices are efficient.

Our measurement approach relies on sufficient statistics which are easily computable. While our measure is based on a model, we do not need to estimate most model parameters to arrive at a data value. In Section 1, we set up a noisy rational expectations model with rich heterogeneity in investors, assets and data types and derive the expected utility of data in dollar amounts. We show that a few sufficient statistics – average conditional and unconditional returns, variances and forecast errors – are all that is needed to price a piece or a stream of data. This is true regardless of whether the data is public, private, or known by some. Our sufficient statistics are also a valid measure regardless of how heterogeneous other investors' preferences, data or investment styles are. They can be used to value data about asset fundamentals or about sentiment. Finally, with a small adjustment, they can be used in imperfectly competitive markets as well.

One could apply this tool to any finance-relevant data series, or any bundle of data series – all it requires is knowledge of the relevant investor characteristics and access to a history of market prices and data realizations. We present a small number of examples that highlight the importance of accounting for investor heterogeneity in data valuation. In Section 3, we compute the value of median analyst forecasts for earnings growth for investors with different wealth levels, different investment styles and facing different market conditions.

These exercises highlight the flexibility of our approach and its ability to accommodate various dimensions of heterogeneity. In Section 4, we then apply our tool to assign a dollar value to information about macroeconomic fundamentals. Specifically, we compute the value to an investor from being able to perfectly observe real GDP growth one quarter in advance. We again highlight the role of investor heterogeneity in arriving at this data value.

Our first exercise in Section 3 explores the role of investor wealth and risk preferences. To do this, we consider two investors with the same relative risk aversion and different initial wealth levels. This implies that the wealthier investor has lower absolute risk aversion and as a result, values the same data by more. But, the extent depends on market structure, i.e. on whether their trades have price impact or not. When markets are competitive, i.e. a trade has no impact on the market price, data values increase sharply with wealth: an investor with $250 million in initial wealth values data by almost 300 times compared to one with $0.5 million. Accounting for price impact, in line with empirical estimates, dramatically reduces the value of data for all investors, but has noticeably larger effects on the investor with higher wealth/lower absolute risk aversion. This illustrates a general pattern we see – there is enormous heterogeneity in willingness to pay for data, that is substantially tempered by a modest degree of market illiquidity.

The high sensitivity of data to price impact is interesting in its own right. It suggests that market liquidity is crucial for the value of financial data. Small changes in market conditions can thus lead to large variation in data value and through that, in the valuation of firms whose main asset is financial data. This suggests a new avenue of how liquidity effects in asset markets. We typically think of market liquidity as something that affects only the value of financial assets, not directly affecting the real value of a firm. As data becomes a more important asset for financial firms, the prices of financial firms may become increasingly sensitive to market liquidity.

Our second exercise considers investors with different investment styles. Specifically, we analyze the value of analyst forecast data for investors who trade only in a single portfolio,

such as the S&P 500 portfolio, or a portfolio consisting of only small firms, only large firms, only growth stocks or only value stocks. Our benchmark is an investor who trades all five of these portfolios. Because each of these types uses a piece of data differently, they value the same piece of information differently. We find that investors in large and growth stocks (as well as investors who trade all five portfolios) value analyst forecast data substantially more than a value or small-firm investor.

Our third exercise quantifies how much the value of analyst forecast data depends on what other data is in an investor's database. We find considerable variation in data values when we vary the other data variables used. In general, the more series we add to the investor's information set, the lower is the value gained by having access to analyst forecasts and these effects are sizable. This result illustrates the importance of accounting for many facets of investor heterogeneity. It also suggests that this dimension of heterogeneity can induce sizable variation in data valuations, and potentially, a low price elasticity of data demand.

Our fourth exercise considers the effect of trading horizon on the value of analyst forecast data. Our toolkit can easily accommodate such differences with higher frequency observations on the data series and asset returns. We illustrate this by computing the value of data to an investor who trades over a quarterly horizon (our baseline calculations are for an annual horizon). We find that a shorter horizon makes analyst forecast data somewhat less valuable, i.e., they turn out to be less useful in forecasting returns over shorter horizons. It is worth highlighting that this exercise is about trading horizon, not frequency – it is possible that an investor who trades or rebalances his portfolio more frequently ascribes a higher value to the data compared to one who trades less often. In principle, our procedure can be extended to this type of heterogeneity as well, but in part due to data limiations, we do not explore it in this paper.

In Section 4, we explore how investors with different characteristics value macroeconomic information. In contrast to the analyst forecast data, we find less variation across different

investment styles for this type of data. The estimated dollar values are sizeable, suggesting all investors in general find macroeconomic information quite useful for portfolio choice. Alternatively, similar to analyst forecast data, we find that market illiquidity not only decreases value of data to all investors, but also significantly compresses data valuation across investors with different wealth levels.

As such, these exercises not only highlight how our toolkit can be applied in practice but also yield new insights about financial asset markets.

Why do we need to estimate the value of data? Why not look at prices for data directly? One reason is that not all data prices are observed, either because the data is not traded, or it is traded privately. In other words, the data is an asset, and if it is owned by a firm but never traded, it does affect the value of the firm while its price is unknown. But even if all prices were observed, just like assets can be mispriced, data can be mispriced. Finally, a firm's willingness to pay for data depends on what data it already owns. A market or transaction price for data does not necessarily reflect how any one firm values the data.

Relationship to the literature. Data is information. Therefore, our approach to valuing financial data draws primarily on the literature exploring information in financial markets. A few papers have examined the value of information or skill, for a representative agent or in an economy with one aggregate risk (Kadan and Manela, 2019; Savov, 2014; Dow, Goldstein, and Guembel, 2017; Morris and Shin, 2002). Kacperczyk, Nosal, and Sundaresan (2021), Kyle and Lee (2017), and Kyle (1989) add imperfect competition. What we add is a richer asset structure, a richer information structure, but most importantly, heterogeneous investors who value information differently. This last ingredient is essential to understand what the aggregate data demand function looks like.

Enriching the information structure to allow for public, private or correlated signals is also important for real-world measurement. Such rich information structures are commonly studied in settings with quadratic payoffs (Ozdenoren and Yuan, 2008; Albagli, Hellwig, and Tsyvinski, 2014; Amador and Weill, 2010). But they have substantially complicated

previous asset market models to the point that most authors assume fully private (Barlevy and Veronesi, 2000; He, 2009; Kondor, 2012) or fully common (Grossman and Stiglitz, 1980) information.[1] In addition, investors may choose between asset valuation-relevant data or data about other investors' order flow (Farboodi and Veldkamp, 2017). The idea that all these types of information can be valued with one set of sufficient conditions is a new idea that substantially broadens the empirical applicability of these tools.

The main point of the paper is that heterogeneity in investor characteristics matters. Some version of all these characteristics exist in some noisy rational expectations model (Kacperczyk, Nosal, and Sundaresan, 2021; Peress, 2004; Mondria, 2010), most of which look daunting to estimate.[2] This project shows that, despite all these degrees of heterogneity among investors, data types and equilibrium effects, there is a simple procedure to compute a value for data.

Measures of the information content of prices, like those in Bai, Philippon, and Savov (2016) and Davila and Parlatore (2021) are used to infer how much the average investor in an asset knows. Such measures are related, in that they arise from a similar noisy rational expectations framework. But they answer a question about the quantity of information, not its value. Farboodi, Matray, Veldkamp, and Venkateswaran (2019)'s "initial value" of a unit of precision is not the value a firm would pay, is only valid for private signals about orthogonal assets, and does not account for any particular firm's preferences, portfolio, existing data set or price impact. Our sufficient statistics approach is more relevant for demand estimation, much simpler to estimate and more robust to heterogeneity.

[1] Exceptions include Goldstein, Ozdenoren, and Yuan (2013) and Sockin (2015).

[2] Heterogeneity also arises in micro models like (Bergemann, Bonatti, and Smolin, 2016), who value information in a bilateral trade, where sellers do not know buyers' willingness to pay, but without the equilibrium considerations about what others know.

1 A Framework for Valuing Data

Since data is information, we build on the standard workhorse model of information in financial markets, the noisy rational expectations framework. To the framework, we add long-lived assets, imperfect competition, heterogeneity of preferences, wealth effects, investment styles, public, private or partly public signals and arbitrary correlation between assets and between various signals. We include these features because each one affects the value of information. Model extensions consider data about sentiment or order flow.

Our contribution is not the modeling. Our contribution lies in showing how to estimate data valuations in such a rich and flexible model. The goal of the model is to show how, despite all the heterogeneity, the value of data can be reduced to a few sufficient statistics that are easy to compute. Later, we justify this rich modeling structure by showing that heterogeneity matters for data valuations.

Assets We have N distinct risky assets in the economy indexed by j, with net supply given by $\bar{x}$. Each of these assets are claims to stream of dividends $\{d_{jt}\}_{t=0}^{\infty}$, where the vector d_t is assumed to follow the auto-regressive process

$$d_{t+1} = \mu + G(d_t - \mu) + y_{t+1}.$$

Here, the exogenous dividend innovation shock $y_{t+1} \sim \mathcal{N}(0, \Sigma_d)$ is assumed to be i.i.d. across time. We use subscript t for variables that are known before the end of period t. Thus, the dividend d_{t+1} and its innovation shock y_{t+1} both pertain to assets that are purchased in period t; both these shocks are observed at the end of period t.

Investors and investment styles In each period t, n overlapping generations investors, $i \in [0, 1]$, are born, observe data, and make portfolio choices. The number of investors may be finite, which implies that markets are imperfectly competitive. We will also consider the limiting economy as n becomes infinite. In the following period $t + 1$, investors sell

their assets, consume the dividends and the proceeds of their asset sale and exit the model. Each investor i born at date t has initial endowment $\bar{w}_{it}$ and utility over total, end-of-life consumption c_{it+1}. At date t, investors choose their portfolio of risky assets, which is a vector q_{it} of the number of shares held or each asset. They also choose holdings of one riskless asset with return r, subject to budget constraint

$$c_{it+1} = r\left(\overline{w}_{it} - q_{it}'\theta_i p_t\right) + q_{it}'\theta_i\left(p_{t+1} + d_{t+1}\right). \tag{1}$$

An investor i may also be subject to an investment style constraint, which limits the set of risky assets they purchase. We denote this set of investable assets as $\mathcal{Q}_i$. Following, Koijen and Yogo (2019), we do not model the source of the constraint. However, many investors do describe their strategy as small-firm investing or value investing, which limits the assets they hold. We consider sets $\mathcal{Q}_i$ that either set the holdings of some assets to zero, or allow the entire real line. For example, long-only portfolios would restrict $\mathcal{Q}_i$ to the non-negative realm of Re^N. Of course, it is possible that an investor is unrestricted, in which case $\mathcal{Q}_i = \mathrm{Re}^N$. The matrix θ_i is an $m_i \times N$ matrix of zeros and ones, where $m_i \equiv |\mathcal{Q}_i|$ is the number of investable assets for investor i. Each row of θ_i has a single 1 entry, with all other entries zero. If asset j is in investor i's style class, then that asset is investable and there will be one row of θ_i with ith column entry equal to 1.

Data Each investor has access to H distinct data sources. Signals from each of these data sources (indexed by h) provides information about dividend innovations y_{t+1}, possibly from a linear combination ψ_h of assets:

$$\eta_{iht} = \psi_h y_{t+1} + \Gamma_h e_{it}$$

Here, $e_{it} \sim \mathcal{N}(0, I)$ is iid across time, but not necessarily independent across investors or across assets. In other words, data can have public and private signal noise. Public signal

noise captures the idea that many data sources are available to, observed and used by many investors. In addition, all investors know the variance and covariance of prices, dividends and the data they observe.

External Demand Some source of noise in prices is necessary to explain why some investors know information that others do not. We assume the economy is populated by a unit measure of noise traders. Their demand could come from hedging motives, estimation error, cognition errors or sentiment.[3] Each noise trader buys x_{t+1} shares of the asset, where $x_{t+1} \sim N(0, \Sigma_x)$ is independent of other shocks in the model and independent over time. The noise can be arbitrarily small, as long as $\Sigma_x > 0$. Similar to the dividend d_{t+1} and its innovation shock y_{t+1}, the shock x_{t+1} is observed at the end of period t.

Equilibrium An equilibrium is a sequence of prices $\{p_t\}_{t=0}^{\infty}$ and portfolio choices $\{q_{it}\}_{t=0}^{\infty}$, such that

1. At the beginning of each period t, all investors have information set $\mathcal{I}_t^- = \{\mathcal{I}_{t-1}, y_t, d_t, x_t, z_t\}$, where $\mathcal{I}_{t-1}$ is the information set of the average investor at time $t-1$ (averaged over private signal realizations).

2. Investors use Bayes' Law to combine prior information $\mathcal{I}_t^-$ with data $\{\eta_{iht}\}$, and p_t to update beliefs. The information set at the time of portfolio choice is $\mathcal{I}_{it} = \{\mathcal{I}_t^-, \eta_{it}, p_t\}$.

3. Investors choose their risky asset investment q_{it} to maximize $\mathbb{E}\left[U(c_{it+1})|\mathcal{I}_{it}\right]$, taking the actions of other investors as given, subject to the budget constraint (1) and the investment style constraint $q_{it} \in \mathcal{Q}_i$.

4. At each date t, the risky asset price vector p equates demand plus noise x_{t+1} to a vector

[3] In other words, x_{t+1} includes whatever is unrelated to payoffs. If it is persistent, and therefore payoff relevant, the persistent component should be included in the payoff structure. In previous work, micro-founded heterogeneous investor hedging demand has been shown to rationalize this trading behavior. See Kurlat and Veldkamp (2015).

$\bar{x}$ units of supply:

$$\int_i q_{it} di + x_{t+1} = \bar{x} \qquad \forall t. \tag{2}$$

Equilibrium Solution To solve the model and derive the value of data, we first apply Bayes' law to investors' prior beliefs and data to form posterior beliefs about asset payoffs. Getting this combination of private, public and price information is equivalent to getting an unbiased signal s_{it} about the dividend innovation y_{t+1}, with private signal noise ξ_{it} and public signal noise z_{t+1}.

$$s_{it} = y_{t+1} + \zeta_{it} z_{t+1} + \xi_{it}$$

The term $z_{t+1} \sim \mathcal{N}(0, \Sigma_z)$ comes from the noise in public component of the any data. It is iid across time, with precision Σ_z^{-1}. This public signal noise z_{t+1} pertains to assets that are purchased in period t and is observed at the *end* of period t. If investor i learned nothing from any public sources of information at date t, then $\zeta_{it} = 0$ and this becomes a standard private signal. Similarly, $\xi_{it} \sim \mathcal{N}(0, K_{it}^{-1})$ is the noise in the private component of the signal (iid across individuals and time), which has the precision K_{it}, orthogonal to the noise of the public component.

Next, we take a second-order approximation to the utility function. This allows us to write the unconditional and conditional expected utility at time t as

$$\mathbb{E}\left[U(c_{it+1})\right] = \rho_i \mathbb{E}\left[c_{it+1}\right] - \frac{\rho_i^2}{2} \mathbb{V}\left[c_{it+1}\right], \text{ and} \tag{3}$$

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i \mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2} \mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right]. \tag{4}$$

Here, ρ_i denotes the coefficient of absolute risk aversion for investor i, which can be an arbitrary function of their endowment $\overline{w}_{it}$. Note that even though $\overline{w}_{it}$ is time-varying, ρ_i can be a function of this endowment since the investors are born in overlapping generations, and only live for one period.

Finally, for a perfectly competitive market ($n \to \infty$), we show in Appendix A that the exists an equilibrium price schedule that is linear in current dividend d_t, future dividend innovations y_{t+1} that investors learn about through data, demand shocks x_{t+1} and the noise in public data z_{t+1}.

$$p_t = A_t + B(d_t - \mu) + C_t y_{t+1} + D_t x_{t+1} + F_t z_{t+1} \tag{5}$$

Mapping Data Utility to Sufficient Statistics Our first result uses the law of iterated expectations to compute unconditional expectation (3) in terms of means and variances of the vector of asset profits Π_{it}, defined below. Since we have substituted out the optimal consumption, we replace the direct utility function which takes consumption as its argument, with an indirect expected utility function $\tilde{U}$ which takes an information set $\mathcal{I}_{it}$ as its argument.

In order to state the main result we need to define Π_{it}, the vector of profits from buying each asset in investor i's feasible investment set, at time t,

$$\Pi_{it} \coloneqq \theta_i \left[p_{t+1} + d_{t+1} - r p_t \right]. \tag{6}$$

Lemma 1. *In a competitive market ($n \to \infty$), investor unconditional expected utility can be expressed as*

$$\tilde{U}(\mathcal{I}_{it}) = \frac{1}{2} \mathbb{E}\left[\Pi_{it}\right]' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} \mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2} Tr\left[\mathbb{V}\left[\Pi_{it}\right] \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\rho_i \bar{w}_{it} \tag{7}$$

where Tr is the matrix trace and $\bar{w}_{it}$ is investor i's exogenous endowment.

Proof is in Appendix B.

Equation (7) illustrates the basis for our measurement strategy. The value of data is this expected utility with the piece of data, minus this expected utility without that piece of data.

The first term is the expected profit on individual i's portfolio. The role of more or better data is to reduce conditional variance $\mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]$. In other words, an investor's utility rises with data if she can use the data to make forecasts with smaller squared forecast errors. Smaller forecast errors are valuable because they allow the investor to buy more of assets that will ultimately have higher returns — the first term captures utility gain through *expected profit*. The second term captures the benefit of data lowering the risk of the portfolio, which increases utility for a risk-averse investor — the second term represents utility gain through *variance reduction*.

One might object that data should also enter in the expected payoff. Data will affect the conditional beliefs about asset profits $\mathbb{E}[\Pi_{it} \mid \mathcal{I}_{it}]$, but not the unconditional, ex-ante, expected profit $\mathbb{E}[\Pi_{it}]$. The reason data cannot affect our *ex-ante* expected profit is simply that beliefs are martingales: If before seeing data, I believe that such data will make me more optimistic about an asset's return, then I should raise my expectation of that return right now.

In a imperfectly competitive market, expected utility takes a similar form, but with price-impact-adjusted variances.

Lemma 2. *Unconditional expected utility, for an investor with price impact dp/dq_i is*

$$\tilde{U}(\mathcal{I}_{it}) = \mathbb{E}[\Pi_{it}]' \hat{V}_i^{-1} \mathbb{E}[\Pi_{it}] + Tr\left[\left(\mathbb{V}[\Pi_{it}] - \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]\right) \hat{V}_i^{-1}\right] + r\rho_i \bar{w}_{it}. \tag{8}$$

where $\hat{V}_i^{-1} := \tilde{V}_i^{-1}\left(1 - \frac{1}{2}\mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]\tilde{V}_i^{-1}\right)$ *and* $\tilde{V}_i := \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}] + \frac{1}{\rho_i}\frac{dp}{dq_i}$.

Proof is in Appendix B.

Notice that if $dp/dq_i = 0$, then $\frac{\hat{V}_i}{2} = \tilde{V}_i = \mathbb{V}[\Pi_{it} \mid \mathcal{I}_{it}]$ and we get the expression in Lemma 1.

This formula explains another important features of our results. Multiplying dp/dq_i is an investor's risk tolerance $1/\rho_i$. Since this is absolute risk aversion and we know that absolute risk aversion declines in wealth, one can interpret this as a proxy for investor wealth. In

equilibrium, an investor with lower absolute risk aversion will have larger trade sizes, and their equilibrium trades will have more price impact.

The price impact of *all* investors' trades would seem to matter for the value of data. It does. But once again, it is captured by the variances. Other investors' price impact enters this expression through the equilibrium price coefficient C. This, in turn, shows up in the mean and variance of Π_{it}. Since we measure then mean and variance of Π_{it} directly, we do not need to know the extent of other investors' market power or explicitly account for it. This effect is already incorporated in our sufficient statistics.[4] As long as we can measure these sufficient statistics, and we know investor i's market power, we can accurately compute the value of investor i's data.

As before, data value is the difference between expected utility with and without the data. When we make this calculation, we are calculating the value of an investor taking as given the best responses of all other investors. Of course, if the investor is large, it is possible that knowledge of this data choice will change the behavior of other investors – we abstract from this possibility, by positing a surprising, one-time deviation.

The two key assumptions behind both the competitive and market power results are that price can be approximated as a linear function of innovations as in equation (5), and that individual i maximizes risk-adjusted return. In other words, this calculation is accurate as long as investors use linear factor models and maximize risk-adjusted return, even with potentially heterogeneous prices of risk.

Private, Public and Correlated Information At first pass, this result is unsurprising. This type of expected utility expression shows up in many noisy rational expectations models, dating back to Grossman and Stiglitz (1980). But what is perhaps surprising is that the exprssion for utility does not depend on a number of potentially complicating factors and heterogeneity.

[4]Market power does change the interpretation of C as a measure of price informativeness. But how one interprets the price coefficient C, in this case, does not affects its use in assessing data value.

In particular, the expression for data value is the same for public and private information – regardless of who else knows the data, it is valuable only for its ability to change the conditional forecast errors. This might seem to contradict what we know about information value, e.g. (Glode, Green, and Lowery, 2012). The reconciliation comes from the fact that the publicity of the data does matter for the conditional variance. Private information, which is less likely to be impounded into price, is typically more valuable compared to information that the market already knows (and is therefore uncorrelated with $p_{t+1} + d_{t+1}$). Public information about $p_{t+1} + d_{t+1}$ is already impounded in rp_t.

In short, knowing the forecast errors fully captures the way in which knowledge matters: conditional variances, or in other words, the properties of forecast errors, are sufficient statistics. This is an incredibly helpful property because it relieves the econometrician of having to figure out who knows what.

Similarly, the risk preferences and investment styles of all market participants matter for the value of data. However, the expected profit $\mathbb{E}[\Pi_{it}]$ captures the way in which risk preferences and investment mandates matter.

Mapping Utility to a Dollar Value The dollar value of data is the amount of risk-free return an investor would require to be indifferent between having the data, or not having the data but getting the additional riskless wealth. Our utility function takes the form of risk aversion times expected wealth, minus a risk-adjustment. Thus, dividing the difference in utility by the coefficient of absolute risk aversion delivers a certainty equivalent amount:

$$\$\text{Value of Data}_i = \frac{1}{\rho_i}\left(\tilde{U}(\mathcal{I}_{it} \cup \text{data}) - \tilde{U}(\mathcal{I}_{it})\right) \tag{9}$$

Of course, that leaves open the question of what an investor's absolute risk aversion is. One way to impute such a value is to assume the investor has constant relative risk aversion (CRRA), with a risk aversion coefficient of σ. Then, we can compute the level of absolute risk aversion that corresponds to relative risk aversion of σ. We will use $\sigma = 2$, a conservative

estimate. We equate a standard power utility function (CRRA) to a standard exponential utility function (CARA), and then solve for the absolute risk aversion ρ that equates the two functions at relative risk aversion of $\sigma = 2$ and a wealth level of c.

Thus, absolute risk aversion is the value of ρ that equates

$$\frac{c^{1-\sigma}}{1-\sigma} = -\exp^{-\rho c}.$$

For a relative risk aversion $\sigma = 2$, the absolute risk aversion is

$$\rho = \frac{1}{c} ln(c).$$

For example, the imputed coefficient of absolute risk aversion for an investor with wealth level $c = \$500,000$ will be $\rho = 2.6 \times 10^{-5}$, while the imputed coefficient for an investor with wealth level $c = \$250$ million takes the value $\rho = 7.7 \times 10^{-8}$.

An alternative approach to estimating ρ could be to use the market price of risk. Using the formulas for the equilibrium price coefficients, one could map the value of ρ to an equity premium and choose the value that matches a preferred estimate of the equity premium. We do not follow that approach for two main reasons. First, this would give us an estimate of the market's risk aversion and therefore, on how an average investor in the market values data. We are interested in how an individual investor, with particular characteristics should value data and in understanding how investor heterogeneity matters for data valuation. Second, it requires estimating most of the structural parameters of the model. As such, the estimates becomes much more sensitive to the exact model structure and choices of how to estimate each object, and counteracts the advantage of our simple sufficient statistics approach.

Data About Order Flow or Sentiment Many new data sources teach us about how others investors feel about an asset. For example, analyzing a twitter feed is unlikely to turn up new dividend information. But it might well correlate with the current price because it

detects sentiment. Sentiment is something unrelated to the fundamental asset value, that affects current demand. In our model, the variable that moves current price in a way that is orthogonal to value is x_{t+1}. So, we interpret sentiment as something that shows up in x, thus sentiment data are time-t signals about price noise x_{t+1}.

Put differently, our base model is set up to value data which are signals about future cash flows of a firm. But this tool can also be used to value data series about sentiment, order flow, or aspects of demand that are orthgonal to future cash flows but may affect the current price. In fact, Appendix D shows that such data can be valued using Equation (7) and Equation (9), just as if this were cash flow data.

Of course, many structural aspects of this model with sentiment data change. If we were to estimate the underlying parameters from order flow data, many adjustments would be necessary. But the essence of Farboodi and Veldkamp (2020) is to show that such data can be used to remove the noise from the price signal and thus better forecast earnings. Doing this is functionally equivalent to trading against dumb money, a common practice for sophisticated traders with access to retail order flow. The fact that such trading activity can be formally represented as if sentiment/order flow data were being used in a linear combination with current prices to forecast cash flows, means that estimating cashflows conditional on prices and sentiment data yields a valid estimate of data value.

2 Data and Estimation Procedure

In this section, we describe our estimation procedure in detail and the data series used.

Excess Returns To build a tighter connection with the asset pricing literature, we reformulate our data value expression in terms of returns. Excess return on assets in the investment set is defined as:

$$R_{it} \coloneqq \theta_i \left[(p_{t+1} + d_{t+1}) \oslash p_t - r \right] = \Pi_{it} \oslash \theta_i p_t, \tag{10}$$

where $\oslash$ represents the Hadamard (element-by-element) division of two matrices. The binary θ_i matrix pre-multiplying returns selects out only the subset of returns that are for assets the investor can hold, given their investment style constraint. This ensures that investors do not get expected utility from assets they cannot hold, and drops out ($\theta_i = I$) for investors with no constraints.

The investor unconditional expected utility in Lemma 1 and Lemma 2 are expressed in terms of Π_{it}. In Appendix C, we derive expressions for *ex-ante* expected utility expressions in Lemma 1 and Lemma 2 in terms of moments of returns.[5] In the case of perfect competition ($n \to \infty$), expected utility is

$$\tilde{U}(\mathcal{I}_{it}) \approx \frac{1}{2}\left\{\mathbb{E}\left[R_{it}\right]' \mathbb{E}\left[\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1}\right] \mathbb{E}\left[R_{it}\right]\right\} + \frac{1}{2}\operatorname{Tr}\left[\mathbb{V}\left[R_{it}\right] \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\overline{w}_{it}\rho_i. \tag{11}$$

If investors have price impact, expected utility is

$$\tilde{U}(\mathcal{I}_{it}) \approx \mathbb{E}\left[R_{it}\right]' \hat{\tilde{V}}_{it}^{-1} \mathbb{E}\left[R_{it}\right] + \operatorname{Tr}\left[\left(\mathbb{V}\left[R_{it}\right] - \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\right) \hat{\tilde{V}}_{it}^{-1}\right] + r\rho_i\bar{w}_{it}, \tag{12}$$

where $\hat{\tilde{V}}_{it} \coloneqq \tilde{\tilde{V}}_{it}\left(I - \frac{1}{2}\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\tilde{\tilde{V}}_{it}^{-1}\right)^{-1}$ and $\tilde{\tilde{V}}_{it} \coloneqq \left(\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i} \oslash \theta_i p_t p_t' \theta_i'\right)$.

Estimation Procedure The first step is to construct a time series of the return vector R_t by computing returns for each asset type, R_{jt}. The estimates for unconditional expected return $\mathbb{E}\left[R_t\right]$ and variance $\mathbb{V}\left[R_t\right]$ are obtained from the corresponding time series moments, i.e., $\widehat{\mathbb{E}\left[R_t\right]} = \frac{1}{T}\sum_{t=1}^{T} R_t$ and $\widehat{\mathbb{V}\left[R_t\right]} = \frac{1}{T-1}\sum_{t=1}^{T}\left(R_t - \widehat{\mathbb{E}\left[R_t\right]}\right)^2$.

Our strategy requires a historical time series of the data-set they are interested in valuing. The next step is to project R_t on the available time series of the data along with any other data that the investor already has access to. In our empirical implementation, we will use standard controls (such as the S&P 500 dividend-price ratio) as a proxy for such existing

[5]This requires an assumption about the *ex-ante* variability of p_t. The key patterns in data valuation described in the following sections hold even when we work directly with profits using the expressions in Lemma 1 and Lemma 2.

data. The procedure is a ordinary least squares regression of returns R_t on all the variables, already owned and new, in the data set. The estimated variance of the residuals is then our estimate for $\mathbb{V}\left[R_t \mid \mathcal{I}_{it}\right]$.

Using these objects, we can compute $\mathbb{E}\left[U(c_{it+1})\right]$. We then repeat this procedure excluding the data series of interest, i.e., with only the already-owned data. The difference between these two expected utilities is the utility gain from having access to that data source.

Formally, given data, denoted X_t, and existing data, denoted Z_t, we can estimate the data added precision $\mathbb{V}\left[R_t \mid X_t, Z_t\right]^{-1}$ and $\mathbb{V}\left[R_t \mid Z_t\right]^{-1}$ by estimating the following two regressions:

$$R_t = \beta_1 X_t + \beta_2 Z_t + \varepsilon_t^{XZ} \tag{13}$$

$$R_t = \gamma_2 Z_t + \varepsilon_t^{Z} \tag{14}$$

From these two vector regressions, an estimate for $\mathbb{V}\left[R_t \mid \mathcal{I}_{it}\right]$ would be $\widehat{\text{Cov}}(\varepsilon_t^{XZ})$. For a data set with observations $1, \ldots, T$, this estimate is $\frac{1}{T-|X|-|Z|}\sum_{t=1}^{T} \varepsilon_t^{XZ}\varepsilon_t^{XZ\prime}$. Similarly, the estimate for $\mathbb{V}\left[R_t\right]$ would be $\widehat{\text{Cov}}(\varepsilon_t^{Z})$. With a finite sample, the approximate variance-covariance matrix of residuals is $\frac{1}{T-|Z|}\sum_{t=1}^{T} \varepsilon_t^{Z}\varepsilon_t^{Z\prime}$, where $|X|$ and $|Z|$ are the number of data series that comprise X_t and Z_t, including the constant in Z_t. For most of the calculations that follow, $|X| = |Z| = 2$. Substituting in the mean return and the estimated variance-covariance matrices in Equation (7) yields the estimated value of data, in utils.

One might question how a Bayesian theory corresponds to a procedure that uses OLS. When variables are normal and relationships are linear, Bayesian estimates are the efficient, unbiased estimates. Since OLS estimates are the unique efficient, unbiased linear estimates, they must coincide with the Bayesian ones, in the specific case of normal variables in a linear relationship. Thus, in this case, OLS estimators are Bayesian weights on information. In cases where variables are not normal or the expected relationship between the data and R_t is not linear, there are a few possible solutions: 1) Transform the data to make it normal or

linear; 2) use OLS or non-linear least squares as an approximation to the Bayesian forecast, or 3) perform a full-fledged Bayesian estimation.

Data Sources for Asset Prices and Cashflows All data are for the U.S. equity market, over the period 1985–2015. Stock prices come from CRSP (Center for Research in Security Prices). All accounting variables are from Compustat. For our annual calculations, we measure prices at the end of the calendar year and dividends per share paid throughout the calendar year. In line with common practice, we exclude firms in the finance industry (SIC code 6).

The equity valuation measure, i.e., the empirical counterpart for the price p_{jt} in the model, is market capitalization over total assets for the calendar year. Our cash-flow variable, d_{jt}, is proxied using total dividends paid over assets.

We make a couple of adjustments to the raw data. The first is to deal with inflation, which can create predictability in nominal dividends and prices. We adjust all cash-flow variables with a GDP deflator, deflating all nominal values to 2010 USD values. The second pertains to exiting firms. Our preferred solution is to only consider periods during which a firm has non-missing information. Next, we winsorize the deflated values for assets, market capitalization and total dividends at 0.01% level.

Henceforth, we refer to the market capitalization at the end of year for stock j divided by the assets in that year for stock j as the price p_{jt}, and the total dividends normalized by assets in that year as d_{jt}. We calculate the excess returns as $R_{jt} = \left(p_{jt+1} + d_{jt+1} - p_{jt}\right)/p_{jt} - r_t^f$, where we use the yield on Treasury bills (constant maturity rate, hereafter CMT) with one year maturity as the risk-free rate.

Forming Asset Portfolios The procedure described above can be used for any number and type of assets, including individual stocks. However, for expositional purposes, and to show more clearly the patterns in data value, we group assets into a small number of commonly-used portfolios, rather than work with a large number of individual stocks/assets.

This leaves us with a more manageable number of data values to compute and compare.

Our first two portfolios are based on size. We group firms into Large and Small, based on whether their market capitalization is above or below the median value for all firms in our sample, in a given year. Next, we construct Growth and Value portfolios, using the book-to-market ratio (defined as the difference between total assets and long-term debt, divided by the firm's market capitalization). Firms above the median value of book-to-market in a year are assigned to the Value portfolio, while those below the median are part of the Growth portfolio. In addition to these four portfolios – Small, Large, Growth and Value – we also include a market index (specifically, the S&P500) as a portfolio. We use value-weighted averages for excess returns for each portfolio as the return measure, where we weigh each firm's return by its market capitalization.

Measuring Price Impact In order to use our formulae for data value, we need a estimate for price impact. In practice, an investor who wants to use our data valuation framework, (s)he should use the price impact applicable to his/her context. For our purposes here, we will use an average estimate from the literature. Our starting point is the estimate from Hasbrouck (1991), who finds that a \$20,000 trade moved prices by 0.3% on average. Using a reference price per share of one, a 0.3% price increase corresponds to a price that is 0.003 units higher. Therefore, we explore imperfectly competitive markets where $dp_j/dq_j = 0.003/20000 = 1.5 \times 10^{-7}$. While this might seem like a small number, we will see that it has substantial impact on data valuations.[6]

Data Timing As discussed above, our return measure for year t for an asset j is the cum-dividend excess return on that asset over the year t – using prices at the end of year t and at the end of year $t-1$, along with dividends paid out over year t. We are interested

[6]Data limitations force us to make the following simplifying assumptions: (i) price impact is the same for all portfolios we analyze (ii) trading in one asset portfolio can only move the prices of that portfolio. Thus, the matrix dp/dq for Equation (12) is λI, where $\lambda = 1.5 \times 10^{-7}$ is the price impact magnitude as listed in the text, and I is the identity matrix.

in understanding the value of data available to an investor *before* year t, in predicting the value of this return measure for year t.

The value of any control variable in Z_t used for the purpose of this calculation is obtained for year $t-1$, since these values will be in the investor's information set while predicting the profits for year t. Similarly, the data signal in X_t that we are valuing needs to be in the information set of the investor *before* year t. To predict profits over year t, we use the data signals which are produced *before* year t starts, which give information about growth in earnings of firms between year $t-1$ and year t.

Our toolkit can used to value any finance-relevant data stream or bundle of data streams. In the rest of the paper, we show how it can be used to value two different data streams. The first, discussed in the following section, values earnings forecast data put out by stock analysts. We discuss how the value varies with investor heterogeneity along various dimensions and market conditions. The second, in Section 4, estimates the value of a hypothetical data source that allows investors to perfectly forecast GDP. Before turning to that analysis, we describe the two data sources of interest in more detail.

The Financial Data Stream We Value: I/B/E/S Forecasts The data series of interest in our first exercise is earnings forecasts provided by the Institutional Brokers' Estimate System (I/B/E/S). We use earnings forecasts for 5,506 unique firms from 1985–2015, with 1,018 firm observations per year on average.[7]

We use annual earnings forecasts from I/B/E/S. In our baseline model, investors have a horizon of a year and use the latest available one-year-ahead earnings forecast at each date. Later, we explore how different trading horizons affect the data value.

The Macro Data Stream We Value: *Ex-post* GDP Growth For realized GDP growth, we use the second release estimates of quarterly GDP growth from BEA, as reported

[7]We use the Summary Statistics series from I/B/E/S, accessed through WRDS, `https://wrds-www.wharton.upenn.edu/pages/get-data/ibes-thomson-reuters/ibes-academic/summary-history/summary-statistics/`.

by the Federal Reserve Bank of Philadelphia[8].

3 Valuing Financial Data

In this section, we first estimate the utility gain that investors would assign to I/B/E/S forecasts, given what they already know, and then convert this into a dollar amount. The latter is the monetary value of I/B/E/S data, or equivalently investors' willingness to pay for this data. In most cases, these private valuations look nothing like a price that any investor actually pays for an I/B/E/S subscription. Some valuations are orders of magnitude higher, others much lower. Recall that these are not predicted transactions prices. They are private valuations that trace out a demand curve. The qualitative patterns are mostly intuitive, which suggests that our measurement strategy/toolkit is a sensible one.

When we value a stream of data, we need to take a stand on what else an investor already knows, i.e. the publicly available information. Obviously, as econometricians, we do not observe information sets directly, so in our implementation, cannot control for this perfectly. Of course, this is not a problem for a practitioner or investor who wishes to use our toolkit to value a stream of data (e.g. one that she is considering buying), since she would know exactly what other data she already has access to. For the purposes of illustrating the use of the tool, we endow our hypothetical investor with some commonly-used and publicly-available data series. Specifically, we assume that they already observe the dividend yield (D/P ratio) for S&P500[9].

In additional results, we also consider and investor who also has access to one or more of the following pieces of data: the yield on a 1-year Treasury bill (constant maturity rate)[10], the consumption-wealth ratio (CAY) from Lettau and Ludvigson (2001) and a sentiment index from Baker and Wurgler (2006).

[8] `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/routput`

[9] Obtained from NASDAQ Quandl `https://data.nasdaq.com/data/MULTPL/SP500_DIV_YIELD_MONTH-sp-500-dividend-yield-by-month`

[10] Obtained from FRED series `DGS1`

We use annual earnings forecasts from I/B/E/S as the first data stream that we value. In our baseline model, investors have a horizon of a year and use the latest available one-year-ahead earnings forecast at each date. Later, we explore how different trading horizons affect the data value.

For each firm, we use the median consensus analyst forecast for earnings per share (hereafter EPS). We discard all forecast values which have been calculated during or after the calendar year for which the forecast is being made. For example, any forecast we use for earnings in 2015 has to be issued before the year 2015 starts. We then drop all but the latest consensus forecasts for each firm-year observation, which gives us a single consensus forecast for EPS over the next year. Using this forecast, we calculate a forecasted growth rate: the forecasted EPS for the coming year, divided by the realized value of EPS from the last year.

Our goal is to explore data valuation patterns, to gain intuition for how large this amount is and what makes it vary. Therefore, in order to keep the analysis manageable, we collapse the large number of assets into a few portfolios. Specifically, we analyze five portfolios: Small, Large, Growth, and Value firms, as well as the S&P500 index. We find that most of the value of the I/B/E/S data comes from signals about growth firms and those in the S&P500 index.

Therefore, the data value numbers we report in this section for two annual signals, one about earnings of all firms in the Growth portfolio and one about the earnings of all firms in the S&P500 index. Specifically, these are the portfolio value-weighted average values of median forecasted growth rates for earnings per share – for the Growth and S&P500 portfolios. Note that we are valuing a forecast of a payoff of a particular portfolio of assets. [11]

[11] We could have performed this calculation under many alternative assumptions. For example, one could value growth firms' data from the perspective of an investor who invests only in growth firms. In that case, one would regress the growth firm asset payoffs on the relevant data and use means variances and forecast errors of growth asset payoffs. We did not take that approach because if we vary the investment set and the data together, we would not know whether data was more/less valuable because of the data or the investment restriction. But, it is certainly another dimension of investor heterogeneity that might be interesting to explore.

3.1 Wealth and Risk Tolerances

One obvious dimension along which investors differ is the size of their portfolios. We consider investors with two wealth levels— $\$500,000$, and $\$250$ million, each with the same relative risk aversion of $\sigma = 2$. In terms of the wealth level, the former group is similar in magnitude to the wealth level of the mean US household (Badarinza, Campbell, and Ramadorai, 2016), while the latter investor group has comparable wealth level to the size of the mean US hedge fund (Yin, 2016). The resulting difference in absolute risk aversion give rise to different willingness to pay for the same data.

To value data for a particular investor, we need to know what else they already know and what they can invest in. The investor whose value we are calculating already knows the previous year's S&P500 dividend/price ratio. They can invest in any combination of the following five portfolios: S&P500, Small, Large, Growth, and Value. However, we make no assumption about what any other investors know or trade.

Table 1 reports the dollar value of the I/B/E/S forecasts for two investors with different wealth levels, with and without price impact. The results illustrate show that wealthier investors attach a higher dollar value to the same data. In our setting, this occurs through the dependence of the curvature parameter ρ on wealth. Under our calibration, an investor with $\$250$ million in wealth operating in a competitive setting would be willing pay almost 300 times more for this data compared to one with half a million dollars of wealth.

Next, as one would expect, price impact attenuates the value of data. This effect is quite significant, even for the relatively modest levels of price impact in our calibration and increases with wealth. This is especially true for wealthier investors: taking price impact into account cuts the value of the I/B/E/S data for an investor with $\$250$ million in wealth by almost 80%. To see why, recall that in Lemma 2, price impact (dp/dq) gets scaled by $1/\rho_i$. Since wealthier investors are assumed to have a lower degree of absolute risk aversion (a lower ρ_i), price impact has a disproportionate effect on their payoffs and data valuations.

To better understand the sources of data value, Table 1 also reports the expected return

Table 1: **Risk Tolerance.** Annual data between 1985–2015. The dependent variables in (13) and (14) is the vector of returns, in excess of a 1-year treasury (CMT), for five portfolios – {Small, Large, Growth, Value, S&P500}. All specifications include a constant and a control variable (the S&P500 D/P ratio). Data variables being valued are the I/B/E/S median forecasts for annual value-weighted earnings for Growth and S&P500 portfolios, normalized by assets and growth over last year's realized earnings for each ticker. The case with price impact assumes Kyle's Lambda $\lambda = \frac{dp}{dq} = 1.5 \times 10^{-7}$. Dollar values are reported in thousands of 2010 USD.

	Perfect Competition	With Price Impact
Panel A: Investor with $500,000 Wealth.		
Utility Gain	0.0919	0.0363
Expected Profit	0.0365	0.0084
Variance Reduction	0.0554	0.0279
Dollar Value (in $000)	3.50	1.38
Time Periods	31	31
Panel B: Investor with $250m Wealth.		
Utility Gain	0.0919	0.0196
Expected Profit	0.0365	0.0029
Variance Reduction	0.0554	0.0167
Dollar Value (in $000)	1188.50	253.62
Time Periods	31	31

and the variance reduction on the investor's portfolio. The expected profit is the ex-ante expected return on the optimal, diversified portfolio of the five assets the investor can hold. The variance reduction is the difference between the raw variance of this return and the conditional variance, which is the average squared residual of the predicted return, after conditioning on the data. This is a measure of how much one learns from data. Notice that price impact lowers both components of data value and has a more pronounced effect when wealth is higher (or equivalently, absolute risk aversion is lower).

3.2 Liquidity Affects Data Value

A consistent theme throughout our results is the importance of price impact. For expositional purposes, we have treated price impact as a single, time-invariant number. In reality, it

fluctuates with market liquidity. Our estimates suggest that such fluctuations will have a dramatic impact of the value of data, especially for large investors.

Now consider a financial firm whose business model revolves around the use or sale of data. That firm's market value is based largely on the value of their data. Changes in market liquidity will thus affect the real value of this firm's data assets through this channel.

As firms' data stocks grow larger, the magnitude of liquidity shocks to data values should grow. The reason is that price impact enters additively with conditional variance. This additive form comes from first order condition for the optimal portfolio choice of investor i: $q_{it} = (\rho_i \mathbb{V}[p_{t+1} + d_{t+1}|\mathcal{I}_{it}] + dp/dq_i)^{-1} (\mathbb{E}[p_{t+1} + d_{t+1}|\mathcal{I}_{it}] - rp_t)$. If the conditional variance $\mathbb{V}[p_{t+1} + d_{t+1}|\mathcal{I}_{it}]$ is large (high uncertainty), then small changes in price impact dp/dq_i have little effect. Those changes are swamped by the variance term and the inverse of this large number is small. However, if conditional variance is small, meaning that asset payoff forecasts are precise, then that first term, the inverse of a potentially small number, may be large. In this case, the effects of price impact can be substantial. Over time, if firms have more data and thus smaller forecast errors, their data valuations become more and more susceptible to changes in the price impact of a trade.

The high and growing sensitivity of data value to market liquidity suggests a new channel through which market liquidity matters. Since the value of a financial firm depends on its ability to trade profitably, the value of data is an input into the valuation of a financial firm. As financial firms become more data-centric, the firm's value becomes more sensitive to the value of its data. At the same time, growing data abundance makes the value of data more sensitive to market liquidity. These two margins of increasing sensitivity amplify each other. This suggests that changes in market liquidity may affect the real value and the equity value of financial firms through a new channel, through the value of their data. In a world in which data is becoming increasingly abundant, this new liquidity-data effect could grow much stronger. These findings suggest that, because of the rising abundance and importance of data for financial firms, market liquidity may become more important than

ever before.

3.3 Investment Styles

Another dimension along which investors differ is their investment style. To understand the implications of this type of heterogeneity for data valuation, we value exactly the same data as before, the median earnings growth forecasts from I/B/E/S, from the perspective of investors who only trade in individual portfolios. We will refer to these investors by the portfolios they trade. For example, the Small investor is one who only buys and sells the portfolio of small stocks that we constructed. Same for the Large, Growth, Value and S&P investors. They each use the earnings forecast data to determine how much to trade in their respective portfolios. We compare these data values to the value of the investor who trades in all five portfolios, which corresponds to the case analyzed in Table 1.

Table 2 shows that among the investors who invest in a single portfolio, I/B/E/S forecast data is most valuable for investors in Growth, Large or the S&P500 portfolios. While the investor wealth and price impact raise and lower the dollar value of the data, respectively, this pattern of growth and large or S&P500 investors valuing earnings forecast data by more emerges consistently. This is because the I/B/E/S data lacks relevance for the Small portfolio. More precisely, it does little to reduce return forecast errors and in that sense, provide little guidance to an investor about when to buy and sell the Small portfolio. So, despite the high unconditional expected returns of the Small portfolio, the value of this particular data stream for such investors is quite low. The relevance of the I/B/E/S forecasts is low for the Value portfolio as well. The Large and Growth portfolios on the other hand have medium expected returns, but their returns are predicated to a larger degree by the analyst forecast data. We can see that in the difference between $\mathbb{V}[R]$ with and without data, in rows 3 and 4. Therefore, this data is most valuable to those who invest in growth and large-firm equities.

As we saw in the previous set of results, price impact reduces the value of data, but

Table 2: **Investment Styles.** Annual data between 1985–2015. Dependent variables in (13) and (14) are returns, in excess of a 1-year treasury (CMT), for five portfolios – {Small, Large, Growth, Value, S&P500}. All specifications include a constant and the S&P500 D/P ratio as a control. Data variables being valued are the I/B/E/S median forecasts for annual value-weighted earnings growth for Growth and S&P500 portfolios, normalized by assets (growth is computed over last year's realized earnings for each portfolio). Panel B (estimates with price impact) assumes Kyle's Lambda $\lambda = \frac{dp}{dq} = 1.5 \times 10^{-7}$. Dollar values are reported in thousands of 2010 USD.

	Portfolio Type					
	Small	Large	Growth	Value	S&P500	All
Panel A: Perfect Competition						
$\mathbb{E}[R]$	0.2058	0.0802	0.1047	0.0273	0.0350	–
$\mathbb{V}[R]$	0.1333	0.0223	0.0255	0.0269	0.0144	–
$\mathbb{V}[R]$ (controls)	0.1371	0.0231	0.0263	0.0270	0.0145	–
$\mathbb{V}[R]$ (controls+data)	0.1375	0.0215	0.0241	0.0264	0.0133	–
Utility Gain	0.0000	0.0438	0.0653	0.0127	0.0498	0.0919
Dollar Value (in $000) for:						
Investor with $500,000 Wealth	0.00	1.67	2.49	0.49	1.90	3.50
Investor with $250m Wealth	0.00	566.41	844.09	164.71	643.62	1188.50
Time Periods	31	31	31	31	31	31
Panel B: With Price Impact						
Investor with $500,000 Wealth:						
Utility Gain	0.0000	0.0433	0.0650	0.0107	0.0480	0.0363
Dollar Value (in $000)	0.00	1.65	2.48	0.41	1.83	1.38
Investor with $250m Wealth:						
Utility Gain	0.0000	0.0019	0.0044	0.0001	0.0012	0.0196
Dollar Value (in $000)	0.00	23.93	57.00	1.45	15.98	253.62
Time Periods	31	31	31	31	31	31

also reduces the dispersion in valuations. The investors who value data most are the same investors who would liek to trade aggressively on the data, but are prevented from doing so when price impact is large.

3.4 Previously Purchased Data

A third dimension along which investors differ enormously is in the data they already own. While large, institutional investors have access to enormous libraries of data, households may

Table 3: **Previously Purchased Data.** Annual data between 1985–2015. Dependent variables in (13) and (14) are excess returns (over CMT-1yr) for five portfolios – {Small, Large, Growth, Value, S&P500}. All specifications include a constant. Data variables being valued are the I/B/E/S median forecasts for annual value-weighted earnings for Growth and S&P500 portfolios, normalized by assets and growth over last year's realized earnings for each ticker. Values in each column represent the additional value of I/B/E/S data *on top of* the control variable(s) listed in the header. Higher value for data indicates I/B/E/S data adds more value over the control variable. Dollar values are reported in thousands of 2010 USD.

	No Controls	Real CMT-1yr	BW Sentiment	cay	S&P500 D/P ratio	All Controls
Utility Gain	0.163	0.147	0.145	0.104	0.092	0.073
Expected Profit	0.065	0.066	0.060	0.045	0.037	0.046
Variance Reduction	0.098	0.080	0.086	0.059	0.055	0.027
Time Periods	31	31	31	31	31	31
Dollar Value (in $000) for:						
Investor with $500,000 Wealth	6.21	5.60	5.54	3.96	3.50	2.80
Investor with $250m Wealth	2106.62	1900.07	1880.90	1343.14	1188.50	949.75

know only a few summary statistics about each asset. We illustrate both how to incorporate differences in existing data sets and their quantitative importance through a simple exercise. So far, we have valued the I/B/E/S data assuming that investors already have access to S&P500 dividend/price ratio. In this set of results, we ask: How valuable would the same I/B/E/S forecasts be if, instead of the S&P500 dividend/price ratio, the investor had some other variable in his or her existing data set? Of course, that does not nearly capture the extent of the difference between the knowledge of investors. But even these minor differences in which macro variable the investor already knows can significantly change the value of a new data stream.

In Table 3, the first column reports the value of the I/B/E/S forecasts to the two classes of investors we consider —an investor with a wealth level of $500,000, and an investor with $250m wealth—who have no other sources of information. The next four columns report the value of data when investors already have access to a single prior data series: Real CMT-1y, BW Sentiment, CAY, and S&P500 D/P ratio, respectively. In the last column, the investor already has access to all five of these data series.

Unsurprisingly, access to prior data decreases the value of I/B/E/S data for investors. The I/B/E/S forecasts are more than twice as valuable to the investor who knows nothing, relative to the investor who already knows all five series. This is just an illustration of the diminishing marginal returns to data. However, Table 3 shows that value of the I/B/E/S data is relatively insensitive to knowledge of Real CMT-1y and BW Sentiment data. This insensitivity means that I/B/E/S contains information that is not highly correlated with the information in either series. On the other hand, data about CAY and S&P500 D/P ratio are closer substitutes for the I/B/E/S data and attenuate the value of the latter more visibly.

Among the alternative pieces of data that the investors can use, S&P500 D/P ratio is by far the most informative one. The additional I/B/E/S data has the lowest value to investors who already have access to S&P500 D/P ratio. Furthermore, for investors who have S&P500 D/P ratio prior data, access to the rest of the macroeconomic data series does not attenuate the value of the I/B/E/S data much more.

3.5 Trading Horizon

Finally, investors differ in their trading horizons. Our data valuation tool can be applied to various trading horizons. However, for the data we are exploring, this dimension of investor heterogeneity seems to matter less than the others.

Our calculations so far have assumed that investors trade over an annual horizon. Next, we measure the value of the same data – the median I/B/E/S forecast – for an investor who trades the same portfolio but with a quarterly horizon. This does not change the data value formula; it does change how we implement it. The procedure is to compute residuals from (13) and (14) where R_t is quarterly return, the prior information Z_t is a constant and quarterly dividend-price ratios, and where X_t is the median forecast of the earnings growth for Growth and S&P500 portfolios over the year.[12] The resulting regression residuals (ε_t^{XZ} and ε_t^{Z}) are then used to construct the variance matrices and substitute these variances,

[12]We also re-did the estimation using forecasts of quarterly earnings growth. It produced similar, but somewhat smaller, data value estimates.

Table 4: **Trading Horizon.** Data between 1985–2015. Dependent variables in (13) and (14) are returns, in excess of a 1-year treasury (CMT), for five portfolios – {Small, Large, Growth, Value, S&P500}. All specifications include a constant and the S&P500 D/P ratio. Data variables (X_t in (13)) are the I/B/E/S median forecasts, in growth rates, for annual value-weighted earnings for Growth and S&P500 portfolios, normalized by assets. Numbers reported in each column represent the additional value of annual I/B/E/S data (9) on top of the control variable (S&P500 D/P ratio) for an investor trading at the trading horizon listed in the table header. Dollar values are reported in annualized thousands of 2010 USD, and utility gain numbers are annualized.

	Annual	Quarterly
Utility Gain (ann.)	0.092	0.067
Dollar Value (in $000, ann.) for Investor with $500,000 Wealth	3.50	2.54
Dollar Value (in $000, ann.) for Investor with $250m Wealth	1188.50	862.46
Time Periods	31	124

along with expected quarterly returns, into the expected utility formula (7). We convert expected utility to data value as before, using (9).

The expected asset payoff and its variance will typically be smaller for shorter horizons. This just reflects the fact that there is less asset appreciation and smaller changes over a shorter period of time. The utility of an equally precise forecast is smaller because that information will be used only for a lower potential payoff. Therefore, in order to facilitate comparison with our baseline annual horizon numbers, we annualize our estimated quarterly horizon data values by multiplying them by four.

Table 4 reports the value of the I/B/E/S forecasts for both annual and quarterly investors. The first column is the same values reported in Table 1. The second column shows that investors who trade more frequently, on a quarterly basis, would be less willing to pay for data each year. The reason for the lower quarterly valuation is that quarterly returns are considerably more noisy. Earnings data is not very useful for quarterly portfolio adjustment. Trading on this data only creates more noise.

The effect of trading horizon surely depends on the data source. For example, high-frequency data is useful for high-frequency traders, but will likely be worthless after a year.

Table 5: **Macroeconomic Information.** Quarterly data between 1985–2015. Dependent variables in (13) and (14) are returns, in excess of a 1-year treasury (CMT), for five portfolios – {Small, Large, Growth, Value, S&P500}. All specifications include a constant and controls (the S&P500 D/P ratio, the realized real GDP growth in the previous quarter and the median forecasted growth rate in real GDP from the Survey of Professional Forecasters). Data variables (X_t in (13)) are the second release estimates of real quarterly GDP numbers as reported by by the BEA, expressed as growth rates over the previous quarter. Numbers reported in each column represent the additional value of *ex-post* real GDP growth data (9) on top of the control variables for an investor trading at the quarterly horizon. The case with price impact assumes Kyle's Lambda $\lambda = \frac{dp}{dq} = \frac{1}{4} \times 1.5 \times 10^{-7}$. Dollar values are reported in annualized thousands of 2010 USD, and utility gain numbers are annualized.

	Portfolio Type					
	Small	Large	Growth	Value	SP500	All
Panel A: Perfect Competition						
Utility Gain (ann.)	0.1058	0.1073	0.0770	0.0975	0.0480	0.1369
Dollar Value (in $000, ann.) for:						
Investor with $500,000 Wealth	4.03	4.09	2.93	3.71	1.83	5.22
Investor with $250m Wealth	1367.65	1387.57	995.22	1260.04	620.80	1769.76
Panel B: With Price Impact						
Investor with $500,000 Wealth:						
Utility Gain (ann.)	0.1057	0.1067	0.0769	0.0714	0.0468	0.0936
Dollar Value (in $000, ann.)	4.03	4.06	2.93	2.72	1.78	3.57
Investor with $250m Wealth:						
Utility Gain (ann.)	0.0156	0.0069	0.0130	0.0005	0.0015	0.0703
Dollar Value (in $000, ann.)	201.35	89.62	167.54	6.90	19.14	909.20
Time Periods	123	123	123	123	123	123

The more important take-away is that trading horizon can matter for how an investor values their data. By adjusting the input data and the interpretation of the results, our data valuation tool can be used to value data used by investors who trade at various frequencies.

4 Valuing Macroeconomic Information

How do different investors value information about macroeconomic variables (e.g. GDP)? We now use our framework to provide an answer to this question. In Table 5, we compute the value of a hypothetical data source which allows investors to perfectly forecast GDP.

Formally, we use the realized (i.e. *ex-post*) real GDP growth as our data series of interest[13] and calculate its value to investors with different trading styles, defined in Section 3.3. We control for the S&P500 dividend-price ratio, as before, as well as two additional controls—the realized real GDP growth rate in the previous quarter and the median forecasted growth rate in real GDP for the current quarter by the Survey of Professional Forecasters (SPF)[14]. Adding these two additional control variables allows us to find the value of the new information in *ex-post* GDP growth.

The last column of Table 5 shows that a fund with assets of $250 million, trading all five portfolios under perfect competition, would be willing to pay $1.77 million for the ability to perfectly forecast quarterly GDP growth in advance. The other columns show values for more restricted trading styles. They are all sizable, albeit with some variation. There are some interesting cross-sectional differences relative to the value of earnings forecasts analyzed in Table 2. For example, better information about GDP is quite valuable for investors trading only the Small portfolio. This is because, unless the earnings forecasts, GDP growth turns out to be a valuable predictor of returns on the Small portfolio, i.e. this data has high relevance for such an investor. In fact, macroeconomic information of this form shows relatively high data relevance for all five assets – unlike the earnings forecasts data, which showed high relevance mostly for Large and Growth portfolios.

The bottom panel shows the value of data with price impact.[15] As with the earnings forecast data, price impact significantly attenuates the value of macroeconomic information as well and the effects are more pronounced for wealthier, less risk-averse investors. The value of a perfect GDP forecast for the aforementioned $250 million fund trading all five portfolios is cut almost in half once price impact is taken into account. The drop in valuations is even more significant for some of the individual portfolios, again underscoring the importance of market liquidity for the value of data.

[13]We use the second release revised estimates of realized real GDP growth for this calculation.
[14]`https://www.philadelphiafed.org/surveys-and-data/rgdp`
[15]These calculations use Kyle's $\lambda = 0.375 \times 10^{-7}$.

5 Conclusion

Data is one of the most valuable assets in the modern economy. Yet the tools we have to quantify that value are scant. We offer a tool that an investor or financial firm can use to value its existing data, or a potential stream of data that it is considering to acquire. Along with information about the distribution of investor characteristics, researchers can use this tool to trade out the demand curve for data.

We uncover important investor wealth and trading style effects, the importance of an investor's existing data, and the role of trading horizon. Jointly, these effects point toward enormous heterogeneity, spanning multiple orders of magnitude, in the value different investors assign to the same data. The dispersion in valuations suggests that marginal changes in the price of data will have little effect on demand. With such dispersed valuations, few data customers would be on the margin. This low price elasticity of demand is significant because it points to one reason why data markets might not evolve to be very competitive.

We further uncover a new channel through which market liquidity matters for the real value of data, which is an important new class of assets. As firms accumulate more data and data technologies improve, more and more of the value of a financial firm will depends on the value of the data it possess. The sensitivity of the value of data to price impact of a trade could introduce a new source of financial fragility, brought on by data accumulation, and exacerbated by data technologies that improve financial forecasting.

The advantage of our measurement tool is its simplicity. While our measure of the value of data is derived from a structural model, computing it does not require estimating structural parameters. Instead, the relevant sufficient statistics are simple means and variances of linear regression residuals. No matter whether the data is public, private, or known only to a fraction of investors, these methods are valid. Even if the data is about sentiments or order flows, as long as it is measured along with the market prices in the observable data set, our data value measure offers a meaningful assessment of its value to an investor.

References

Albagli, E., C. Hellwig, and A. Tsyvinski (2014): "Risk-Taking, Rent-Seeking, and Investment When Financial Markets Are Noisy," Yale Working Paper. 6

Amador, M., and P.-O. Weill (2010): "Learning from prices: Public communication and welfare," *Journal of Political Economy*, forthcoming. 6

Badarinza, C., J. Y. Campbell, and T. Ramadorai (2016): "International Comparative Household Finance," *Annual Review of Economics*, 8(1), 111–144. 25

Bai, J., T. Philippon, and A. Savov (2016): "Have Financial Markets Become More Informative?," *Journal of Financial Economics*, 122 (35), 625–654. 7

Baker, M., and J. Wurgler (2006): "Investor sentiment and the cross-section of stock returns," *Journal of Finance*, 61 (4), 1645–1680. 23

Barlevy, G., and P. Veronesi (2000): "Information Acquisition in Financial Markets," *Review of Economic Studies*, 67(1), 79–90. 7

Bergemann, D., A. Bonatti, and A. Smolin (2016): "The Design and Price of Information," CEPR Discussion Papers 11412, C.E.P.R. Discussion Papers. 7

Davila, E., and C. Parlatore (2021): "Identifying Price Informativeness," NYU Working Paper. 7

Dow, J., I. Goldstein, and A. Guembel (2017): "Incentives for Information Production in Markets where Prices Affect Real Investment," *Journal of the European Economic Association*, 15(4), 877–909. 6

Farboodi, M., A. Matray, L. Veldkamp, and V. Venkateswaran (2019): "Where Has All the Data Gone?," Working Paper. 7

Farboodi, M., and L. Veldkamp (2017): "Long Run Growth of Financial Technology," Working Paper, Princeton University. 7, 49

——— (2020): "Long-run Growth of Financial Data Technology," Discussion Paper 8. 17

Glode, V., R. Green, and R. Lowery (2012): "Financial Expertise as an Arms Race," *Journal of Finance*, 67(5), 1723–1759. 15

Goldstein, I., E. Ozdenoren, and K. Yuan (2013): "Trading frenzies and their impact on real investment," *Journal of Financial Economics*, 109(2), 566–82. 7

Grossman, S., and J. Stiglitz (1980): "On the impossibility of informationally efficient markets," *American Economic Review*, 70(3), 393–408. 7, 14

Hasbrouck, J. (1991): "Measuring the Information Content of Stock Trades," *The Journal of Finance*, 46(1), 179–207. 21

He, Z. (2009): "The Sale of Multiple Assets with Private Information," *Review of Financial Studies*, 22, 4787–4820. 7

Kacperczyk, M., J. Nosal, and S. Sundaresan (2021): "Market Power and Informational Efficiency," Working Paper, Imperial College London. 2, 6, 7

Kadan, O., and A. Manela (2019): "Estimating the Value of Information," *Review of Financial Studies*, 32 (3), 951–990. 6

Koijen, R. S. J., and M. Yogo (2019): "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 127(4), 1475 – 1515. 9

Kondor, P. (2012): "The more we know about the fundamental, the less we agree," *Review of Economic Studies*, 79(3), 1175–1207. 7

Kurlat, P., and L. Veldkamp (2015): "Should we regulate financial information?," *Journal of Economic Theory*, 158, 697–720. 10

Kyle, A., and J. Lee (2017): "Toward a Fully Continuous Exchange," SSRN Working Paper. 6

Kyle, A. S. (1989): "Informed Speculation with Imperfect Competition," *Review of Economic Studies*, 56(3), 317–355. 6

Lettau, M., and S. Ludvigson (2001): "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," *Journal of Political Economy*, 109(6), 1238–1287. 23

Mondria, J. (2010): "Portfolio choice, attention allocation, and price comovement," *Journal of Economic Theory*, 145, 1837–1864. 7

Morris, S., and H. S. Shin (2002): "Social value of public information," *The American Economic Review*, 92(5), 1521–1534. 6

Ozdenoren, E., and K. Yuan (2008): "Feedback Effects and Asset Prices," *The Journal of Finance*, 63(4), 1939–1975. 6

Peress, J. (2004): "Wealth, information acquisition and portfolio choice," *The Review of Financial Studies*, 17(3), 879–914. 7

Savov, A. (2014): "The price of skill: Performance evaluation by households," *Journal of Financial Economics*, 112(2), 213–231. 6

Sockin, M. (2015): "Not So Great Expectations: A Model of Growth and Informational Frictions," Princeton Working Paper. 7

Yin, C. (2016): "The Optimal Size of Hedge Funds: Conflict between Investors and Fund Managers," *The Journal of Finance*, 71(4), 1857–1894. 25

Appendix

A Model Solution

Portfolio Choice Since we have a linear Gaussian system, we conjecture an equilibrium price which is linear in the aggregate shocks,

$$p_t = A_t + B(d_t - \mu) + C_t y_{t+1} + D_t x_{t+1} + F_t z_{t+1} \tag{15}$$

Assuming price of the form given in Equation (15), the investor derives an unbiased signal η_{pt} of y_{t+1} from the price as,

$$\eta_{pt} \equiv C_t^{-1}\left(p_t - A_t - B(d_t - \mu)\right) = y_{t+1} + C_t^{-1} D_t x_{t+1} + C_t^{-1} F_t z_{t+1}$$

This price signal has the conditional variance,

$$V\left(\eta_{pt} \mid \mathcal{I}_{it}\right) \equiv \Sigma_{pt} = C_t^{-1} D_t \Sigma_x D_t' C_t^{-1\prime} + C_t^{-1} F_t \Sigma_z F_t' C_t^{-1\prime}$$

Note that the variance of this price signal is a fixed quantity (since the coefficients are artifacts of the model, known ex ante to all investors). Given the information set $\mathcal{I}_{it}$, the investors update their beliefs of the dividend innovation y_{t+1} as per Bayesian updating to get,

$$\begin{aligned}
\mathbb{E}\left[y_{t+1} \mid \mathcal{I}_{it}\right] \equiv \mu_{it} &= \Sigma_{it}\left(\Sigma_d^{-1} \times 0 + \Sigma_{pt}^{-1}\eta_{pt} + (\zeta_{it}^2\Sigma_z + K_{it}^{-1})^{-1} s_{it}\right) \\
&= \Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1} s_{it}\right) \\
\mathbb{V}\left[y_{t+1} \mid \mathcal{I}_{it}\right] \equiv \Sigma_{it} &= \left\{\Sigma_d^{-1} + \Sigma_{pt}^{-1} + (\zeta_{it}^2\Sigma_z + K_{it}^{-1})^{-1}\right\}^{-1}
\end{aligned}$$

Further, we can express the gross payout at the end of period $t+1$ as,

$$\begin{aligned}
p_{t+1} + d_{t+1} &= A_{t+1} + B(d_{t+1} - \mu) + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2} + d_{t+1} \\
&= A_{t+1} + \mu + (B + I)\left(d_{t+1} - \mu\right) + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2} \\
&= A_{t+1} + \mu + (B + I)\left[G(d_t - \mu) + y_{t+1}\right] + C_{t+1} y_{t+2} + D_{t+1} x_{t+2} + F_{t+1} z_{t+2}
\end{aligned}$$

Hence, the conditional moments of the gross payout can be expressed as,

$$\begin{aligned}
\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] &= A_{t+1} + \mu + (B + I)G(d_t - \mu) + (B + I)\mu_{it} \\
\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] &= (B + I)\Sigma_{it}(B + I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'
\end{aligned}$$

We first note that the shocks y_{t+2}, x_{t+2} and z_{t+2} do not contribute towards the conditional expectation, but are driving the conditional variance of the gross payout. On the other hand, investors

form imprecise estimate for the end-of-period shock y_{t+1}, resulting in a contribution in both the conditional moments.

In the perfect competition equilibrium (as per Lemma 1), investor i selects the optimal portfolio q_{it} given by the first order condition

$$q_{it} = \frac{1}{\rho_i}\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right]^{-1}\left\{\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] - rp_t\right\}.$$

Hence, the optimal portfolio is given as,

$$q_{it} = \frac{1}{\rho_i}\left\{(B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right\}^{-1} \times \left[\underbrace{A_{t+1} + \mu + (B+I)G(d_t - \mu) - rp_t}_{\star} + \underbrace{(B+I)\mu_{it}}_{\dagger}\right] \tag{16}$$

Market Clearing We now impose market clearing, $\int_i q_{it}di = \bar{x} + x_{t+1}$. First, note that the terms marked by $\star$ in Equation (16) are constants for the integration. Hence, we define the factor multiplying these terms – the risk tolerance weighted average precision of the gross payout,

$$\Omega_t \equiv \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di$$

We next simplify the remaining term marked by $\dagger$ in the integration in Equation (16) as,

$$\begin{aligned}
&\int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\mu_{it}di \\
&= \int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1} s_{it}\right)di \\
&= \left\{\int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}di\right\}\Sigma_{pt}^{-1}\eta_{pt} \\
&+ \int_i \rho_i^{-1} V\left(p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right)^{-1}(B+I)\Sigma_{it}\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}(y_{t+1} + \zeta_{it}z_{t+1} + \xi_{it})di \\
&= \Gamma_t\Sigma_{pt}^{-1}\eta_{pt} + \Phi_t y_{t+1} + \Psi_t z_{t+1}
\end{aligned}$$

Here, we used the fact that ξ_{it} is distributed independently of all other variables with mean zero, and defined the additional covariance terms Γ_t, Φ_t and Ψ_t (with Ω_t duplicated for reference) as,

$$\begin{aligned}
\Omega_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di \\
\Gamma_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \underbrace{(B+I)\Sigma_{it}} di \\
\Phi_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it} \underbrace{\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}} di \\
\Psi_t &\equiv \int_i \rho_i^{-1} \left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it} \left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1} \underbrace{\zeta_{it}} di
\end{aligned}$$

As noted before, Ω_t is the risk tolerance weighted average precision of the gross payout. The terms highlighted with $\underbrace{\quad}$ indicate the additional terms in each subsequent covariance term. First, Γ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1}. Similarly, Φ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1} and the signal precision $\left(\zeta_{it}^2\Sigma_z + K_{it}^{-1}\right)^{-1}$. Lastly, Ψ_t is the covariance of the gross payout precision with the posterior variance of the dividend shock y_{t+1}, the signal precision and the exposure to the public signal ζ_{it}.

We can now subsitute the covariance terms Ω_t, Γ_t, Φ_t, Ψ_t and the price signal $\eta_{pt} = C_t^{-1}(p_t - A_t - B(d_t - \mu))$ in the market clearing equation to get,

$$\begin{aligned}
\bar{x} + x_{t+1} &= \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}\left(p_t - A_t - B\left(d_t - \mu\right)\right) + \Phi_t y_{t+1} + \Psi_t z_{t+1} \\
&\quad + \Omega_t\left[A_{t+1} + \mu + (B+I)G(d_t - \mu) - rp_t\right] \\
\implies \left(\Gamma_t\Sigma_{pt}^{-1}C_t^{-1} - r\Omega_t\right)p_t &= \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}A_t + \Gamma_t\Sigma_{pt}^{-1}C_t^{-1}B\left(d_t - \mu\right) \\
&\quad - \Omega_t A_{t+1} - \Omega_t\mu - \Omega_t(B+I)G(d_t - \mu) \\
&\quad - \Phi_t y_{t+1} - \Psi_t z_{t+1} + \bar{x} + x_{t+1}
\end{aligned}$$

Let $M_t = \Gamma_t\Sigma_{pt}^{-1}C_t^{-1} - r\Omega_t$. Using the linear conjecture for the price p_t, we match coefficients as follows:

- A_t to all the constant terms: $A_t = M_t^{-1}\left[\Gamma_t\Sigma_{pt}^{-1}C_t^{-1}A_t - \Omega_t A_{t+1} - \Omega_t\mu + \bar{x}\right]$
- B to all terms with $d_t - \mu$: $B = M_t^{-1}\left[\Gamma_t\Sigma_{pt}^{-1}C_t^{-1}B - \Omega_t(B+I)G\right]$
- C_t to all terms with y_{t+1}: $C_t = -M_t^{-1}\Phi_t$
- D_t to all terms with x_{t+1}: $D_t = M_t^{-1}$
- F_t to all terms with z_{t+1}: $F_t = -M_t^{-1}\Psi_t$

Solving this yields,

$$\begin{cases} A_t = \frac{1}{r}\left\{A_{t+1} + \mu - \Omega_t^{-1}\bar{x}\right\} \\ B = (r-G)^{-1}G \\ C_t = -M_t^{-1}\Phi_t \\ D_t = M_t^{-1} \\ F_t = -M_t^{-1}\Psi_t \end{cases} \tag{17}$$

Special Cases We consider some special cases, where our expressions should reduce to more familiar forms.

1. $K_{it} = K$: In case all investors share the same precision of the private component of signal, none of the expressions change substantially.

$$\begin{aligned}
\Sigma_{it} &= \left\{\Sigma_d + \Sigma_{pt}^{-1} + \left(\zeta_{it}^2\Sigma_z + K^{-1}\right)^{-1}\right\}^{-1}, \ \ \mu_{it} = \Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta_{it}^2\Sigma_z + K^{-1}\right)^{-1} s_{it}\right) \\
\Omega_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di \\
\Gamma_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}(B+I)\Sigma_{it} di \\
\Phi_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it}\left(\zeta_{it}^2\Sigma_z + K^{-1}\right)^{-1} di \\
\Psi_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} \\
&\quad \times (B+I)\Sigma_{it}\left(\zeta_{it}^2\Sigma_z + K^{-1}\right)^{-1}\zeta_{it} di
\end{aligned}$$

2. $\zeta_{it} = 0$: In case none of the investors read the public signal, some of our expressions change to indicate that the public signal noise is no longer relevant to the problem.

$$\begin{aligned}
\Sigma_{it} &= \left\{\Sigma_d + \Sigma_{pt}^{-1} + K_{it}\right\}^{-1}, \ \ \mu_{it} = \Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + K_{it}s_{it}\right) \\
\Omega_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}'\right)^{-1} di \\
\Gamma_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}'\right)^{-1}(B+I)\Sigma_{it} di \\
\Phi_t &= \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}'\right)^{-1}(B+I)\Sigma_{it}K_{it} di \\
\Psi_t &= 0
\end{aligned}$$

3. $\zeta_{it} = 1$: In case all investors read the public signal, some of our expressions change to indicate that the investors do not fully disentangle the public signal noise from the dividend innovation

(since the private signal is essentially an unbiased signal for $y_{t+1} + z_{t+1}$ in this case).

$$\Sigma_{it} = \left\{\Sigma_d + \Sigma_{pt}^{-1} + \left(\Sigma_z + K_{it}^{-1}\right)^{-1}\right\}^{-1}, \;\; \mu_{it} = \Sigma_{it}\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\Sigma_z + K_{it}^{-1}\right)^{-1} s_{it}\right)$$
$$\Omega_t = \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1} di$$
$$\Gamma_t = \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}(B+I)\Sigma_{it} di$$
$$\Phi_t = \int_i \rho_i^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}$$
$$\times (B+I)\Sigma_{it}\left(\Sigma_z + K_{it}^{-1}\right)^{-1} di$$
$$\Psi_t = \Phi_t$$

For the remaining exposition, we consider the special case where all investors have the same exposure to the public signal $\zeta_{it} = \zeta$ and the same precision of the orthogonal private component of the signal $K_{it} = K$. The only source of individual level variation in the model solution remains in the risk tolerance and the signal realization. Hence, the covariance expressions simplify to reflect this, only aggregating across individuals using the average risk tolerance (since the signal realizations don't affect the covariances).

$$\Sigma_{it} = \Sigma_t = \left\{\Sigma_d + \Sigma_{pt}^{-1} + \left(\zeta^2\Sigma_z + K^{-1}\right)^{-1}\right\}^{-1}, \;\; \mu_{it} = \Sigma_t\left(\Sigma_{pt}^{-1}\eta_{pt} + \left(\zeta^2\Sigma_z + K^{-1}\right)^{-1} s_{it}\right)$$
$$\Omega_t = \bar{\rho}^{-1}\left((B+I)\Sigma_t(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}$$
$$\Gamma_t = \bar{\rho}^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}(B+I)\Sigma_t$$
$$\Phi_t = \bar{\rho}^{-1}\left((B+I)\Sigma_{it}(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right)^{-1}$$
$$\times (B+I)\Sigma_t\left(\zeta^2\Sigma_z + K^{-1}\right)^{-1}$$
$$\Psi_t = \Phi_t\zeta$$

Here, we use the average risk tolerance $\bar{\rho} = \left(\int_i \rho^{-1} di\right)^{-1}$, which is simply the harmonic mean of the risk tolerance across individuals.

B Proofs for Lemmas 1 and 2

In order to prove Lemma 1, we first state and prove an interim utility result.

Lemma 3. *In a perfectly competitive market ($n \to \infty$), investor expected utility at date t, conditional on all date-t data is*

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tag{18}$$

Proof of Lemma 3.

From Equation (1) and Equation (10), end-of-period consumption for an investor can be represented as

$$c_{it+1} = r(\overline{w}_{it} - q'_{it}\theta_i p_t) + q'_{it}\theta_i(p_{t+1} + d_{t+1}) = r\overline{w}_{it} + q'_{it}\Pi_{it}.$$

The ex ante utility of the investor is,

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_t^-\right] = \mathbb{E}\left[\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] \mid \mathcal{I}_t^-\right]$$

That is, we calculate the ex ante utility from the interim utility using the law of iterated expectations. From Equation (4), the interim utility is given as

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i \mathbb{E}\left[r\overline{w}_{it} + q'_{it}\Pi_{it} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2}\mathbb{V}\left[r\overline{w}_{it} + q'_{it}\Pi_{it} \mid \mathcal{I}_{it}\right].$$

The first order condition for optimal portfolio choice implies $q_{it} = \rho_i^{-1}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$. Here, we used the fact that the only variable term in Π_{it} is $p_{t+1} + d_{t+1}$ at the interim stage. The first term of the interim utility is,

$$\rho_i \mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{19}$$

The second term of the interim utility can be written as

$$\frac{\rho_i^2}{2}\mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' V\left(\Pi_{it} \mid \mathcal{I}_{it}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{20}$$

Taking the difference of the first term and the second term yields the result in Lemma 3

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' V\left(\Pi_{it} \mid \mathcal{I}_{it}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \tag{21}$$

□

Proof of Lemma 1. Expand the expression for profit Π_{it} as,

$$\begin{aligned}
\Pi_{it} &= \theta_i\left[p_{t+1} + d_{t+1} - rp_t\right] \\
&= \theta_i\left[A_{t+1} + B(d_{t+1} - \mu) + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} + (d_{t+1} - \mu) + \mu - rp_t\right] \\
&= \theta_i\left[A_{t+1} + \mu + (B + I)\left[G(d_t - \mu) + y_{t+1}\right] + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} - rp_t\right] \\
&= \theta_i\left[A_{t+1} + \mu + (B + I)G(d_t - \mu) + (B + I)y_{t+1} + C_{t+1}y_{t+2} + D_{t+1}x_{t+2} + F_{t+1}z_{t+2} - rp_t\right]
\end{aligned}$$

The interim variance of the profit is given as,

$$\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] = \theta_i \left[(B+I)\Sigma_t(B+I)' + C_{t+1}\Sigma_d C_{t+1}' + D_{t+1}\Sigma_x D_{t+1}' + F_{t+1}\Sigma_z F_{t+1}'\right]\theta_i' \tag{22}$$

Here, we use the posterior variance of the dividend innovation $\Sigma_t = \mathbb{V}\left[y_{t+1} \mid \mathcal{I}_{it}\right]$. Further, it is clear from Equation (22) that the interim variance of consumption $\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ is a known quantity – it is only a function of ζ_{it} and K_{it} (in our case, ζ and K), and not a function of information revealed at the interim stage p_t or s_{it}. That is, it is a function only of the model primitives and the information set $\mathcal{I}_0$.

Next, in the expression for the conditional expected utility from Lemma 3, we decompose the conditional expected profit (4) into an expected $\mathbb{E}\left[\Pi_{it}\right]$ and a surprise component $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]$,

$$\begin{aligned}
\mathbb{E}\left[U(c_{it+1})\right] &= \mathbb{E}\left[\mathbb{E}\left[U(c_{it+1} \mid \mathcal{I}_{it})\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left\{\mathbb{E}\left[\Pi_{it}\right]' + \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' - \mathbb{E}\left[\Pi_{it}\right]'\right)\right\}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left\{\mathbb{E}\left[\Pi_{it}\right] + \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right\}\right] \\
&\quad + r\overline{w}_{it}\rho_i \\
&= \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \underbrace{\mathbb{E}\left[\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right]}_{=0} \\
&\quad + \frac{1}{2}\mathbb{E}\left[\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)\right] + r\overline{w}_{it}\rho_i
\end{aligned} \tag{23}$$

We are interested in the second term of the ex ante expected utility in Equation (23). We will use the fact that the mean of a random variable with the central chi-square distribution is the trace of the covariance matrix of the underlying normal variable,

$$\mathbb{E}\left[U(c_{it+1})\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2}\mathrm{Tr}\left[\mathbb{V}\left[\Upsilon_{it}\right]\right] + r\overline{w}_{it}\rho_i \tag{24}$$

$$\text{where, } \Upsilon_{it} = \left(\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right)'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-\frac{1}{2}} \tag{25}$$

We can express $\mathbb{V}\left[\Upsilon_{it}\right]$ as,

$$\begin{aligned}
\mathbb{V}\left[\Upsilon_{it}\right] &= \mathbb{V}\left[\left\{\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right\}'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-\frac{1}{2}}\right] \\
&= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}
\end{aligned}$$

Hence, the term of interest is the prior variance of the ex ante stochastic quantity $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$, since the prior expectation of this quantity $\mathbb{E}\left[\Pi_{it}\right]$ is a known variable ex ante. Hence, we can use the law of total variance, which says that the prior variance of the posterior expectation $\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ is

equal to the prior variance minus the posterior variance for Π_{it},

$$\begin{aligned}\mathbb{V}\left[\Upsilon_{it}\right] &= \left\{\mathbb{V}\left[\Pi_{it}\right] - \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right]\right\}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} \\ &= \mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\end{aligned}$$

Hence, we can express the ex ante expected utility as,

$$\mathbb{E}\left[U(c_{it+1})\right] = \frac{1}{2}\mathbb{E}\left[\Pi_{it}\right]'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\mathbb{E}\left[\Pi_{it}\right] + \frac{1}{2}\mathrm{Tr}\left[\mathbb{V}\left[\Pi_{it}\right]\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\overline{w}_{it}\rho_i$$

□

Proof of Lemma 2. Differentiating expected interim utility, when price p_t depends on investor i's demand yields a first order condition,

$$\begin{aligned}q_{it} &= \left[\rho_i\mathbb{V}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] + \frac{dp}{dq_i}\right]^{-1}\left\{\mathbb{E}\left[p_{t+1} + d_{t+1} \mid \mathcal{I}_{it}\right] - rp_t\right\} \\ &= \left(\rho_i\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{dp}{dq_i}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]. \qquad (26)\end{aligned}$$

The term dp/dq_i, often referred to as "Kyle's lambda" is the measure of how much effect investor i's demand has on the market price of an asset.

Interim utility still takes the form

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = \rho_i\mathbb{E}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2}\mathbb{V}\left[r\overline{w}_{it} + q_{it}'\Pi_{it} \mid \mathcal{I}_{it}\right].$$

However, substituting in the new expression for q_{it} from Equation (26), the first term of the interim utility is now

$$\rho_i\mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_t \mid \mathcal{I}_{it}\right]'\left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1}\mathbb{E}\left[\Pi_t \mid \mathcal{I}_{it}\right]$$

The second term of the interim utility can be written as

$$\begin{aligned}\frac{\rho_i^2}{2}\mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right] &= \frac{\rho_i^2}{2}q_{it}'\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]q_{it} \\ &= \frac{1}{2}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]'\left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\left(\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}\right)^{-1}\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\end{aligned}$$

Let $\tilde{V}_i \coloneqq \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i}$. Note that all terms in $\tilde{V}_i$ are known ex ante to investor i. Taking the

difference of the first term and the second term yields interim expected utility

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] = r\overline{w}_{it}\rho_i + \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]' \tilde{V}_i^{-1} \left(I - \frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tilde{V}_i^{-1}\right) \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tag{27}$$

To compute ex-ante utility, we follow the same steps as in the proof for Lemma 1. The solution is also similar, except that we replace $\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ with $\hat{V}_i \coloneqq \tilde{V}_i \left(I - \frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \tilde{V}_i^{-1}\right)^{-1}$ in Equation (24) and in Equation (25). Similar to $\tilde{V}_i$, all terms in $\hat{V}_i$ are known to investor i ex ante. In this case,

$$\begin{aligned}\mathbb{V}\left[\Upsilon_{it}\right] &= \mathbb{V}\left[\left\{\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right\}' \hat{V}_i^{-\frac{1}{2}}\right] \\ &= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] - \mathbb{E}\left[\Pi_{it}\right]\right] \hat{V}_i^{-1}\end{aligned}$$

Applying the law of total variance,

$$\mathbb{V}\left[\Upsilon_{it}\right] = \left(\mathbb{V}\left[\Pi_{it}\right] - \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right) \hat{V}_i^{-1}.$$

Substituting $\hat{V}_i$ for $\frac{1}{2}\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]$ in Equation (24) and using the new expression for $\mathbb{V}\left[\Upsilon_{it}\right]$ yields

$$\tilde{U}(\mathcal{I}_{it}) = \mathbb{E}\left[\Pi_{it}\right]' \hat{V}_i^{-1} \mathbb{E}\left[\Pi_{it}\right] + \mathrm{Tr}\left[\left(\mathbb{V}\left[\Pi_{it}\right] - \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right) \hat{V}_i^{-1}\right] + r\rho_i\bar{w}_{it}. \tag{28}$$

□

C Unconditional Utility in terms of Excess Returns

In this Appendix, we impose a key approximation which allows us to express the unconditional utility of the investor in terms of moments of excess, as defined in Equation (10), as opposed to profits. Recall from Equation (10)

$$R_{it} = \Pi_{it} \oslash \theta_i p_t. \tag{29}$$

Noting that p_t is known at the interim stage (in the information set $\mathcal{I}_{it}$), we start by writing the expressions for the conditional moments of R_{it}

$$\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right] = \mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t, \text{and} \tag{30}$$

$$\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] = \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t p_t' \theta_i'. \tag{31}$$

We assume that the *ex-ante* variation in $\theta_i p_t$ is small relative to the other terms in the expected utility expression. Formally, this amounts to assuming that $\theta_i p_t$ is a constant from an *ex-ante* perspective. This allows us to use the law of iterated expectations and express the ex ante expectation

of excess return R_{it} as

$$\begin{aligned}\mathbb{E}\left[R_{it}\right]_j &= \mathbb{E}\left[\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right]\right]_j = \mathbb{E}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t\right]_j = \mathbb{E}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]_j / \left(\theta_i p_t\right)_j\right] \\ &\approx \frac{\mathbb{E}\left[\Pi_{it}\right]_j}{\left(\theta_i p_t\right)_j}. \end{aligned} \tag{32}$$

Or equivalently,

$$\mathbb{E}\left[R_{it}\right] = \mathbb{E}\left[\Pi_{it}\right] \odot \left(\theta_i p_t\right)^{\circ(-1)}, \tag{33}$$

where $\odot$ is the Hadamard (element-wise) product of two matrices and $W^{\circ(-1)}$ represents the Hadamard (element-wise) inverse of a matrix W. Further, we use the law of total variance to express the unconditional variance of R_{it} as

$$\begin{aligned}\mathbb{V}\left[R_{it}\right] &= \mathbb{V}\left[\mathbb{E}\left[R_{it} \mid \mathcal{I}_{it}\right]\right] + \mathbb{E}\left[\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\right] \\ &= \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t\right] + \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right] \oslash \theta_i p_t p_t' \theta_i'\right] \\ &\approx \mathbb{V}\left[\mathbb{E}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right] \oslash \theta_i p_t p_t' \theta_i' + \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]\right] \oslash \theta_i p_t p_t' \theta_i' \\ &= \mathbb{V}\left[\Pi_{it}\right] \oslash \theta_i p_t p_t' \theta_i' \end{aligned} \tag{34}$$

Perfectly Competitive Markets We can now use Equations (31), (33) and (34) to express the unconditional expected utility from Lemma 1 in terms of R_{it}. We get the expression for the ex ante expected utility in terms of excess returns as

$$\begin{aligned}\tilde{U}(\mathcal{I}_{it}) &= \frac{1}{2}\left\{\mathbb{E}\left[\Pi_{it}\right]' \mathbb{E}\left[\mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1}\right] \mathbb{E}\left[\Pi_{it}\right]\right\} + \frac{1}{2}\operatorname{Tr}\left[\mathbb{V}\left[\Pi_{it}\right] \mathbb{V}\left[\Pi_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\overline{w}_{it}\rho_i \\ &\approx \frac{1}{2}\left\{\mathbb{E}\left[R_{it}\right]' \mathbb{E}\left[\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1}\right] \mathbb{E}\left[R_{it}\right]\right\} + \frac{1}{2}\operatorname{Tr}\left[\mathbb{V}\left[R_{it}\right] \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]^{-1} - I\right] + r\overline{w}_{it}\rho_i \end{aligned} \tag{35}$$

Imperfectly Competitive Markets Using Equation (31), we can express the modified variance $\tilde{V}_i$ as

$$\begin{aligned}\tilde{V}_i &= \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] \odot \theta_i p_t p_t' \theta_i' + \frac{1}{\rho_i}\frac{dp}{dq_i} \\ &= \left(\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i} \oslash \theta_i p_t p_t' \theta_i'\right) \odot \theta_i p_t p_t' \theta_i' \\ &= \tilde{\tilde{V}}_{it} \odot \theta_i p_t p_t' \theta_i', \end{aligned} \tag{36}$$

where

$$\tilde{\tilde{V}}_{it} \coloneqq \left(\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] + \frac{1}{\rho_i}\frac{dp}{dq_i} \oslash \theta_i p_t p_t' \theta_i'\right). \tag{37}$$

Similarly, we restate $\hat{V}_i$ as

$$\hat{V}_i = \tilde{\tilde{V}}_{it}\left(I - \frac{1}{2}\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right] \tilde{\tilde{V}}_{it}^{-1}\right)^{-1} \odot \theta_i p_t p_t' \theta_i' = \hat{\tilde{V}}_{it} \odot \theta_i p_t p_t' \theta_i', \tag{38}$$

where

$$\hat{\tilde{V}}_{it} \coloneqq \tilde{\tilde{V}}_{it}\left(I - \frac{1}{2}\mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\tilde{\tilde{V}}_{it}^{-1}\right)^{-1}. \tag{39}$$

We can now use Equations (31), (33), (34), (37) and (39) to express the unconditional expected utility from Lemma 2 in terms of R_{it}. We get the expression for the ex ante expected utility in terms of excess returns as

$$\tilde{U}(\mathcal{I}_{it}) \approx \mathbb{E}\left[R_{it}\right]'\hat{\tilde{V}}_{it}^{-1}\mathbb{E}\left[R_{it}\right] + \mathrm{Tr}\left[\left(\mathbb{V}\left[R_{it}\right] - \mathbb{V}\left[R_{it} \mid \mathcal{I}_{it}\right]\right)\hat{\tilde{V}}_{it}^{-1}\right] + r\rho_i\bar{w}_{it}. \tag{40}$$

D Valuing Order Flow Data

Consider an extension of the model where investors can observe data on sentiment shocks from H different data sources. Investors have the same preference and choose their risky asset investment q_{it} to maximize $\mathbb{E}\left[U(c_{it+1})|\mathcal{I}_{it}\right]$, taking the asset price and the actions of other investors as given, subject to the budget constraint (1). A given piece of data m from data source h is now a signal about x_{t+1}: $\eta_{iht}^{mx} = \psi_h^x x_{t+1} + \Gamma_h^x e_{it}^x$, with $e_{it}^x \overset{iid}{\sim} \mathcal{N}(0, I)$.

Information on sentiment shocks allows an investor i to extract a more precise signal about dividends from prices $s_{it}^p = y_{t+1} + C_t^{-1}D_t\left(x_{t+1} - \mathbb{E}\left[x_{t+1} \mid s_{it}^x\right]\right)$. While investors probably do not think about using order flow data to learn about fundamentals, they often trade against uniformed order flow (sentiment). This is mathematically equivalent to using sentiment to extract clearer fundamental information from price and then trading on that fundamental information.

The solution of this model is a straightforward n-asset extension of the model with order flow information in Farboodi and Veldkamp (2017). Given an $N \times 1$ unbiased signal s_{it}^y about the dividend innovations y_{t+1} with precision matrix k_{it}^y and an $N \times 1$ unbiased signal s_{it}^x about the sentiment shocks y_{t+1} with precision matrix k_{it}^x, investors apply Bayes' law. They combine their prior, information in the sentiment-adjusted market price, and information on dividend innovation obtained from the data to form a posterior view about the $(t+1)$-period dividend d_{t+1}. The posterior precision is $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]^{-1} = \Sigma_0^{-1} + C_t^{-1}D_t\left(\Sigma_x + (k_{it}^x)^{-1}\right)^{-1}D_t'C_t^{-1\prime} + k_{it}^y$.

At each date t, the risky asset price equates demand with noise trades plus one unit of supply, as described by Equation (2). The equilibrium price is still a linear combination of past dividends d_t, the t-period dividend innovation y_{t+1}, and the sentiment shock x_{t+1}, as in Equation (2).

Ex-ante utility is still given by Equation (3). The precision variables k_{it}^y and k_{it}^x enter through the posterior variance $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$ and $\mathbb{V}\left[\Pi_t \mid \mathcal{I}_{it}\right]$. In the second term, k_{it}^y and k_{it}^x enter only through $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$. Thus, $\mathbb{V}\left[d_{t+1} \mid \mathcal{I}_{it}\right]$ is a sufficient statistic for expected utility. The fact that the uncertainty about dividends is a sufficient statistic, and the formulation of Bayes' law for posterior precision (the inverse of uncertainty), implies that k_{it}^y and k_{it}^x affect utility in the same way, except that k_{it}^x is multiplied by $C_t^{-1}D_tD_t'C_t^{-1\prime}$. This ratio of price coefficients represents the squared signal-to-noise ratio in prices, where C is the price coefficient on the signal (future dividend) and D is the coefficient on noise (sentiment). The bottom line is that the value of sentiment data

is exactly the same as the value of fundamental data, after adjusting for the signal-to-noise ratio in prices. That signal-to-noise adjustment is exactly what an OLS procedure does.