

# Where Has All the Big Data Gone?

Maryam Farboodi\* Adrien Matray† Laura Veldkamp‡

December 22, 2017§

## Abstract

As ever more technology is deployed to process and transmit financial data, this could benefit society, by allowing capital to be allocated more efficiently. Recent work supports this notion. Bai, Philippon, and Savov (2016) document an improvement in the ability of S&P 500 equity prices to predict firms' future earnings. We show that most of this "price informativeness" rise comes from a composition effect. S&P 500 firms are getting older and larger. In contrast, the average public firm's price information is deteriorating. Do these facts imply that big data failed to price assets more efficiently? To answer this question, we formulate a model of data-processing choices. We find that big data growth, in conjunction with a change in the firm size distribution, can trigger a concurrent surge in large firm price informativeness and decline in informativeness for small firms. The implication is that ever-growing reams of data processed by the financial sector might not deliver efficiency benefits, for the vast majority of firms.

---

\*Princeton University; farboodi@princeton.edu

†Princeton University; amatray@princeton.edu

‡Department of Economics Stern School of Business, NBER, and CEPR, New York University, 44 W. 4th Street, New York, NY 10012; lveldkam@stern.nyu.edu; <http://www.stern.nyu.edu/~lveldkam>.

§We thank John Barry, Matias Covarrubias, Ye Zhen and Joseph Abadi for their excellent research assistance, Pete Kyle for his insightful and helpful suggestions, and participants in the 2017 NBER Summer Institute for their comments. JEL codes: Keywords: financial technology, big data, capital misallocation.

Does the adoption of financial technology add social value? The answer to this basic question lies at the heart of many policy and regulatory debates. Recent evidence that the informativeness of asset prices has been increasing (Bai, Philippon, and Savov, 2016) suggests that the labor, technology, and human capital growth in the financial sector is yielding real benefits, in terms of more efficient capital allocation for firms. However, this rosy headline result of greater price informativeness pertains to firms in the S&P 500. More accurately, large, old firms are priced, and have always been priced, more efficiently. These large firms have simply become more prevalent in the S&P 500. For the universe of publicly traded firms as a whole, price informativeness has, in fact, declined. This paper explores these competing facts, teases out composition effects from trends, explores the reasons for the shifts in market efficiency, and concludes that most firms may miss out on the financial benefits of the big data revolution.

Section 1 start by exploring the question: What is it about S&P 500 firms that explains why their prices became more informative when other firms' price informativeness has fallen? Is current membership in the S&P important? No, we compare the set of all firms that have been in the S&P 500 at some time, and find that price of firms currently in the S&P 500 have neither a higher level nor a steeper trend in price informativeness. Is this an industry effect? No, we find that for industries most represented in the S&P, the firms in those industries that are not themselves S&P 500 members have witnessed no rise in price information. Perhaps this is a shift to more high-tech firms, which are harder to price. Yes, but that only explains a small fraction of the effect. What can explain the divergence is the change in firm size. We find that the set of firms currently in the S&P 500 are getting larger over time. Since larger firms have more informative prices, there is a size composition effect. We show that this shift in firm size can account for most of the rise in the informativeness of S&P 500 prices.

The conclusion one might draw from these facts is that big data has not helped financial markets to better price assets at all. This is all just a change in the composition of firm size. That conclusion would raise a few questions. First, why would financial firms pour resources into data technology, if not to trade in a more informed way? The other puzzle is why is overall price informativeness declining? The universe of public firms is not shifting toward smaller firms. How is it possible that despite the deluge of data, some firms' prices contain

less information today than they did 30 years ago? Does this imply that the data is useless, or that rational or behavioral market inefficiencies prevent big data from benefitting most firms?

To answer these questions, Section 3 uses a simple model to work out the logical consequences of big data growth and large firm growth for data allocation and price informativeness. We use a portfolio choice model, with multiple, risky assets, where investors may choose to process data about any or all of those assets. Data comes in the form of binary strings that encode information about the future value of the risky assets. What investors are choosing is the length of the binary code, for each asset. Given their encoded data, investors update beliefs about risk and return and make portfolio investment choices. Asset prices clear the market for each asset. We find that the divergence in price informativeness is compatible with optimizing agents and markets that have no friction, other than imperfect information.

If increasing data processing were the only force at work, the firms' price informativeness should rise across the board. A key force is that large firms' data is particularly valuable to an investor. If the largest firms grow larger, they become more attractive targets for data processing and they draw attention away from the relatively less attractive small firms. This can explain why large firm prices become more informative and small firm prices less informative.

These findings are consistent with a financial sector that has improved its ability to process data and use that knowledge to price assets. But, by no means, do they prove that overall efficiency did, in fact, increase. The facts and model together do allow us to bound the increase in data productivity. If the growth in data processing is too large, relative to the increase in the size of large firms, then such a combination of forces would be unable to explain the decline in price informativeness of small firms. Section 4 concludes that, while big data may be helping investors to price assets more accurately, the technological gains are modest, and are failing to help many smaller firms that might well be our future engines of growth.

**Our Contribution Relative to the Literature** Examinations of the effects of big data are scarce. Empirical work primarily examines whether particular data sources, such as social media text, predict asset price movements (Ranco, Aleksovski, Caldarelli, Grcar, and Mozetic, 2015). In contrast, many papers have developed approaches to measuring stock market informativeness across countries (Edmans, Jayaraman, and Schneemeier, 2016), or (Durnev, Morck, and Yeung, 2004). The novelty of our approach, compared to these studies is that we study how price informativeness evolves over time and in the cross-section, because it reveals changes in financial efficiency.

Explorations of how information production affects real investment (Bond, Edmans, and Goldstein, 2012; Goldstein, Ozdenoren, and Yuan, 2013; Dow, Goldstein, and Guembel, 2017; Bond and Eraslan, 2010; Ozdenoren and Yuan, 2008) complement our work by showing how the financial information trends we document could have real economic effects. Our work also contributes to the debate on the sources of capital misallocation in the macroeconomy.<sup>1</sup> Like David, Hopenhayn, and Venkateswaran (2016), our focus is on the role financial markets play in informing these real investment choices. We add an explanation for why financial markets may be providing better guidance over time for some firms, but not for others.

On the theoretical side, the information theory (computer science) based measures we use to quantify big data flows are similar to those used in work on rational inattention (Sims, 2003; Maćkowiak and Wiederholt, 2009; Kacperczyk, Nosal, and Stevens, 2015).<sup>2</sup> Our model extends Farboodi and Veldkamp (2017) in two ways. First, our information processing constraint corresponds to computer science measures of data processing, based on bits. That change allows us to map processing directly to CPU speed. Second, instead of a single risky asset, we have heterogeneous asset characteristics. This is essential for our model to speak to the cross-sectional data. It allows us to explain how firm size and

---

<sup>1</sup>See e.g., Hsieh and Klenow (2009) or Restuccia and Rogerson (2013) for a survey.

<sup>2</sup>More broadly, equilibrium models with information choice have been used to explain income inequality (Kacperczyk, Nosal, and Stevens, 2015), information aversion (Andries and Haddad, 2017), home bias (Mondria, Wu, and Zhang, 2010; Van Nieuwerburgh and Veldkamp, 2009), mutual fund returns (Pástor and Stambaugh, 2012; Stambaugh, 2014), among other phenomena. Related microstructure work explores the frequency of information acquisition and trading (Kyle and Lee, 2017; Dugast and Foucault, 2016; Chordia, Green, and Kottimukkalur, 2016; Crouzet, Dew-Becker, and Nathanson, 2016). Davila and Parlato (2016) share our focus on price information, but do not examine its time trend or cross-sectional differences. Empirical work (Katz, Lustig, and Nielsen, 2017) finds evidence of rational inattention like information frictions in the cross section of asset prices.

technology affect data processing and thus price informativeness.

# 1 Data and Measurement of Price Informativeness

## 1.1 Data

The data we are using are for the U.S. over the period 1964–2015. Stock prices come from CRSP (Center for Research in Security Prices). All accounting variables are from Compustat. We take stock prices as of the end of March and accounting variables as of the end of the previous fiscal year, typically December. This timing convention ensures that market participants have access to the accounting variables that we use as controls.

The main equity valuation measure is the log of market capitalization  $M$  over total assets  $A$ ,  $\log(M/A)$  and the main cash flow variable is earnings measured as EBIT (earnings before interest and taxes, denoted  $EBIT$  in Compustat). This measure includes current and future cash flows, and investment by current total assets. All ratio variables are winsorized at the 1%.

Since we are interested in how well prices forecast future earnings, and future earnings are affected by inflation, we need to consider how to treat inflation. We adjust for inflation with GDP deflator to ensure that differences in future nominal cash flows do not pollute our estimation of stock price informativeness.

## 1.2 Measuring Price Informativeness

While there is a debate in the empirical literature about how to best measure price informativeness (e.g. Philippon, 2015), the measure suggested by Bai, Philippon and Savov (2016) is closest to our model’s measure. It captures the extent to which asset prices in year  $t$  are able to predict future cash-flows in year  $t + k$ .

Their informativeness measure is constructed by running cross-sectional regressions of future earnings on current market prices. Controlling for other observables limits the risk of confounding public information impounded in prices with markets foresight. For each firm  $j$ , in year  $t$ , we estimate  $k$ -period ahead informativeness as

$$\frac{E_{j,t+k}}{A_{j,t}} = \alpha + \beta_t \log\left(\frac{M_{j,t}}{A_{j,t}}\right) + \gamma X_{j,t} + \epsilon_{i,t}, \quad (1)$$

where  $E_{j,t+k}/A_{j,t}$  is the cash-flow of firm  $j$  in year  $t+k$ , scaled by total assets of the firm in year  $t$ ;  $\log(M_{j,t}/A_{j,t})$  is firm market capitalization, scaled by total assets; and  $X_{j,t}$  are controls for firm  $j$  that capture publicly available information. In the main specification, the controls are current earnings and industry sector (SIC 1) fixed effects. When we estimate price informativeness at the industry level (SIC3 or SIC2), we need to drop the industry fixed effect as a control.

The parameter  $\beta_t$  measures the extent to which firm market capitalization in year  $t$  can forecast the firm cash-flow in year  $t+k$ . To map this coefficient into a proxy of price informativeness, we follow Bai et al. (2016) and do the following adjustment:

$$\left(\sqrt{P_{Info}}\right)_t = \beta_t \times \sigma_t(\log(M/A)) \quad (2)$$

where  $\sigma_t(\log(M/A))$  is the cross-sectional standard deviation of the forecasting variable  $\log(M/A)$  in year  $t$ . The use of square root gives the measure an economic interpretation as dollars of future cash flows per dollar of current total assets.

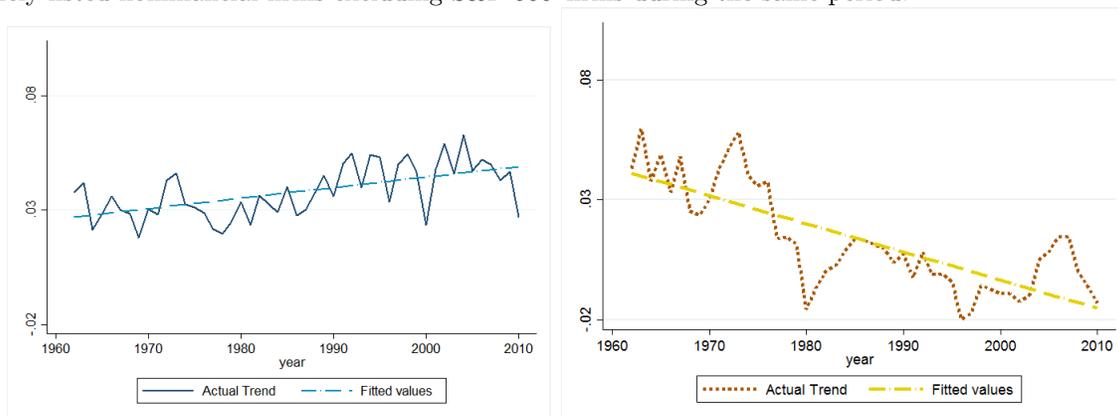
### 1.3 Aggregate Trends in Price Informativeness

We first establish the empirical puzzle that motivates our analysis. Price informativeness increases over time for firms in the S&P 500 (Bai, Philippon, and Savov (2016)'s headline result), but it decreases when we look at all the other publicly listed nonfinancial firms, excluding S&P 500 firms. Figure 1 illustrates the contrast between the increase in informativeness for S&P 500 firms (left figure) and the decrease in price informativeness for all non-S&P 500 firms (right figure). We observe a similar decline if we look at the universe of listed firms (including both S&P 500 and non-S&P 500 firms).

Similar plots in the Appendix reveal that the trends are nearly identical for 3-year and 5-year horizons. Therefore we proceed by looking only at 5-year price informativeness.

Table 1 quantifies these trends and demonstrates the statistical significance of the difference between the S&P 500 and all-public-firm samples. Both trends are economically large.

**Figure 1: Price Informativeness is Rising for S&P 500 Firms but Falling for All other Public Firms.** Results from the cross-sectional forecasting regression (eqn 1):  $E_{i,t+k}/A_{i,t} = \alpha + \beta_t \log(M_{i,t}/A_{i,t}) + \gamma X_{i,t} + \epsilon_{i,t}$ , where  $M$  is market cap,  $A$  is total asset,  $E$  is earnings before interest and taxes ( $EBIT$ ) and  $X$  are a set of controls that captures information publicly available. We run a separate regression for each year  $t = 1964, \dots, 2010$  and horizon  $k = 5$ . Price informativeness is  $\beta_t \times \sigma_t(\log(M/A))$ , where  $\sigma_t(\log(M/A))$  is the cross-sectional standard deviation of  $\log(M/A)$  in year  $t$ . Above each plot is a linear trend normalized to zero and one at the beginning and end of the sample (plotted in dashed lines). The left figure contains S&P 500 nonfinancial firms from 1964 to 2008, while the right figure contains all publicly listed nonfinancial firms excluding S&P 500 firms during the same period.



For the S&P 500 sample, the mean of price informativeness is 0.041 and its time-series standard deviation is 0.01. Between 1962 and 2010, price informativeness rose 70% relative to its mean, or 2.1 standard deviations. For non-S&P 500 firms, the average level of price informativeness is 0.028 with a time-series standard deviation of 0.012. So the fall in price informativeness is 100% relative to the mean and 2.5 times the standard deviation.<sup>3</sup>

## 2 Where Is Information Flowing?

The divergent aggregate informativeness trends offer a puzzling and mixed message about whether the financial sector is becoming more efficient or not. To understand what is going on and why, this section cuts the sample of firms in different ways, to understand which prices are becoming more informative and which less, or if this is a composition effect.

<sup>3</sup>For the S&P 500 sample, the interquartile range in price informativeness is 0.0162. The rise in price informativeness is about two times this interquartile distance. For non-S&P 500 firms, the interquartile range is 0.011. The fall in price informativeness is more than twice this interquartile distance.

**Table 1: Price Informativeness Trends over Time**

<i>Dep. Var</i>	100× Price Informativeness			
	S&P 500		All Listed Firms	
<i>Sample</i>	<i>k=3</i>	<i>k=5</i>	<i>k=3</i>	<i>k=5</i>
	(1)	(2)	(3)	(4)
Time Trend	1.76*** (0.33)	2.38*** (0.41)	-3.11*** (0.49)	-2.96*** (0.52)
Observations	45	45	45	45

This table presents time series regressions of price informativeness by horizon. Price informativeness is calculated as in Eq. 2 using estimates from the cross-sectional forecasting regression 1. For this table, we regress the time series of price informativeness at a given horizon  $k = 3, 5$  years on a linear time trend normalized to zero and one at the beginning and end of the sample. Newey - West standard errors, with five lags are in parentheses. \*\*\* denotes significance at the 1% level.

## 2.1 The Role of Firm Size

One possible explanation is that firms in the S&P 500 are, on average, getting larger, relative to other firms. Could differences in firm size explain the different trends in informativeness? Perhaps big data enabled us to improve analysis of large firms more than small ones? This hypothesis holds more promise. There are systematic differences in the level and trend of informativeness between small and large firms. But, this does not explain all of the difference between S&P 500 and non-S&P 500 firms.

We compute price informativeness into ten size bins in the following way: For the whole sample period, we compute bins based on firm size (market value deflated in 2009 dollars).<sup>4</sup> Bins are defined such that they contain roughly the same number of observations to avoid having biased estimates coming from large differences in sample size. Then we run separate cross-sectional regressions of price informativeness. Each regression takes the same form as (1), but with an additional  $y$  subscript for each size bin:

$$\frac{E_{i,y,t+k}}{A_{i,y,t}} = \alpha + \beta_{t,y} \log\left(\frac{M_{i,y,t}}{A_{i,y,t}}\right) + \gamma X_{i,y,t} + \epsilon_{i,y,t} \quad (3)$$

<sup>4</sup>This is the size variable that has been shown to matter in the context of CEO compensation for instance (e.g. Gabaix and Landier, 2008).

where  $E_{i,y,t+k}/A_{i,y,t}$  is the cash-flow of firm  $i$  belonging to size-bin  $y$  in year  $t+k$  scaled by total asset of the firm in year  $t$ .<sup>5</sup>

Figure 2: **Large Firms Have More Informative Prices.** Price informativeness is the ability to forecast future earnings (Eq 2). We run a separate regression for each year  $t = 1962, \dots, 2010$ , horizon  $k = 5$  and bin interval  $[1/10), \dots, [10/10]$  partitioned by 1/10 deciles. Firms are split by size. Price informativeness is the average value of  $\beta_{t,y} \times \sigma_{y,t} \log(M/A)$ , where  $\sigma_{y,t} \log(M/A)$  is the cross-sectional standard deviation of  $\log(M/A)$  in year  $t$  and size interval  $y$ . Future earnings are measure here at 5-year horizons. The sample contains publicly listed nonfinancial firms from 1962 to 2010.

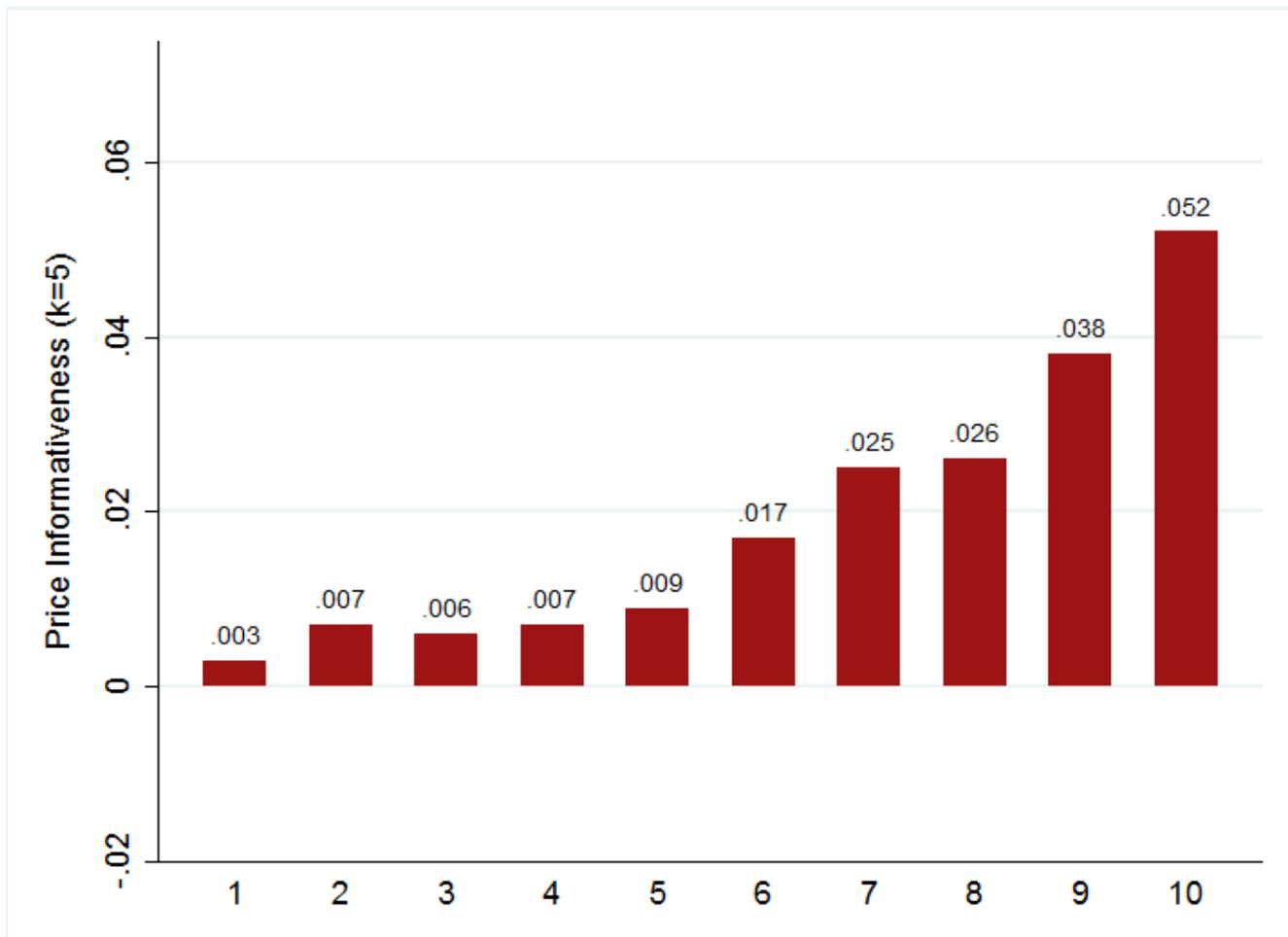


Figure 2 shows that larger firms have more informative prices. The effect is large. Moving from the first decile to the last decile of size implies an 17-fold increase in price informativeness.

It is possible that this result is driven by shifts of firms within decile bins. To make

<sup>5</sup>Adding year fixed effects to the cross-sectional specification does not change the result.

sure that the bin construction is not responsible for our results, we also estimate a similar regression using firm size as a continuous variable, over the whole sample. To see how the predictive power of firm stock price varies with firm size, we estimate

$$\frac{E_{i,t+k}}{A_{i,t}} = \alpha + \beta \log\left(\frac{M_{i,t}}{A_{i,t}}\right) \times M_{i,t} + \gamma_1 \log\left(\frac{M_{i,t}}{A_{i,t}}\right) + \gamma_2 M_{i,t} + \gamma_3 X_{i,t} + \epsilon_{i,t}.$$

The interaction between  $\log(M_{i,t}/A_{i,t})$  and  $M_{i,t}$  tells us how the ability of  $\log(M_{i,t}/A_{i,t})$  to predict firm  $i$ 's future cash-flow, varies with its size. Because we demean firm size, the interaction term can be interpreted as the marginal effect of firm size.<sup>6</sup> Table 2 reports the results when we cluster standard errors by industry and year. In Column (1), we find that  $\log(M_{i,t}/A_{i,t})$  is positive and significant at the 1% level. This clearly supports the idea that equity valuations forecast earnings. We also find that the interaction between  $\log(M_{i,t}/A_{i,t})$  and firm size is significant and positive. In other words, equity prices for large firms are better forecasters of those firms' earnings. In terms of magnitude, the largest decile of firms has more than twice the correlation between  $\log(M_{i,t}/A_{i,t})$  and future earnings, of the smallest decile of firms. Columns (2) and (3) confirm that the result is robust to year and industry fixed effects. Finally, Column (4) interacts all the variables with a time trend. The finding that this time interaction is positive and significant implies that the gap between large firms' and small firms' price informativeness has been growing over time.

Taken together, these results teach us that the increase in price information for S&P 500 firms may arise because of a change in the size composition of the S&P 500 . Given that large firms have more informative prices, a change in the composition of the S&P 500 toward larger firms can explain why the S&P 500 is becoming more informative, even if the informativeness of the largest firms is not rising. We explore this possible composition effect next.<sup>7</sup>

**Is this a composition effect?** Perhaps financial markets are not getting better at pricing larger firms over time, or any kind of firm in particular. It's simply that small firms have always been hard to price accurately and the composition of the S&P 500 changed so that

---

<sup>6</sup>In this case  $\log(M_{i,t}/A_{i,t})$  measures the effect for the median firm of demean size zero.

<sup>7</sup>Note that what we call a size effect could also be an age effect. Since the effect of size and age are similar and the attributes are highly correlated across firms, the two effects are hard to distinguish. We have replicated the same exercise with age, instead of size and obtained similar results.

**Table 2: Large Firms Have More Informative Prices**

<i>Dep. Var</i>	Earning <sub><i>t</i>+5</sub>			
	(1)	(2)	(3)	(4)
Size × log( <i>M/A</i> )	0.004*** (5.41)	0.003*** (5.32)	0.003*** (5.36)	-0.224*** (-2.97)
Size × log( <i>M/A</i> ) × Time Trend				0.224*** (3.00)
Size	0.000 (0.04)	0.000 (0.14)	0.001 (1.08)	-0.334*** (-3.36)
log( <i>M/A</i> )	0.013*** (2.61)	0.013*** (2.72)	0.015*** (3.12)	2.769*** (5.90)
Obs.	97778	97778	97778	97778
Sector FE	–	Yes	Yes	Yes
Year FE	–	–	Yes	Yes

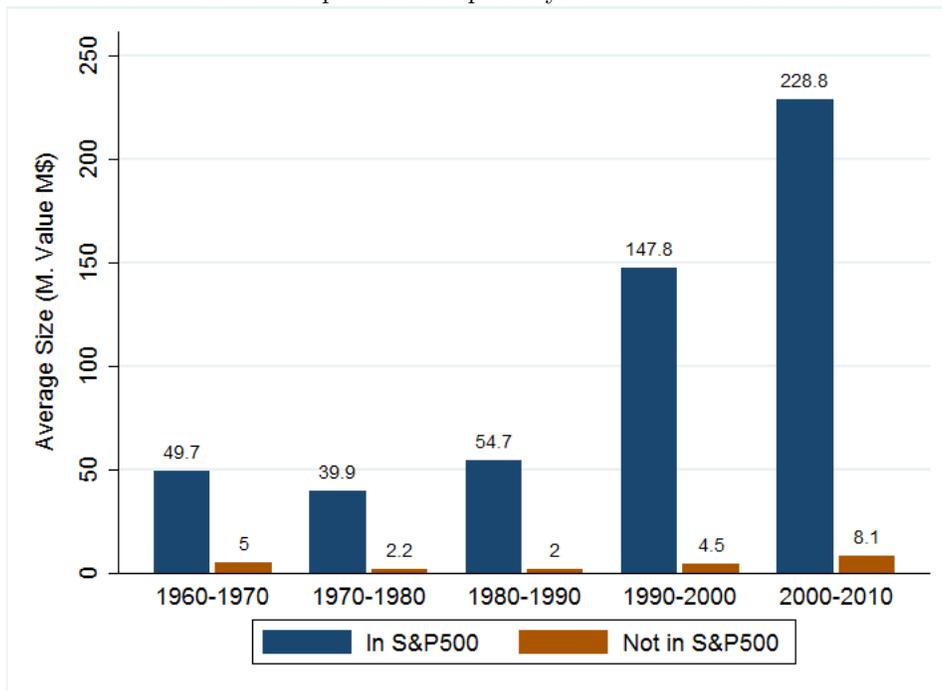
This table presents a cross-sectional regression of price informativeness as calculated as in Eq. 1. Earning of firm  $i$  in  $t + 5$  (measured by EBIT) is regressed on the natural logarithm of firm market capitalization scaled by total assets:  $\log(M/A)$ . *Size* is defined as the deflated firm market value in \$K. We control for earnings in  $t$  and include progressively year and industry fixed effects. Standard errors are clustered by industry and year. \*\*\* denotes significance at the 1% level.

there are fewer small firms in the index. In other words, S&P 500 price efficiency is rising because the average S&P 500 firm is getting larger. For this composition effect to explain the decline in overall price efficiency for all firms, it would have to be that the average non-S&P 500 firm is getting smaller.

Figure 3 supports the first hypothesis that S&P 500 firms are getting larger. But it does not support the second hypothesis that non-S&P 500 firms are getting smaller.

**How much of the trend can changing size composition explain?** The change in composition of S&P 500 firms to larger firms clearly favors an increase in price informativeness. But does the compositional change explain the entire trend? To determine this, we proceeded in three steps. First, we define size deciles from all firm-years in our sample and compute the average price informativeness in each decile as in Figure 2. Second, for each year, we compute the share of S&P 500 firms and the share of all firms that are in each

Figure 3: **S&P 500 Firms Became Larger. Non-S&P 500 Firms Grew Less.** We compare the average size of firms that are in S&P 500 and firms there are in S&P 500 over time. Firm size is defined as firm’s total asset. The sample contains publicly listed nonfinancial firms from 1960 to 2010.



decile.<sup>8</sup> Third, to get a size-predicted price informativeness trend, we multiply the share of each size decile by the average informativeness of firms in that decile to get the trend in price informativeness that changing size alone would explain. Formally, we compute the following equation:

$$\hat{\beta}_t^{size} = \sum_{y \in [1, \dots, 10]} \beta_y \times ShareFirms_{y,t}$$

where  $\beta_y$  is estimated in the cross-section over all firms belonging to size decile  $y$  (Equation 3) and  $ShareFirms_{y,t}$  correspond to the fraction of firms in year  $t$  that belongs to size decile  $y$ . The value of each  $\beta_y$  is displayed in Figure 2.

Figure 4 compares the measured price informativeness series (measured as in Figure 1, dark-blue line) and the size-predicted price informativeness ( $\hat{\beta}_t^{tech}$ , light-blue line). Of

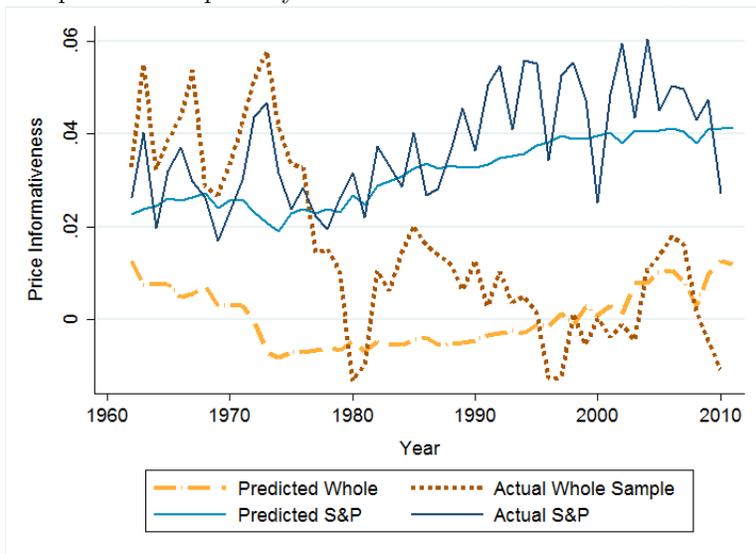
<sup>8</sup>Confirming the results in Section 2.1, we observe an increase in the fraction of S&P 500 firms in the top size decile. During the period, the fraction of S&P 500 firms in the top size decile grew from roughly 40% to almost 100%, while this fraction for the entire firm sample remained stable.

course, the measured price informativeness series fluctuates more. However, the trends of the actual and size-predicted series are pretty well aligned. This fact suggests that most of the increase in price informativeness in the S&P 500 can be explained by the change in firm size composition. Firms in S&P 500 getting are larger and the price informativeness of large firms is higher.

We do the same exercise for the whole sample. Since full-sample firms are not getting smaller, the predicted evolution of price informativeness for the whole sample (yellow dashed-line) does not explain the decline in measured informativeness (orange dotted-line).

**Figure 4: Predicted Evolution of PI based on Size: S&P 500 and Whole Sample.**

This figure shows the evolution of predicted and actual price informativeness for S&P 500 firms and the whole sample. For firms in the S&P 500, we show in the dark-blue line the coefficient  $\beta_t$  estimated from the cross-sectional forecasting regression defined in eqn 1. The orange dotted-line reports the same result when  $\beta_{t,5}$  is estimated for every listed firms (instead of restricting to S&P 500). The light-blue line and yellow dashed-line plot the evolution of the predicted  $\hat{\beta}_{t,5}^{size}$  computed in eqn 2.1.  $\hat{\beta}_{t,5}^{size}$  is the weighted sum of  $\beta_y$ , where  $y$  corresponds to a size-decile (eqn 3) and weights correspond to the fraction of firms in the same size-decile in a given year. The light-blue line plots the evolution of  $\hat{\beta}_{t,5}^{size}$  when we use as weights the fraction of firms in the S&P 500. The yellow dashed-line plots the same weighted average, except with weights that are the fraction of firms in the whole sample, at date  $t$ . Future earnings are measured here at 5-year horizons. The sample contains publicly listed nonfinancial firms from 19642 to 2010.



In sum, the result that S&P 500 price informativeness is rising, can be mostly explained by an increase in firm size, because larger firms are easier to price. While a compositional shift to larger firms can explain the upward trend of S&P 500 price informativeness, there is no downward trend in size to explain the fall in informativeness of the non-S&P 500 firms.

This leaves open the question of why small, non-S&P 500 firms face less informed investors, in a world when data has become so much more abundant.

## 2.2 The Role of High Tech

Part of the story of the decrease in price informativeness of the whole sample of firms is that the share of high-tech firms has increased over time and these high-tech firms are hard to price. At the same time, the S&P 500 has also become more tech-heavy. But the price information for these large tech firms is not more scarce relative to their non-tech counterparts. The data reveals that the combination of being small and high-tech depresses price information.

Figure 5: **Price informativeness for decile of R&D Intensity: S&P 500 vs Whole Sample.** Price informativeness is the ability to forecast future earnings (Eq 2). We run a separate regression for each year  $t = 1962, \dots, 2010$ , horizon  $k = 5$  and bin interval  $[1/10), \dots, [10/10]$  partitioned by 1/10 deciles. Firms are split by R&D intensity measured as the firm average R&D spending scaled by its assets. Price informativeness is the average value of  $\beta_{t,y} \times \sigma_{y,t} \log(M/A)$ , where  $\sigma_{y,t} \log(M/A)$  is the cross-sectional standard deviation of  $\log(M/A)$  in year  $t$  and R&D intensity interval  $y$ . Future earnings are measure here at 5-year horizons. The sample contains publicly listed nonfinancial firms from 1962 to 2010.

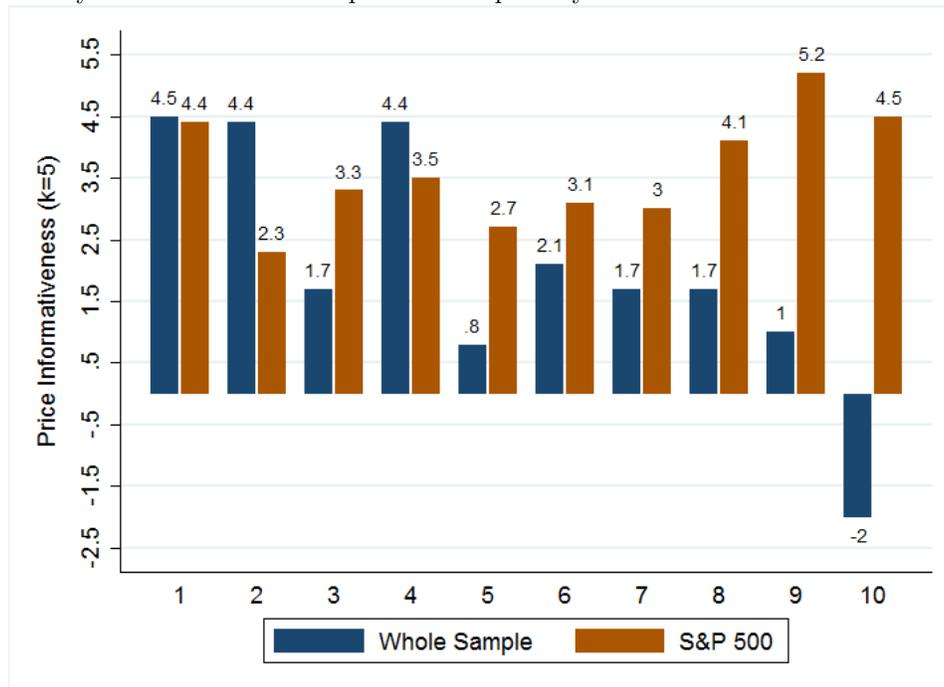
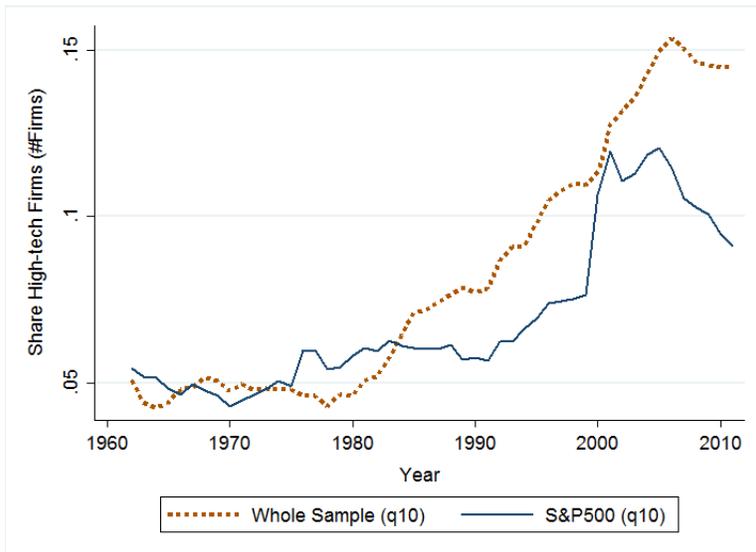


Figure 5 shows the average price informativeness for firms in the S&P 500 (orange bars) and firms in the whole sample (blue bars) by decile of R&D intensity (R&D spending scaled

by total asset). We estimate price informativeness by decile in the following way. First, we sort all observations in the full sample into decile of R&D intensity over the whole period, such that each decile has an equal number of observations. We then estimate price informativeness for each bin using the same method before. Second, for S&P 500 firms, we select out only the S&P 500 firms in each bins, such that we keep the same thresholds of R&D intensity for S&P 500 firms and for the whole sample. We then re-estimate the price informativeness of each bin on this sub-sample.

Figure 5 reveals two striking features. First, price informativeness of firms in the whole sample strongly declines with R&D intensity for firms above the 5th-decile. Second, this pattern disappears if we look at S&P 500 firms (the orange bars). In this case, the price information of the highest tech decile in the S&P 500 differs little from other S&P 500 firms and if anything, is slightly higher at the end of the R&D intensity distribution. Therefore, while high tech firms in the full sample have much less future earnings information impounded in their prices, this is not the case for S&P 500 firms.

Figure 6: **Share of High-Tech Firms based on Decile of R&D Intensity: S&P 500 vs Whole Sample.** We compare the average research intensity of firms that were ever in S&P 500 and firms there were never in S&P 500 over time. Research intensity is defined as a firm’s R& D annual expenditures, divided by the firm’s total assets. The sample contains publicly listed nonfinancial firms from 1960 to 2010.



Next, we calculate the fraction of firms in each tech decile, at each date. Figure 6 plots the share of firms in the whole sample that are in the top decile and the share of S&P 500

firms in that same top decile, at each date.

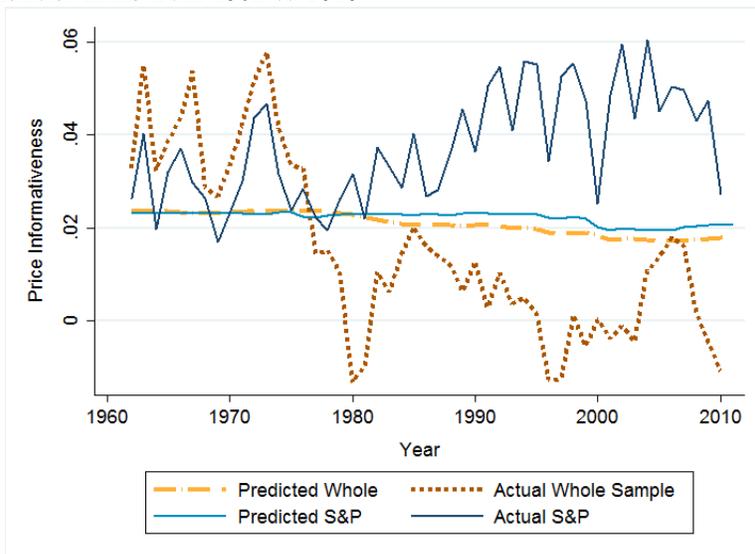
In both groups, the fraction of firms investing more in R&D is increasing steadily. The share of high-tech has grown slightly more rapidly in the full sample than in the S&P 500 sample. Until the early 80's, the high-tech shares for S&P 500 and non-S&P 500 firms track either other closely. Then, in the mid-80's trends diverge. The share of high-tech firms increases more in the whole sample, essentially driven by a rapid entry rate of tech firms (extensive margin), rather than an increase in R&D effort by incumbents (Begenau and Palazzo, 2017). Then, in the early 2000's, the share of tech firms in the S&P 500 increases and the shares converge again. In the last decade, tech shares diverge, with more entry of smaller tech firms in the whole sample. But for both samples, this trend toward more research or more tech suggests that firms, on average, should be getting harder to value.

To quantify how much this technology composition change can explain of the price informativeness trends, we do the same type of prediction exercise as we did in the last section, for size. At each date, we multiply the share of whole sample firms in each tech decile by the average price informativeness for that decile  $\beta_y$  and add them together. That gives us  $\hat{\beta}_t^{tech}$ , which is the degree of price informativeness that the tech composition alone would explain. Then, we do the same for only the S&P 500 firms. We calculate tech-predicted informativeness by multiplying the share of the S&P 500 that each tech bin comprises at each date, by the average informativeness of the S&P 500 firms that tech decile. Formally, tech-predicted informativeness  $\hat{\beta}_t^{tech}$ , is:

$$\hat{\beta}_t^{tech} = \sum_{y \in [1, \dots, 10]} \beta_y \times ShareFirms_{y,t}.$$

The  $\beta_y$  is the average price informativeness, for all firms belonging to R&D intensity decile  $y$ , R&D deciles are estimated using the whole panel of observations. The  $\beta_y$  coefficients are reported in Figure 5.  $ShareFirms_{y,t}$  is the fraction of firms in year  $t$  that belongs to R&D intensity decile  $y$ .

Figure 7: **Predicted PI based on High-Tech: S&P 500 and Whole Sample.** This figure shows the evolution of predicted and actual price informativeness for S&P 500 firms and the whole sample. For firms in the S&P 500, we show in the dark-blue line the coefficient  $\beta_t$  estimated from the cross-sectional forecasting regression defined in eqn 1. The orange dotted-line reports the same result when  $\beta_{t,5}$  is estimated for every listed firms (instead of restricting to S&P 500). The light-blue line and yellow dashed-line plot the evolution of the predicted  $\hat{\beta}a_t^{tech}$  computed in eqn 2.1.  $\hat{\beta}a_t^{tech}$  is the weighted sum of  $\beta_y$ , where  $y$  corresponds to a tech-decile and weights correspond to the fraction of firms in the same tech-decile in a given year. The light-blue line plot the evolution of  $\hat{\beta}a_t^{tech}$  use the fraction of S&P 500 firms in each tech bin, at each date  $t$ , as weights; the yellow dashed-line uses the fraction of whole sample firms in each tech bin as weights. Future earnings are measured here at 5-year horizons. The sample contains publicly listed nonfinancial firms from 1962 to 2010.



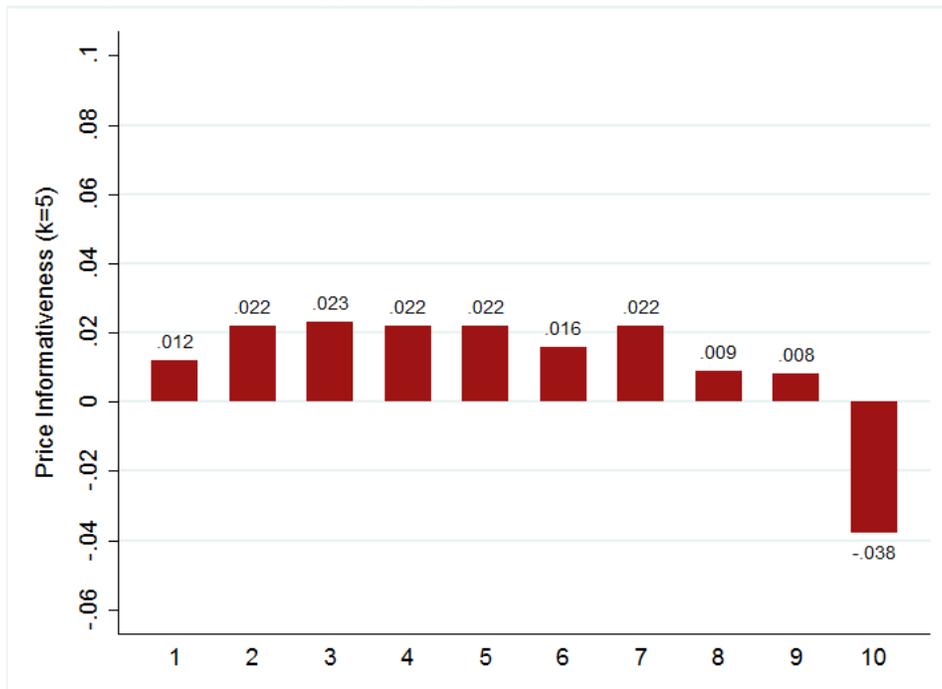
## 2.3 Ruling Out Potential Explanations

**Are stock prices more informative for less volatile firms?** Perhaps a change in the composition of high- and low-volatility firms can explain the divergence in S&P 500 and non-S&P 500 price informativeness. To examine this hypothesis, we define firm volatility as the standard deviation of its earnings (measured by EBIT) scaled by firm total assets. Then, we sort the whole panel of data into deciles of cash-flow volatility. We find that the correlation between size bins and volatility bins is indeed negative. In other words, larger firms tend to be less volatile. However, the correlation is very small. For instance, a firm in the largest decile of firm size has a two percentage point higher probability of being in the first (lowest) decile of volatility.

Another way to gauge the importance of volatility is to compute price informativeness for each of the ten volatility bins, as we did for size bins. As before, we run separate

cross-sectional regression as in Eq. 3 for each volatility bin. Figure 8 shows no difference in price informativeness across cash-flow volatility bins, with the exception of the highest decile, which displays a slightly lower level of price informativeness. This force is nowhere near strong enough to explain the large divergence in S&P 500 and non-S&P 500 price informativeness.

**Figure 8: Price Informativeness across Cash-Flow Volatility Bins.** Price informativeness is the ability to forecast future earnings (Eq 2). We run a separate regression for each year  $t = 1962, \dots, 2010$ , horizon  $k = 5$  and bin interval  $[1/10), \dots, [10/10]$  partitioned by 1/10 deciles. Firms are split by cash-flow volatility measured as the standard deviation of EBIT scaled by total asset. Price informativeness is the average value of  $\beta_{t,y} \times \sigma_{y,t} \log(M/A)$ , where  $\sigma_{y,t} \log(M/A)$  is the cross-sectional standard deviation of  $\log(M/A)$  in year  $t$  and volatility interval  $y$ . Future earnings are measure here at 5-year horizons. The sample contains publicly listed nonfinancial firms from 1962 to 2010.



**Is information flowing to S&P 500 industries?** One plausible explanation is that the market is getting better at pricing some types of firms. Perhaps health care or online firms were hard to price initially as they are more intensive in research and development, or some changes in industry-specific regulation made S&P 500 firms easier to price. These features are all highly correlated with a firm's industry. So, we begin by asking if the growth or decline in price informativeness is determined by an industry effect.

There are 253 different SIC3 codes in Compustat and 173 in S&P 500. The median number of firms per industry is 12, but the distribution is very skewed. Looking at the industries with strictly more than the median number of firms, we end up with only 24 distinct industries. We call these 24 industries *SPindustries*. Then, we restrict our sample only to firms in these 24 *SPindustries*. Within this restricted sample, we compare price informativeness trends of firms that appear in the S&P 500, at some point (542 firms) with those in the same industries, that do not (7,768 firms).

Non-S&P 500 firms in S&P 500 industries do not experience a rise in price informativeness. From 1962 to 2010, price informativeness for these firms falls from 0.03 to 0.01. S&P 500 firms in these same industries do experience the improvement in price efficiency. Over the same period, their trend price informativeness rises from 0.07 to 0.11.

If we do the same exercise with every industry represented in the S&P, instead of just the 24 most represented industries, we get similar results. This difference in price informativeness does not appear to be driven by differences in industries. This evidence suggests that the increase in price information for S&P 500 firms does not result from S&P 500 firms being in more informative industries.

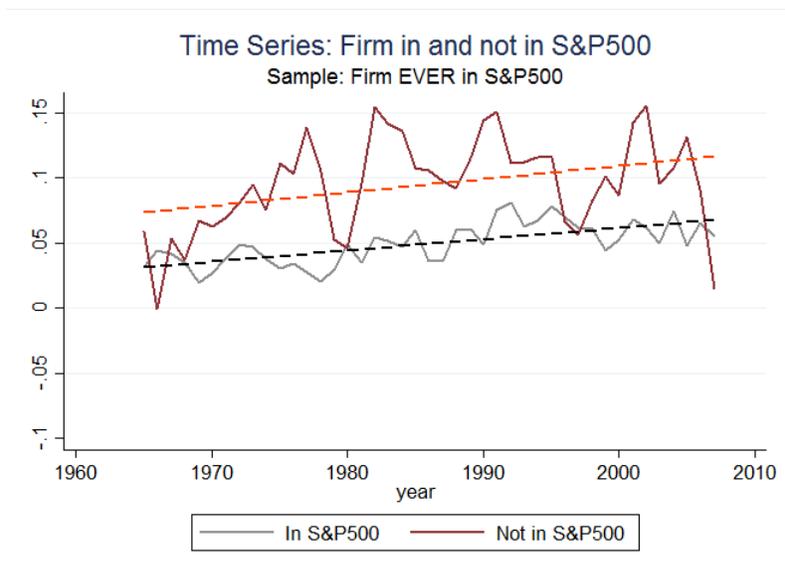
**Is information flowing specifically to firms currently in the S&P 500?** No, this does not seem to be a result about firms currently in the S&P 500 having greater price informativeness or a different trend. Instead, the rise in price informativeness seems to affect the type of firm that would be in the S&P 500. We show this result in two ways and then continue to investigate the question of what firm characteristics determine rising or declining price informativeness.

To look at the question of whether there is something specific to firms in the S&P 500, we perform two different tests. First, we look at firms, which at some point will be part of the S&P 500, and compare their price informativeness trend, (a) during the period where they are in the S&P 500; and (b) during the period where they are not. Second, we look at firms that share similar characteristics to S&P 500 firms but will never be part of the S&P 500 and compare their price informativeness trend to firms in the S&P 500 with the same characteristics.

For the first exercise, we estimate two separate regressions of Equation 1 for the period of the firm life when it is in the S&P 500 and when it is not. Figure 9 shows that, among the sample of firms that are in the S&P 500 at some point in their life, the trend in price informativeness is similar for firms currently in and out of the S&P 500. In levels, informativeness is actually higher when a firm is not in the S&P 500, than when they are in.

**Figure 9: Price Informativeness Trend While in and out of S&P 500 is Similar.**

The sample for both lines contains publicly listed nonfinancial firms that have been in the S&P 500 at some time between 1962 and 2010. The grey line (bottom) is the firms currently in the S&P 500, at the date listed on the x-axis. The red line (top) is firms not currently in the S&P 500. The black and red dashed lines are linear trends that fit the grey and red time trends, respectively. Price informativeness is obtained separately for each group by running the forecasting regression (eqn 1) for horizon  $k = 5$  and calculating the product of the forecasting coefficient and the cross-sectional standard deviation of market prices in year  $t$  using eqn 2.



For the second exercise, we want to investigate whether firms with similar characteristics have similar changes in their stock price informativeness. We proceed in two steps. First, for the universe of listed firms every year, we estimate the probability of being part of the S&P 500. To do so, we construct a dummy variable  $SP500_{i,t}$ , which takes the value of one if firm  $i$  is in the S&P 500 at time  $t$  and zero otherwise. Then, we estimate  $\alpha$ ,  $\delta$ ,  $\phi$  and  $\gamma$  in the following equation:

$$SP500_{i,t} = \alpha_i + \delta_t + \phi_t \log(M/A) + \gamma_t \log(Asset) + \epsilon_{i,t} \quad (4)$$

We then use the estimates of  $\alpha$ ,  $\delta$ ,  $\phi$  and  $\gamma$  to construct predicted probabilities of being in

the S&P 500. We denote this probability as  $\widehat{SP500}_{i,t}$ .

Second, we partition the sample into firms *similar* to S&P 500 firms and firms *actually* in the S&P 500 and compute price informativeness for each subsample using Equation 2. The median score of  $\widehat{SP500}_{i,t}$  for firms in the S&P 500 is around 0.6. Therefore, we restrict the sample to *all* firms higher than this threshold. This leaves us with 3,105 distinct firms, among which 60% will be indeed at some point in their life in the S&P 500 and 40% that will not. We call firms not in the S&P 500 but with a  $\widehat{SP500}_{i,t} \geq 0.6$  firms *similar* to S&P 500 firms.

We find that firms that will never be in the S&P 500 but are relatively close in terms of market capitalization and size exhibit a nearly identical rise in price informativeness to the S&P 500 firms. While the level of price informativeness is somewhat different, we learn that there is something about the type of firm in the S&P 500, the size or book-to-market, that draws in more analysis over time. For some reason, firms with similar characteristics that will never be S&P 500 firms have lower informativeness levels, but a similar informativeness growth rate.

**Do the informativeness trends reflect changes in institutional ownership?** The idea that price informativeness rises when more institutional owners hold the asset is appealing, and supported by our data. However, the time trends in institutional ownership of S&P 500, relative to non-S&P 500 firms are not consistent with the changes price informativeness we observe. Institutional owners are better at pricing assets. But they don't explain the trends we see.

For sure, assets that institutions hold in abundance have prices that better predict future earnings. In our data set, the 500 firms with the highest level of institutional ownership have a price informativeness measure that is roughly three times that of the rest of the sample (0.04 vs. 0.01). However, when we estimate the effect of institutional ownership and control for it, we still find that S&P 500 price informativeness increases, while the rest of the sample declines.

In summary, our analysis has uncovered the following trends: Price informativeness rose for the large firms in the S&P 500. For the rest of the sample, price informativeness declined.

Informativeness rose even more for high-tech, S&P 500 firms and fell more for small, high-tech firms. But this difference in tech intensity explains only a small amount of the price informativeness divergence. This is primarily a firm size effect.

The question of whether these facts are consistent with efficient use of data is really a question about what rational agents, who choose data allocations, should or should not choose to process data about. To answer these questions, it is useful to set up a model. The model explains what makes a firm's data valuable and predicts how the decisions to acquire and process data should change over time, as large firms grow and data becomes more abundant. Therefore, the next section sets up, solves and explores the properties of a model of data choice and portfolio choice.

### **3 A Model to Interpret Patterns In Price Information**

The data reveal two opposing trends: an increase in the informativeness of S&P 500 firms, that appears to be driven by a composition effect, and a decline in the informativeness of small firms, especially small high-tech firms. Given the growth in computing speed, data availability, and human resources devoted to the financial sector, it is puzzling that many firms are priced less accurately today than such firms were in the past. Why wouldn't rising data processing ability lift all price informativeness? Furthermore, why does the trend to more high-tech firms affect large and small firms differently?

We explore a simple explanation for these facts: Investors are rationally allocating their data processing ability, in the face of growing data processing technology and changing firm characteristics. We use a model of data allocation and portfolio choice to show that this explanation is consistent with our stylized facts. The key insight we get from this model is that the growth in large firm size can draw data analysis away from small firms.

To explore data allocation and its effect on prices, we need an equilibrium model with multiple assets and agents who choose how much data to process about each asset. If we assumed, exogenously, that information processing is directed at particular assets, it would not explain why some prices are becoming more efficient and others are not. Instead, we adapt the framework of Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) to predict

where data should flow. After adding a big data specific constraint and growing large firms, the model teaches us about how a profit maximizing investor should use data processing and invest, and how this should affect the information contained in equilibrium prices.

**Assets** The model features 1 riskless and  $n$  risky assets. The price of the riskless asset is normalized to 1 and it pays off  $r$  at the end of each period. One share of a risky asset is a claim to the random payout  $\tilde{d}_{jt}$  at the end of the period. For simplicity, we assume that these asset payoffs are independent:  $\tilde{d}_{jt} \sim iidN(\mu, \Sigma)$ . The riskless asset pays a known amount  $1 + r$  at the end of the period.

There are  $n$  risky assets, one for each of the firms in the economy. Each share of a risky asset  $j$  is a claim to the payoff  $\tilde{d}_{jt}$ . Each risky asset has a stochastic supply given by  $\bar{x}_j + \bar{x}_{jt}$ , where noise  $\bar{x}_{jt}$  is normally distributed, with mean zero, variance  $\sigma_x$ , and no correlation with other noises: the vector of  $\bar{x}_{jt}$ 's is  $\bar{x}_t \sim \mathcal{N}(0, \sigma_x I)$ . As in most noisy rational expectations equilibrium model, the supply is random to prevent the price from fully revealing the information of informed investors. This randomness might be interpreted as investors in the market trading for hedging reasons that are unrelated to information, as in Manzano and Vives (2010).

**Portfolio Choice Problem** Each period, a new continuum of atomless investors is born. Each investor is endowed with initial wealth,  $W_0$ .<sup>9</sup> They have mean-variance preferences over ex-post wealth, with a risk-aversion coefficient  $\rho$ . Let  $E_i$  and  $V_i$  denote investor  $i$ 's expectations and variances conditioned on all interim information, which includes prices and signals. Thus, investor  $i$  chooses how many shares of each asset to hold,  $q_{it}$  to maximize interim expected utility,  $\hat{U}_{it}$ :

$$\hat{U}_{it} = \rho E[W_{it} | \mathcal{I}_{it}] - \frac{\rho^2}{2} V[W_{it} | \mathcal{I}_{it}] \quad (5)$$

subject to the budget constraint:

$$W_{it} = rW_0 + q_{it}(\tilde{d}_t - p_t r), \quad (6)$$

---

<sup>9</sup>Since there are no wealth effects in the preferences, the assumption of identical initial wealth is without loss of generality. The only consequential part of the assumption is that initial wealth is known.

where  $q_{it}$  and  $p_t$  are  $n \times 1$  vectors of prices and quantities of each asset held by investor  $i$ .

**Data Processing Choice** Investors can acquire information about asset payoffs  $\tilde{d}_t$  by processing digital data. Digital data is coded in binary code. Investors face a constraint  $\bar{B}_t$  on the total length of the binary code they can process. This constraint represents the frontier information technology in period  $t$ . One  $\{0, 1\}$  digit encodes 1 bit of information.<sup>10</sup> Thus units of binary code length are bits.

All data processing is subject to error. The most common model of processing error is the parallel Gaussian channel.<sup>11</sup> For a Gaussian channel, the number of bits required to transmit a message is related to the signal-to-noise ratio of the channel. Clearer signals can be transmitted through the channel, but they require more bits. The relationship between bits and signal precision for a Gaussian channel is  $bits = 1/2 \log(1 + \text{signal-to-noise})$  (Cover and Thomas (1991), theorem 10.1.1). The signal-to-noise is the ratio of posterior precision to prior precision.

Investors choose how to allocate their capacity among  $n$  risky assets. Let  $b_{it}$  be a vector whose  $j$ th entry,  $b_{it}(j) > 0$ , is the number of bits processed by agent  $i$  at time  $t$  about  $\tilde{d}_{jt}$ . Let  $\eta_{it}^b$  represent the realized string of zeros and ones that investor  $i$  observes. The data processing constraint is then

$$\sum_{j=1}^N b_{it}(j) \leq \bar{B}_t \quad \text{where } b_{it}(j) \geq 0 \quad \forall i, j, t. \quad (7)$$

**Information sets and equilibrium** The information set the investor has when he makes investment decisions is  $\mathcal{I}_t = \{\mathcal{I}_{t-1}, \eta_{it}^b, p_t\}$ . The ex-ante information set includes the entire sequence of data processing capacity:  $\mathcal{I}_0 \supset \{\bar{B}_t\}_{t=0}^{\infty}$ .

An equilibrium is a sequence of bit string lengths choices,  $\{b_{it}\}$  and portfolio choices  $\{q_{it}\}$  by investors such that

---

<sup>10</sup>A byte is 8 bits, which allows for 256 possible sequences of zeros and ones, enough for one byte to describe an alpha-numeric character or common keyboard symbol. Megabytes are  $10^6$  bytes. If your computer can store 1GB in its RAM, that is  $10^9$  bytes, or a binary code of length  $8 \times 10^9$ .

<sup>11</sup>As Cover and Thomas (1991) explain, “The additive noise in such channels may be due to a variety of causes. However, by the central limit theorem, the cumulative effect of a large number of small random effects will be approximately normal, so the Gaussian assumption is valid in a large number of situations.”

1. Investors choose bit string lengths  $b_{it} \geq 0$  to maximize  $E[\hat{U}_{it} | \mathcal{I}_{t-1}^+]$ , where  $\hat{U}_{it}$  is defined in (5), taking the choices of other agents as given. This choice is subject to (7).
2. Investors choose their risky asset investment  $q_{it}$  to maximize  $E[U(c_{it}) | \eta_{fit}, p_t]$ , taking asset prices and others' actions as given, subject to the budget constraint (6).
3. At each date  $t$ , the vector of equilibrium prices  $p_t$  equates aggregate demand (left side) with supply (right) to clear the market:

$$\int_i q_{ijt} di = \bar{x}_{jt} + x_{jt}, \quad (8)$$

### 3.1 Solving the Model

We solve the model in four steps. We sketch each step here and relegate details to the appendix for the interested reader. Because units of signal precision are easier to work with than bits, we define  $K_{ijt}$  to be the precision of the signal  $\eta_{ijt}$  inferred from the data processed by investor  $i$  about firm  $j$  at time  $t$ . Let  $K_{it}$  be the diagonal matrix with  $K_{ijt}$  on its  $j$ th diagonal and  $\eta_{it}$  be the vector of all signals observed by  $i$ . Finally, let  $\bar{K}_t \equiv \int_i K_{it} di$  be the matrix of the average investors' signal precision.

**Step 1: Bayesian updating.** There are three types of information that are aggregated in time-2 posteriors beliefs: prior beliefs, price information, and (private) signals from data processing. We begin with price information. We conjecture and later verify that a transformation of prices  $p_t$  generates an unbiased signal about the risk factor payoffs,  $\eta_{pt} = \tilde{d}_t + \epsilon_{pt}$ , where  $\epsilon_p \sim N(0, \Sigma_p)$ , for some diagonal variance matrix  $\Sigma_p$ .

Next, we construct a single signal that encapsulates the information conveyed in bit strings. Recall that in a Gaussian channel with prior information precision  $\Sigma^{-1}$ , the number of bits required to transmit a signal with a given precision  $K_{it}$  is  $bits = 1/2 \cdot \log(1 + \Sigma K_{it})$ . The data contains the true value of  $\tilde{d}_t$ . But data processing is imperfect and introduces Gaussian noise. Processed fundamental data is observed as  $\eta_{fit} = \tilde{d}_t + \tilde{\epsilon}_{fit}$ , where the channel (data processing) noise is a normal, random variable:  $\tilde{\epsilon}_{fit} \sim N(0, K_{it}^{-1})$ . Substituting this mapping into (7) yields a new data processing constraint in terms of signal precisions  $K_{it} \geq 0$ :

$$\sum_{j=1}^N \log(1 + \Sigma K_{ijt}) \leq 2\bar{B}_t. \quad (9)$$

Finally, Bayes' law tells us how to combine price signals, data signals and prior beliefs. The resulting posterior beliefs about  $z$  are normally distributed with variance  $\text{var}[\tilde{d}_t|\mathcal{I}_{it}] = (\Sigma^{-1} + \Sigma_p^{-1} + K_{it})^{-1}$  and mean

$$E[\tilde{d}_t|\mathcal{I}_{it}] = \text{var}[\tilde{d}_t|\mathcal{I}_{it}](\Sigma^{-1}\mu + K_{it}\eta_{it} + \Sigma_p^{-1}\eta_{pt}). \quad (10)$$

**Step 2: Solve for the optimal portfolios, given information sets and issuance.**

Substituting the budget constraint (6) into the objective function (5) and taking the first-order condition with respect to  $q_{it}$  reveals that optimal holdings are increasing in the investor's risk tolerance, precision of beliefs, and expected return:

$$q_{it} = \frac{1}{\rho} \text{var}[\tilde{d}_t|\mathcal{I}_{it}]^{-1} (E[\tilde{d}_t|\mathcal{I}_{it}] - p_t r). \quad (11)$$

**Step 3: Clear the asset market.** Substitute each agent  $j$ 's optimal portfolio (11) into the market-clearing condition (8). Collecting terms and simplifying reveals that the vector of equilibrium asset prices are linear in payoff risk shocks and in supply shocks:

$$p_t r_t = A_t + C_t \tilde{d}_t + D_t \tilde{x}_t \quad (12)$$

where  $\tilde{d}_t$  is the vector of expected dividends and  $\tilde{x}_t$  is the vector of asset supply shocks at time  $t$ . Coefficients  $A_t$ ,  $C_t$ , and  $D_t$  are in the Appendix.

**Step 4: Solve for data processing choices.** The information choice objective comes from substituting in the optimal portfolio choice and equilibrium price rule, and then taking the ex-ante expectation over the signals and price that are not yet observed at the start of the period. This yields an objective that is linear in signal precisions:

$$\max_{K_{i1t}, \dots, K_{int} \geq 0} \sum_{j=1}^n \Lambda(\bar{K}_{jt}, \bar{x}_j) K_{ijt} + \text{constant} \quad (13)$$

$$\text{where } \Lambda(\bar{K}_{jt}, \bar{x}_j) = \bar{\Sigma}_j[1 + (\rho^2/\tau_x + \bar{K}_{jt})\bar{\Sigma}_j] + \rho^2\bar{x}_j^2\bar{\Sigma}_j^2, \quad (14)$$

and  $\bar{\Sigma}_j^{-1} = \int \hat{\Sigma}_{ti}^{-1}(j, j) di$  is the average precision of posterior beliefs about asset  $j$ . Its inverse, average variance  $\bar{\Sigma}_j$  is decreasing in  $\bar{K}_{jt}$ . The appendix shows two important properties. The first is strategic substitutability in data choices:  $\partial\Lambda(\bar{K}_{jt}, \bar{x}_j)/\partial\bar{K}_{jt} < 0$ . The second is returns to asset scale in data processing:  $\partial\Lambda(\bar{K}_{jt}, \bar{x}_j)/\partial\bar{x}_j > 0$ .

Maximizing a weighted sum (13) subject to a concave constraint (9) yields a corner solution. The investor optimally processes data about only one asset. Which asset to learn about depends on which has the highest marginal utility  $\Lambda(\bar{K}_{jt}, \bar{x}_j)$ . If there is a unique asset  $j^* = \operatorname{argmax}_j \Lambda(\bar{K}_{jt}, \bar{x}_j)_j$ , then the solution is to set  $K_{i,j^*,t} = \Sigma^{-1}(e^{\bar{B}_t} - 1)$  and  $K_{ilt} = 0$ ,  $\forall l \neq j^*$ . But when capacity  $\bar{B}_t$  is high enough, there will exist more than one asset  $j$  that is learned about. Let  $\mathcal{M}_t \equiv \{\bar{K}_{jt} > 0\}_{j=1}^n$  be the set of assets learned about. Then an equilibrium is a set of average precisions for each asset  $j$ ,  $\{\bar{K}_{jt}\}_{j=1}^n$  such that

$$\Lambda(\bar{K}_{jt}, \bar{x}_j) = \bar{\Lambda} \quad \forall j \in \mathcal{M}_t \quad (15)$$

In this equilibrium, investors are indifferent about which single asset  $j \in \mathcal{M}$  to learn about. But the aggregate allocation of data processing is unique (Kacperczyk, Van Nieuwerburgh, and Veldkamp, 2016).

### 3.2 Understanding Trends in Price Informativeness

We start by setting up the puzzle that motivates the paper. If faster computers can process ever more data over time, why haven't all firms prices benefited from the increase in price informativeness? We show that, although investors prefer to learn about large firms, more data does not make them want to learn less about small firms. Instead, all firms should experience an increase in price informativeness. Thus an increase in the efficiency of the financial sector in processing data is not a complete explanation for the trends in the data.

The second set of results offers a solution. It shows that if large firms grow larger, as they do in our data, this trend alone can explain the composition effect driving up S&P 500 price informativeness, as well as a decline in the informativeness of smaller assets' prices. What we learn from this is that our empirical findings do not imply ineffectual, inefficient or irrational

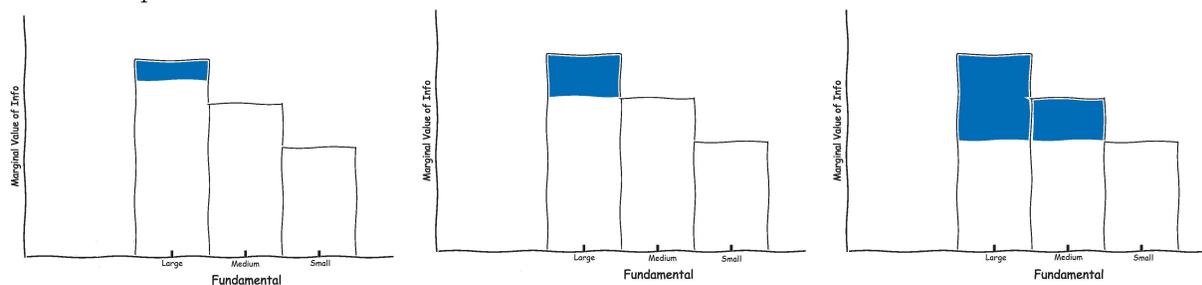
use of technology. Because large firms are growing, modest growth in data technology can be reconciled with the deterioration of small firms' price information.

**Big Data Alone Should Increase Informativeness of All Prices** If investors particularly like processing data about large firms, then perhaps when they have more data processing ability, they direct it towards these large firms. That turns out not to be the case. The next result shows why growth in *data processing alone cannot explain* the facts about price informativeness.

In many cases, after all data processing capacity is allocated, there will be multiple risks with identical  $\Lambda(\bar{K}_{jt}, \bar{x}_j)$  weights. That is because the marginal utility of signal precision,  $\Lambda_i$ , is decreasing in the average information precision  $\bar{K}_i$ . In this case, investors are indifferent about which risk to learn about. When financial data processing efficiency  $\bar{B}_t$  rises, more bits are allocated to all the assets in this indifference class.

**Lemma 1 Technological progress: Intensive Margin.** *As  $\bar{B}_t$  grows, the average investor learns weakly more about every asset  $j$ ,  $\int K_{ij(t+1)} di \geq \int K_{ijt} di$ , with strict inequality for all assets that are learned about:  $\int K_{ij(t+1)} di > \int K_{ijt} di \quad \forall j : K_{ijt} > 0$  for some  $j$ .*

Figure 10: **Equilibrium Allocation of Data Processing.** Shaded area represents aggregate allocation of data processing. Moving from left to right represents an increase in data processing capacity. More processed data lowers the marginal utility of additional data processing. That causes data on other assets to be processed.



This type of equilibrium is called a “waterfilling” solution (Cover and Thomas, 1991). Figure 10 illustrates how the equilibrium allocation maintains indifference (equal marginal utility) between all assets being learned about. The equilibrium uniquely pins down which risk factors are being learned about in equilibrium, and how much is learned about them, but

not which investor learns about which risk factor. Waterfilling arises in other information choice problems, such as Kacperczyk, Nosal, and Stevens (2015).

**Lemma 2 *Extensive Margin: With more bits, more assets are learned about.*** *If  $\bar{x}_i$  is sufficiently large  $\forall i$ , the set of assets learned about  $\mathcal{M}_t$  does not contain all assets, and  $\bar{B}_{t+1} - \bar{B}_t$  is sufficiently large, then the set of assets  $\mathcal{M}_{t+1}$  learned about in  $t + 1$  is larger than the set  $\mathcal{M}_t$ .*

A key force is strategic substitutability in information acquisition, an effect first noted by Grossman and Stiglitz (1980). The more other investors learn about a risk, the more informative prices are and the less valuable it is for other investors to learn about the same risk. If one risk has the highest marginal utility for signal precision, but capacity is high, then many investors will learn about that risk, causing its marginal utility to fall and equalize with the next most valuable risk. With more capacity, the highest two  $\Lambda(\bar{K}_{jt}, \bar{x}_j)$ 's will be driven down until they equal the next highest  $\Lambda$ , and so forth. But when  $K$  increases, the marginal utilities of all risks must remain equated. Since learning about any risk reduces its marginal utility, all risks must have weakly more learning about them so that all their marginal utilities remain equal and the economy stays at an optimum.

**When large firms get larger, the informativeness of small firm prices falls.** The growth in data processing is not the only trend that has changed information processing incentives. At the same time, there has been a change in firm size. It's the growth of large firms that can explain why the informativeness of small firms has not grown. For now, we hold  $\bar{B}_t$  fixed and only consider the change in firm size. After we have explored each force separately, we consider their combined effect.

The following result shows that if an asset grows larger, investors process more data about it, on average. But for other assets whose size does not grow, the amount of data processed falls.

**Lemma 3 *When Big Firms Grow, Small Firm Data Analysis Declines.*** *For  $K_t$  and  $\bar{x}_j$  sufficiently high, an increase in the size of firm  $j$  increases the amount learned about  $j$*

and reduces the amount learned about all other assets:  $\partial \bar{K}_{jt} / \partial \bar{x}_j > 0$  and  $\partial K_i / \partial \bar{x}_j \leq 0 \forall i \neq j$ .

The marginal value of signal precision  $\bar{K}_{jt}$  is  $\Lambda(\bar{K}_{jt}, \bar{x}_j)$ , from (14). Recall that  $\partial \Lambda(\bar{K}_{jt}, \bar{x}_j) / \partial \bar{x}_j = \rho^2 \bar{\Sigma}_j^2 > 0$ . So, larger assets are always more valuable targets for data processing. Next consider the equilibrium data allocation. Equation 14 implies that more capacity is allocated to the larger asset in equilibrium as well.

The fall in data processed about other firms is the consequence of more data about  $j$  and a fixed budget for bits of data. If more bits are processed about  $j$ , less bits must be processed for some other asset. That decline in bits processed is equally spread across other assets so as to equate the marginal utility of bit processing for all.

Fundamentally, this preference for more data about larger assets comes from the fact that information has increasing returns to scale. A larger asset will be a larger fraction of an average investor's portfolio. One could use all data to learn about a small fraction of one's portfolio value. But that is not as valuable as using data processing to reduce uncertainty about an asset that represents a large fraction of one's portfolio risk and a large fraction of potential profit. The same bit of data can evaluate 1 share or 1000 shares equally well. That makes data that can be applied to many units of asset value – data on large firms – more valuable.

**Social welfare.** Underlying the paper is the presumption that higher price informativeness is valuable. While there are many mechanisms that justify that link, one might question whether the improvement in some funds' information can compensate for others' decline. In Appendix C, we explore a model where entrepreneurs' exert more effort in firms whose prices reflect more accurate valuation information. In such a world, investors learn too much about large firms, relative to what a planner would choose. But the extent of the social cost depends on how future computing evolves. The gap between individual incentives and the social optimum will be influenced by how much integrated computing creates efficiency returns to scale in information processing.

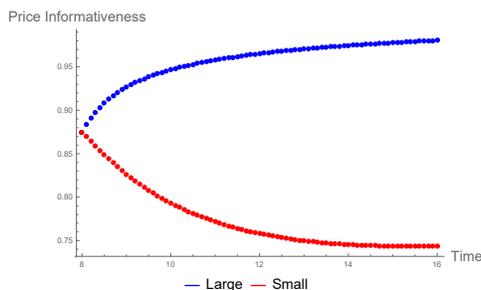
### 3.3 Numerical Example

To provide a visual representation of our results, we consider a two-firm numerical example. We explore the effects of an exogenous increase in data processing and large firm size.

For parameters, risk aversion is  $\rho = 4$ ; the inverse variance of the dividend payoff is  $\Sigma^{-1} = 1$ ; the inverse variance of asset supply shock is  $\tau_x = 3$ . To think about firm size effects, we need a large firm that grows, relative to a small firm.  $\bar{x}_1$  and  $\bar{x}_2$  both start at 1. But  $\text{bar}x_1$ , the large firm grows by 0.1 each period. The small firm size stays constant. The total data processing capacity grows at a constant rate each period, starting at  $K_t = 8$  and ending at  $K_t = 16$ . We did a handful of robustness checks by varying parameters within an order of magnitude and found qualitatively similar results.

Figure 11 shows that when both data processing and large firm size grow, the model can explain the divergence in S&P 500 (large firm) and non-S&P 500 (small firm) price informativeness. Of course, this does not prove that the model is correct or that every parameterization of the model can explain this result. In this example, the key is that data processing growth is slow enough that the amount of data processing about the small firm declines. When data processing about small firms declines, the informativeness of small firm prices declines as well. Thus, this numerical example demonstrates the possibility that this model can speak to the rise in price informativeness of S&P 500 firms and the decline of price informativeness among non-S&P 500 firms, observed in the data.

Figure 11: **Optimal Data Choices Can Explain Informativeness Divergence.** The figure plots the weight on dividend innovations from the price equation,  $C_t$  in eq. (12), from a 2-asset model where  $\rho = 4$ ,  $\Sigma^{-1} = 1$ ,  $\tau_x = 3$ ,  $\bar{x}_1$  starts at 4,  $\bar{x}_2$  starts at 1, and both grow by 0.1 each period.  $K_t$  increases linearly from 8 to 16.



**Information Choices about High-Tech Firms** Since the effect of high-tech firms was not central to explaining the main fact of the paper – the divergence in price informativeness – we did not incorporate it in the model. However, in results available on request, we model high-tech firms as firms with higher payoff uncertainty. We find that for large high-tech firms, size dominates the data choice, and informativeness rises. But for small, high-tech firms, not only does the small size cause these firms’ data to be crowded out, but the effect for high-tech firms is even larger than for low-tech ones. Although the quantitative effect of high-tech in the aggregate informativeness trend was small, it is there, and the same model could explain why.

Of course, none of this proves that these are the forces responsible for this trend. This model is only meant to show that the decline in informativeness for small firms need not imply irrationality or financial frictions. The model does offer one coherent way to think through the kinds of forces that might be at work for optimizing agents to produce the price information trends we document.

## 4 Conclusion

Technology and new ways to use data are transforming financial markets. How might this affect asset prices? Since new technology is primarily information technology, we look for evidence that the information content of prices is changing. It appears that big data technology has uneven effects on large and small firms.

Our data reveal that price informativeness is trending upward, but only for large firms. For others, there is a stagnation, or even slight decline. Thus, the rise in price informativeness of S&P 500 firms appears to be driven by a composition effect, as the firms in the S&P 500 become larger. These larger firms have more informative financial prices. But the informativeness of a firm of a given size has not increased perceptibly.

To understand these facts, we use a model to explore the logical consequences of two long-run trends. One trend is an increase in the efficiency of data processing over time. That is important to consider because when people talk about the financial sector becoming more efficient, often that is associated with greater information efficiency, suggesting the

more information is being discovered, processed and aggregated. The second trend is the well-documented fact that large firms are growing larger. The trend holds also for the publicly-listed firms in our sample. Our model clarifies why such large firms' data is more valuable to process. As they grow larger, investors' optimal allocation of data processing shifts towards these growing firms. As more data is processed and used by investors to trade, the price informativeness of such firms rises. The combination of the two forces can explain the joint evolution of large and small firm price informativeness.

What we learn is that technology does not have a uniform effect on all firms. Like with any technological change, there are winners and losers. Our paper helps explain who wins, who loses, and why.

## References

- ANDRIES, M., AND V. HADDAD (2017): “Information Aversion,” NBER Working Paper.
- BAI, J., T. PHILIPPON, AND A. SAVOV (2016): “Have Financial Markets Become More Informative?,” *Journal of Financial Economics*, 122 (35), 625–654.
- BOND, P., A. EDMANS, AND I. GOLDSTEIN (2012): “The Real Effects of Financial Markets,” *Annual Review of Financial Economics*, 4(1), 339–60.
- BOND, P., AND H. ERASLAN (2010): “Information-based trade,” *Journal of Economic Theory*, 145(5), 1675–1703.
- CHORDIA, T., C. GREEN, AND B. KOTTIMUKKALUR (2016): “Rent Seeking by Low Latency Traders: Evidence from Trading on Macroeconomic Announcements,” Working Paper, Emory University.
- COVER, T., AND J. THOMAS (1991): *Elements of information theory*. John Wiley and Sons, New York, New York, first edn.
- CROUZET, N., I. DEW-BECKER, AND C. NATHANSON (2016): “A Model of Multi-Frequency Trade,” Northwestern University Working Paper.
- DAVID, J., H. HOPENHAYN, AND V. VENKATESWARAN (2016): “Information frictions, Misallocation and Aggregate Productivity,” *Quarterly Journal of Economics*, 131 (2), 943–1005.
- DAVILA, E., AND C. PARLATORE (2016): “Trading Costs and Informational Efficiency,” NYU Working Paper.
- 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.
- DUGAST, J., AND T. FOUCAULT (2016): “Data Abundance and Asset Price Informativeness,” Working Paper, HEC.
- DURNEV, A., R. MORCK, AND B. YEUNG (2004): “Value-enhancing capital budgeting and firm-specific stock return variation,” *The Journal of Finance*, 59(1), 65–105.
- EDMANS, A., S. JAYARAMAN, AND J. SCHNEEMEIER (2016): “The source of information in prices and investment-price sensitivity,” *Journal of Financial Economics*, Forthcoming.
- FARBOODI, M., AND L. VELDKAMP (2017): “Long Run Growth of Financial Technology,” Discussion paper, National Bureau of Economic Research.

- GOLDSTEIN, I., E. OZDENOREN, AND K. YUAN (2013): “Trading frenzies and their impact on real investment,” *Journal of Financial Economics*, 109(2), 566–82.
- GROSSMAN, S., AND J. STIGLITZ (1980): “On the impossibility of informationally efficient markets,” *American Economic Review*, 70(3), 393–408.
- HSIEH, C., AND P. KLENOW (2009): “Misallocation and Manufacturing TFP in China and India,” *Quarterly Journal of Economics*, 124(4), 1403–1448.
- KACPERCZYK, M., J. NOSAL, AND L. STEVENS (2015): “Investor Sophistication and Capital Income Inequality,” Imperial College Working Paper.
- KACPERCZYK, M., S. VAN NIEUWERBURGH, AND L. VELDKAMP (2016): “A Rational Theory of Mutual Funds’ Attention Allocation,” *Econometrica*, 84(2), 571–626.
- KATZ, M., H. LUSTIG, AND L. NIELSEN (2017): “Are Stocks Real Assets? Sticky Discount Rates in Stock Markets,” *The Review of Financial Studies*, 30(2), 539–587.
- KYLE, A., AND J. LEE (2017): “Toward a Fully Continuous Exchange,” SSRN Working Paper.
- MAĆKOWIAK, B., AND M. WIEDERHOLT (2009): “Optimal sticky prices under rational inattention,” *American Economic Review*, 99 (3), 769–803.
- MANZANO, C., AND X. VIVES (2010): “Public and Private Learning from Prices, Strategic Substitutability and Complementarity, and Equilibrium Multiplicity,” CEPR Discussion Papers 7949, C.E.P.R. Discussion Papers.
- MONDRIA, J., T. WU, AND Y. ZHANG (2010): “The determinants of international investment and attention allocation: Using internet search query data,” *Journal of International Economics*, 82 (1), 85–95.
- OZDENOREN, E., AND K. YUAN (2008): “Feedback Effects and Asset Prices,” *The Journal of Finance*, 63(4), 1939–1975.
- PÁSTOR, L., AND R. F. STAMBAUGH (2012): “On the size of the active management industry,” *Journal of Political Economy*, 120, 740–781.
- RANCO, G., D. ALEKSOVSKI, G. CALDARELLI, M. GRGAR, AND I. MOZETIC (2015): “The Effects of Twitter Sentiment on Stock Price Returns,” PLOS working paper.
- RESTUCCIA, D., AND R. ROGERSON (2013): “Misallocation and Productivity,” *Review of Economic Dynamics*, 16, 1–10.
- SIMS, C. (2003): “Implications of rational inattention,” *Journal of Monetary Economics*, 50(3), 665–90.

STAMBAUGH, R. (2014): “Presidential Address: Investment Noise and Trends,” *Journal of Finance*, 69 (4), 1415–1453.

VAN NIEUWERBURGH, S., AND L. VELDKAMP (2009): “Information immobility and the home bias puzzle,” *Journal of Finance*, 64 (3), 1187–1215.

## A Appendix

## B Model Solution Details

*Price Coefficients:* Equating supply and demand from (11), we get

$$\int \frac{1}{\rho} \hat{\Sigma}_t^{-1} \left( E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r \right) = \bar{x} + x_t \quad (16)$$

where  $\hat{\Sigma}_t^{-1} \equiv \text{Var}[\tilde{d}_t | \mathcal{I}_{it}]$ . If we then substitute in the conditional expectation from (10), and use the definition of the price signal  $\eta_p = B_t^{-1}(p_t r - A_t)$ , we obtain

$$p r = \hat{\Sigma}_t \left[ \Sigma^{-1} \mu + \int K_{it} \eta_{it} di + \Sigma_p^{-1} B_t^{-1} (p_t r - A_t) - \rho (\bar{x} + x_t) \right] \quad (17)$$

Notice the price  $p$  on the left and right side of the equation. The term on the right is from the fact that agents use price as a signal. Next, we collect terms in  $p$ . We also use the fact that since signals are unbiased, irrespective of precision,  $\int K_{it} \eta_{it} di = \bar{K}_t \tilde{d}_t$ . The resulting equation is (12), where

$$A_t = \hat{\Sigma}_t (\Sigma^{-1} \mu - \rho \bar{x}) \quad (18)$$

$$C_t = I - \hat{\Sigma}_t \Sigma^{-1} \quad (19)$$

$$D_t = -\hat{\Sigma}_t \left( \rho I + \frac{1}{\rho \sigma_x^2} \bar{K}_t \right) \quad (20)$$

where  $\bar{K}_t \equiv \int K_{it} di$  Price information is the signal about the payoff vector  $\tilde{d}_t$  contained in prices. The transformation of the price vector  $p_t$  that yields an unbiased signal about  $\tilde{d}_t$  is  $\eta_p \equiv B_t^{-1}(p_t r - A_t)$ . The signal noise in prices is  $\varepsilon_p = C_t^{-1} D_t x$ . Since we assume  $x \sim N(0, \sigma_x I)$ , the price noise is distributed  $\varepsilon_p \sim N(0, \Sigma_p)$ , where  $\Sigma_p \equiv \sigma_x C_t^{-1} D_t D_t' C_t^{-1}$ . Substituting in the coefficients  $C_t$  and  $D_t$  shows that the signal precision of prices is  $\Sigma_p^{-1} = \bar{K}_t \bar{K}_t' / (\rho^2 \sigma_x)$  is a diagonal matrix. The  $j$ th diagonal element of  $\bar{K}_t$  is the average capacity allocated to each asset  $j$  at date  $t$ .

*Computing ex-ante expected utility:* Substitute optimal risky asset holdings from equation (11) into the first-period objective function to get:  $U_{1j} = rW_0 + \frac{1}{2} E_1 \left[ (E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r) \hat{\Sigma}_t^{-1} (E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r) \right]$ . Note that the expected excess return  $(E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r)$  depends on signals and prices, both of which are unknown at time 1. Because asset prices are linear functions of normally distributed shocks,  $E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r$ , is normally distributed as well. Thus,  $(E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r) \hat{\Sigma}_t^{-1} (E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r)$  is a non-central  $\chi^2$ -distributed variable. Computing its mean yields:

$$U_{1j} = rW_0 + \frac{1}{2} \text{trace}(\hat{\Sigma}_t^{-1} V_1 [E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r]) + \frac{1}{2} E_1 [E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r]' \hat{\Sigma}_t^{-1} E_1 [E[\tilde{d}_t | \mathcal{I}_{it}] - p_t r]. \quad (21)$$

Note that in expected utility (21), the choice variables  $K_{ijt}$  enter only through the posterior variance  $\hat{\Sigma}_t$  and through  $V_1[E[\tilde{d}_t|\mathcal{I}_{it}] - p_t r] = V_1[\tilde{d}_t - p_t r] - \hat{\Sigma}_t$ . Since there is a continuum of investors, and since  $V_1[\tilde{d}_t - p_t r]$  and  $E_1[E[\tilde{d}_t|\mathcal{I}_{it}] - p_t r]$  depend only on parameters and on aggregate information choices, each investor takes them as given.

Since  $\hat{\Sigma}_t^{-1}$  and  $V_1[E[\tilde{d}_t|\mathcal{I}_{it}] - p_t r]$  are both diagonal matrices and  $E_1[E[\tilde{d}_t|\mathcal{I}_{it}] - p_t r]$  is a vector, the last two terms of (21) are weighted sums of the diagonal elements of  $\hat{\Sigma}_t^{-1}$ . Thus, (21) can be rewritten as  $U_i = rW_0 + \sum_j \Lambda_j \hat{\Sigma}_t^{-1}(j, j) - n/2$ , for positive coefficients  $\Lambda_j$ . Since  $rW_0$  is a constant and  $\hat{\Sigma}_t^{-1}(j, j) = \Sigma^{-1}(j, j) + \Sigma_p^{-1}(j, j) + K_{ij}$ , the information choice problem is (13). From now on, we will use the subindex  $j$  to refer to the  $(j, j)$  element of a matrix, so  $\Sigma^{-1}(j, j) = \Sigma_j^{-1}$ .

## B.1 Proofs

**Proof of Lemma 1.** From step 4 of the model solution, we know that when there is a unique maximum  $\Lambda_{lt}$  the optimal information choice is  $K_{ilt} = K_t = \Sigma^{-1}(\exp \bar{B}_t - 1)$  if  $\Lambda_{lt} = \max_j \Lambda_{jt}$ , and  $K_{ijt} = 0$ , otherwise. If multiple risks achieve the same maximum  $\Lambda_l$  then all attention will be allocated amongst those risks, but each investor would learn about one single risk.

First, we show that value of learning about asset  $j$  falls as the aggregate capacity devoted to studying it increases:  $\partial \Lambda(\bar{K}_{jt}, \bar{x}_j) / \partial \bar{K}_{jt} < 0$ . This is the same strategic substitutability in information as in Grossman and Stiglitz (1980). The solution for  $\Lambda_j$  is given by (14). It is clearly increasing in  $\bar{K}_{jt}$  directly. But there is also an indirect negative effect through  $\hat{\Sigma}_j$ . Recall that by Bayes' Law, the average posterior precision is  $\hat{\Sigma}_j^{-1} = \Sigma_j^{-1} + \Sigma_{pj}^{-1} + \bar{K}_{jt}$  and we know that  $\sigma_{pj}^{-1} = \bar{K}_{jt}^2 / (\rho^2 \sigma_x)$ . Thus,  $\frac{\partial \hat{\Sigma}_j}{\partial \bar{K}_{jt}} < 0$ . To sign the net effect, it is helpful to rewrite  $\Lambda_j$  as

$$\Lambda_j = \hat{\Sigma}_j^2 \left[ \hat{\Sigma}_j^{-1} + \rho^2(1/\tau_x + \bar{x}_j^2) + \bar{K}_{jt} \right]$$

Substituting in  $\hat{\Sigma}_j^{-1} = \Sigma_j^{-1} + \bar{K}_{jt}^2 / (\rho^2 \sigma_x) + \bar{K}_{jt}$ , we get

$$\Lambda_j = \frac{\Sigma_j^{-1} + \bar{K}_{jt}^2 / (\rho^2 \sigma_x) + \rho^2(1/\tau_x + 2\bar{x}_j^2) + 2\bar{K}_{jt}}{(\Sigma_j^{-1} + \Sigma_{pj}^{-1} + \bar{K}_{jt})^2}$$

Finally, the partial derivative with respect to  $\bar{K}_{jt}$  is

$$\begin{aligned} \frac{\partial \Lambda_j}{\partial \bar{K}_{jt}} &= \frac{(2\bar{K}_{jt}^2 / (\rho^2 \sigma_x) + 2)\hat{\Sigma}_j - (\hat{\Sigma}_j + \rho^2(1/\tau_x + 2\bar{x}_j^2) + \bar{K}_{jt})2(2\bar{K}_{jt}^2 / (\rho^2 \sigma_x) + 1)}{\hat{\Sigma}_j^3} \\ &= \frac{-\bar{K}_{jt}^2 / (\rho^2 \sigma_x)\hat{\Sigma}_j - 2(\rho^2(1/\tau_x + 2\bar{x}_j^2) + \bar{K}_{jt})(2\bar{K}_{jt}^2 / (\rho^2 \sigma_x) + 1)}{\hat{\Sigma}_j^3} < 0 \end{aligned}$$

Since the numerator is all terms that can only be negative and the denominator is a sum of precisions, that can only be positive, the sign is negative. This proves that  $\Lambda_j$  is decreasing in  $\bar{K}_{jt}$ .

Now, in addition to  $\frac{\partial \Lambda_j}{\partial \bar{K}_{jt}} < 0$ , we know that all capacity must be used, since we are maximizing a linear objective subject to a concave constraint. Then for some asset attention has to increase, which implies that the new maximum  $\Lambda$  is going to be lower, so by the definition of equilibrium attention on all the assets that are learned about must increase as well. Specifically, lets consider two cases:

*Case 1:*  $\mathcal{M}_t = \mathcal{M}_{t+1}$ : no new assets are added to the set of learned assets  $\mathcal{M}$ . Then allocation of attention in all the assets that are learned about must increase, because if attention to one of those assets decrease or stays the same, his  $\Lambda$  is going to be higher than the  $\Lambda$  of the assets for which attention increased, which would contradicts the definition of equilibrium.

*Case 2:*  $\mathcal{M}_t \subset \mathcal{M}_{t+1}$ : At least one new asset (lets call it  $l$ ) is added to the set of assets that are learned about  $\mathcal{M}$ . Then the new maximum  $\Lambda$  is lower than before, because  $\Lambda(\bar{K}_{l,t+1}, x_l) < \Lambda(0, x_l) < \max \Lambda_t$ . Then, attention in all the assets that were learned about in  $t$  increases, because if not their  $\Lambda$  would be higher than the  $\lambda_l$  of the new asset in  $\mathcal{M}$ , which again contradicts the definition of equilibrium.

■

**Proof of Lemma 2.**

*To show:* If  $\bar{x}_i$  is sufficiently large  $\forall i$ , the set of assets learned about  $\mathcal{M}_t$  does not contain all assets, and  $\bar{B}_{t+1} - \bar{B}_t$  is sufficiently large, then the set of assets  $\mathcal{M}_{t+1}$  learned about in  $t + 1$  is larger than the set  $\mathcal{M}_t$ .

Suppose not. Then, there would be a unique maximum set  $\Lambda_j, \forall j \in \mathcal{M}_t$  that is non-increasing, no matter how large  $\bar{B}_{t+1}$  is. Since there is a unique maximum, the equilibrium solution dictates that all information capacity is used to study this set of risks. Thus the average precision of information,  $\bar{K}_{jt} \equiv \int K_{ijt} di$  becomes arbitrarily large  $\forall j \in \mathcal{M}_t$ .

However, the value of learning about asset  $j$  falls as the aggregate capacity devoted to studying it increases:  $\partial \Lambda_j / \partial \bar{K}_{jt} < 0$ . Furthermore, as the supply of the risk factor  $\bar{x}_j$  becomes large,  $\partial \Lambda_j / \partial \bar{K}_{jt}$  becomes an arbitrarily large negative number. Thus, for a sufficiently large  $\bar{x}_j$ , there exists a  $\underline{K}$  such that if  $\bar{K}_{jt} = \underline{K}$ , then  $\Lambda_j < \Lambda_{j'}$  for some other risk  $j'$ . But then,  $\Lambda_i$  is not a unique maximum in the set of  $\{\Lambda_l\}_{l=1}^N$ , which is a contradiction. Thus the set of assets learned about  $\mathcal{M}_{t+1}$  must grow.

■

**Proof of Lemma 3.** As in the previous lemma, we know that when there is a unique maximum  $\Lambda_{lt}$  the optimal information choice is  $K_{ilt} = K = \Sigma^{-1}(\exp \bar{B}_t - 1)$  if  $\Lambda_{lt} = \max_j \Lambda_{jt}$ , and  $K_{jlt} = 0$ , otherwise. If multiple risks achieve the same maximum  $\Lambda_l$  then all attention will be allocated amongst those risks, but each investor would learn about one single risk. Therefore, there are three cases to consider.

*Case 1:*  $\Lambda_{lt}$  is the unique maximum  $\Lambda_{jt}$ . Holding attention allocations constant, a marginal increase in  $\bar{x}_l$  will cause  $\Lambda_{lt}$  to increase:

$$d\Lambda(\bar{K}_{jt}, \bar{x}_j)/\bar{x}_l|_{\bar{K}_{jt}=\text{constant}} = \rho^2 \bar{\Sigma}_j^2 > 0.$$

The marginal increase in  $\bar{x}_l$  will not affect  $\Lambda_{l't}$  for  $l' \neq l$ . It follows that after the increase in  $\bar{x}_l$ ,  $\Lambda_{lt}$  will still be the unique maximum  $\Lambda_{jt}$ . Therefore, in the new equilibrium, attention allocation is unchanged.

*Case 2:* Prior to the increase in  $\bar{x}_l$ , multiple risks, including risk  $l$ , attain the maximum  $\Lambda_{jt}$ , with  $\mathcal{M}_t$  denoting the set of such risks. If  $\bar{x}_l$  marginally increases and we held attention allocations fixed, then  $\Lambda_{lt}$  would be the unique maximum  $\Lambda_{jt}$ . If  $\Lambda_{lt}$  is the unique maximum, then more investors have to learn about risk  $l$ ,  $\bar{K}_{lt}$  increases, which implies fewer investors learn about any other risk  $l \in \mathcal{M}_t \setminus l$ ,  $\bar{K}_{l't}$  decreases. However, lemma B.1 shows that an increase in  $\bar{K}_{lt}$  would decrease  $\Lambda_{lt}$ . Recall that  $\bar{K}_{lt} = K_{ilt}$  for all investors who learn about asset  $l$ . This effect works to partially offset the initial increase in  $\Lambda_{lt}$  as fewer investors will have an incentive to learn about  $l$ . In the rest of the proof, we construct the new equilibrium attention allocation, following an initial increase  $\Lambda_{lt}$  and show that even though the attention reallocation works to reduce  $\Lambda_{lt}$ , the net effect is a larger  $\bar{K}_{lt}$ .

This solution to this type of convex problem is referred to as a “waterfilling” solution in the information theory literature (See textbook by Cover and Thomas (1991)). To construct a new equilibrium, we reallocate attention from risk  $l' \in \mathcal{M} \setminus l$  to risk  $l$  (increasing the number of investors who learn about  $l$  and as a result  $\bar{K}_{lt}$ , decreasing the number of investors who learn about  $l'$  and as a result  $\bar{K}_{l't}$ ). This decreases  $\Lambda_{lt}$  and increases  $\Lambda_{l't}$ . We continue to reallocate attention from all risks  $l' \in \mathcal{M} \setminus l$  to risk  $l$  in such a way that  $\Lambda_{l't} = \Lambda_{l''t}$  for all  $l', l'' \in \mathcal{M} \setminus l$  is maintained. We do this until either (i) all attention has been allocated to risk  $l$  or (ii)  $\Lambda_{lt} = \Lambda_{l't}$  for all  $l' \in \mathcal{M} \setminus l$ . Note that in the new equilibrium  $\Lambda_{lt}$  will be larger than before and the new equilibrium  $\bar{K}_{lt}$  will be larger than before, while  $\bar{K}_{l't}$ ,  $l' \in \mathcal{M} \setminus l$  will be smaller than before.

*Case 3:* Prior to the increase in  $\bar{x}_l$ ,  $\Lambda_{lt} < \Lambda_{l't}$  for some  $l' \neq l$ . Because  $\Lambda_{lt}$  is a continuous function of  $\bar{x}_l$ , a marginal increase in  $\bar{x}_l$ , will only change  $\Lambda_{lt}$  marginally. Because  $\Lambda_{lt}$  is discretely less than  $\Lambda_{l't}$ , the ranking of the  $\Lambda_{it}$ 's will not change and the new equilibrium will maintain the same attention allocation.

In cases one and three  $\bar{K}_{lt}$  does not change in response to a marginal increase in  $\bar{x}_l$ . In case two  $\bar{K}_{lt}$  is strictly increasing in  $\bar{x}_l$ . Therefore,  $\bar{K}_{lt}$  is weakly increasing in  $\bar{x}_l$ .

■

**Proof of Lemma 4.** Differentiating (19) a second time,

$$\frac{\partial^2 C_{jt}}{\partial K_j^2} = -\frac{2\rho^2 \Sigma_j \left( 3K_j^2 \tau_{xj}^2 + \rho^2 \tau_{xj} (3K_j - \Sigma_j) + \rho^4 \right)}{\left( K_j^2 \tau_{xj} + \rho^2 (K_j + \Sigma_j) \right)^3}$$

So as long as  $K_j \geq \frac{\Sigma_j}{3}$ , the numerator is positive and thus the second derivative is negative, which completes the proof.  $\square$  ■

## C What Data Should Society Be Processing?

To our framework, we add real spillovers that can speak to social efficiency. Our stylized model of the real economy is designed to show one possible reason why financial price informativeness might have economic consequences. In this case, commonly-used compensation contracts that tie wages to firm equity prices (e.g., options packages) also tie price informativeness to optimal effort. Since investors are infinitesimal and take prices as given, they do not internalize the effect of their information and portfolio choices on manager's decision through price informativeness.

**Firm Manager's Problem** At each date  $t$ , each firm manager solves a one-period problem. The key friction is that the manager's effort choice is unobserved by investors. The manager exerts costly effort only because he is compensated with equity, whose value is responsive to his effort. Because asset price informativeness governs the responsiveness of price to effort, it also determines the efficiency of the manager's effort choice.<sup>12</sup>

The profit of each firm  $j$ ,  $\tilde{d}_{jt}$ , depends on the firm manager's effort, which we call labor  $l_{jt}$ . Specifically, the payoff of each share of the firm is  $\tilde{d}_{jt} = g(l_{jt}) + \tilde{y}_{jt}$ , where  $g(l) = l^\phi$ ,  $\phi \leq 1$  is increasing and concave and the noise  $\tilde{y}_{jt} \sim N(0, \tau_0^{-1})$  is i.i.d. and unknown at  $t$ . Because effort is unobserved, the manager's pay  $w_{jt}$  is tied to the equity price  $p_{jt}$  of the firm:  $w_{jt} = \bar{w}_j + p_{jt}$ . However, effort is costly. We normalize the units of effort so that a unit of effort corresponds to a unit of utility cost. Insider trading laws prevent the manager from participating in the equity market. Thus, each period, the manager chooses  $l_{jt}$  to maximize

$$U_m(l_{jt}) = \bar{w}_j + p_{jt} - l_{jt} \tag{22}$$

Each period, the firm  $j$  pays out all its profits  $\tilde{d}_{jt}$  as dividends to its shareholders. We let  $\tilde{d}_t$  denote the vector whose  $j^{\text{th}}$  entry is  $\tilde{d}_{jt}$ .

Both the planner and the investor care more about large firms. The investor values data about large firms more, all else equal, because a large firm, by definition, makes up a larger fraction of

---

<sup>12</sup>Of course, this friction reflects the fact that the wage is not an unconstrained optimal contract. The optimal compensation for the manager is to pay him for effort directly or make him hold all equity in the firm. We do not model the reasons why this contract is not feasible because it would distract from our main point.

the value of an average investor's portfolio. The social planner values data about large firms more because the output of each firm  $E[\tilde{d}_{jt}] - l_{jt}$  is scaled by firm size  $\bar{x}_j$ . In both cases, there are returns to scale in information.

However, investors can rebalance their portfolio to hold more and more of the asset they learn about, whereas the social planner takes the set of firms in the economy as given. This makes returns to scale stronger for investors than for the social planner. Thus, investors prefer to process more data about large firms than what the social planner would prefer. Note that the fact that there are economic externalities is by construction. The result that the social planner favors more data processing about small firms is not.

## C.1 A Planner's Problem with Parallel Investor Processing

The planner maximizes the total output by choosing the allocation of investor information acquisition capacity, taking manager optimal effort decision and investor optimal portfolio decision as given.<sup>13</sup> Formally, the planner chooses aggregate signal precisions  $K_1$  and  $K_2$  to maximize

$$\max_{\{\bar{K}_{jt}\}} \sum_j \bar{x}_j \left( \mathbb{E}[\tilde{d}_{jt}] - l_{jt} \right) \quad (23)$$

$$\text{s.t.} \quad C_{jt} = \Gamma(\bar{K}_{jt}) \quad \forall j \text{ and} \quad (24)$$

$$\sum_j \bar{K}_{jt} = K_t \quad (25)$$

Note that the constraint on processing power for the planner is linear in signal precision. This is different that the constraint facing the individual investor. It represents the idea of parallel computing and a continuum of investors. The computing of different investors is done independently. If each investor can process a total of  $b_I$  bits, which results in a signal of precision  $k_I$ , then producing a signal with double that precision requires two investors, each processing  $b_I$  bits, each producing a conditionally independent signal of precision  $k_I$ . Bayes law tells us that if we combine two conditionally independent, normal signals, each with precision  $k_I$ , the total precision of the optimally combined signals is  $2k_I$ . So, double the precision requires double the resources, implying a linear constraint on signal precision.

The first order condition of this problem with respect to  $\bar{K}_{jt}$  is

$$\bar{x}_j \Gamma^{-2}(\bar{K}_{jt}) \Gamma'(\bar{K}_{jt}) \left( (g')^{-1} \right)'(\bar{K}_{jt}) \left( g'(\bar{K}_{jt}) - 1 \right) = \mu \quad (26)$$

where  $\mu$  is the Lagrange multiplier, or shadow cost, of one additional unit of aggregate signal precision.

---

<sup>13</sup>Assume investors have sufficient wealth that their marginal utility is vanishingly small and drops out of planner objective.

In general, equilibrium outcomes and constrained efficient allocation are different. We can see this from the fact that the solutions to equations 15 and 26 do not coincide. But why are individual and social choices different and what are the economics behind this difference?

If we substitute  $\mathbb{E}[\tilde{d}_{jt}] = g(l_{jt}) = l_{jt}^\phi$  and then use the labor first-order condition ( $p'(l) = 1$ ) to substitute for  $l$ , we get a simplified planner problem:

$$\begin{aligned} \max_{\{\bar{K}_{jt}\}} \quad & \sum_j \bar{x}_j \left( (\phi \Gamma(\bar{K}_{jt}))^{\frac{\phi}{1-\phi}} - (\phi \Gamma(\bar{K}_{jt}))^{\frac{1}{1-\phi}} \right) \\ \text{s.t.} \quad & \sum_j \bar{K}_{jt} = K \end{aligned}$$

Merging the first order condition of the planner with respect to any two assets  $j, j'$  we get

$$\frac{(1 - 1/\Gamma(\bar{K}_{jt})) \Gamma(\bar{K}_{jt})^{\frac{\phi}{1-\phi}} \Gamma'(\bar{K}_{jt})}{(1 - 1/\Gamma(\bar{K}_{j't})) \Gamma(\bar{K}_{j't})^{\frac{\phi}{1-\phi}} \Gamma'(\bar{K}_{j't})} = \frac{\bar{x}_{j'}}{\bar{x}_j} \quad (27)$$

Let  $F$  represent the marginal social value of an additional unit of data precision, per share of the asset,

$$F(\bar{K}_{jt}) = (1 - \Gamma^{-1}(\bar{K}_{jt})) \Gamma(\bar{K}_{jt})^{\frac{\phi}{1-\phi}} \Gamma'(\bar{K}_{jt}).$$

Then with two assets, we can express the social optimum simply as  $F(\bar{K}_{2t})/F(\bar{K}_{1t}) = \bar{x}_1/\bar{x}_2$ .

## C.2 Why the Social Optimum Involves Less Data on Large Firms

For the investor, the potential profits from learning more and more precise information are unbounded. But for a social planner, the gains to information from added efficiency are bounded. From differentiating (19), we learn that  $\frac{\partial C_{jt}}{\partial \bar{K}_{jt}} > 0$  and  $\lim_{\bar{K}_{jt} \rightarrow \infty} C_{jt} = 1$ . Thus, an infinite amount of data can only possibly make price informativeness equal to 1 at most. This offers finite social welfare gains.

**Lemma 4** *The improvement in price informativeness from additional data processing exhibits diminishing returns. If  $K_t$  is sufficiently large, then  $\partial^2 C_{jt} / \partial \bar{K}_{jt}^2 < 0$ .*

To ensure that the second order condition of the planner problem is satisfied, it must be that  $F'(K) < 0$ , which holds when  $K$  is sufficiently large. Inspecting the objective function of the planner, it is easy to verify that the planner allocates more capacity to the larger asset, proportional to its marginal social value, its supply  $\bar{x}_i$ . This observation is also verified in equation 27 using the second order condition. The difference is governed by concavity of the production function.

The fact that the result rests on a sufficiently high level of data processing explains why this phenomenon of informative large firm prices has grown over time. When  $K$  was small, the social

planner valued large firm data more than the investor. As  $K$  grew larger, the stronger increasing returns to data in large firms for investors kicked in, and large firm prices became more informative.

Let  $\{\bar{K}_{jt}^{sp}\}_j$  and  $\{\bar{K}_{jt}^{eq}\}_j$  denote the solution to the constrained planner and equilibrium. With two assets, the following two equations fully characterize the two solutions when both assets are learned about<sup>14</sup>

$$\begin{aligned}\bar{x}_1 F(\bar{K}_{1t}^{sp}) - \bar{x}_2 F(K_t - \bar{K}_{1t}^{sp}) &= 0 \\ \Lambda(\bar{K}_{1t}^{eq}, \bar{x}_1) - \Lambda(K_t - \bar{K}_{1t}^{eq}, \bar{x}_2) &= 0\end{aligned}$$

where  $\Lambda$  is defined in (14).

It is straightforward to verify that  $\forall (\Sigma^{-1}, \tau_x, \frac{\bar{x}_1}{\bar{x}_2}, \phi)$ ;  $1 < \frac{\bar{x}_1}{\bar{x}_2} < x_{\max}$ ,  $0 < \phi < 1$ , if  $\rho > \bar{\rho}$  then  $\bar{K}_{1t}^{eq} > \bar{K}_{1t}^{sp}$  and  $\bar{K}_{2t}^{eq} < \bar{K}_{2t}^{sp}$ . In other words, in equilibrium investors learn too much about the larger firm and the smaller firm remains under unexplored.

Why does equilibrium feature a misallocation of resources away from the smaller risk toward the larger risk? Although it is true that both the constraint social planner and individual investors care about the larger asset more, the investor preferences are more extreme since information has increasing return to scale at the individual level, but only constant return to scale at the aggregate level.

### C.3 What if Investors' Computing Could be Integrated?

The reason that the social planner's problem features a linear constraint is that each investor in the economy produces a conditionally independent signal. They process data independently. When different processors work simultaneously, but independently on a problem, that is called parallel computing. For a given investor, the constraint on computing is not linear because optimal data processing is not parallel. A single processor can accomplish more than two processors, each with half the power, because its processing is integrated. With integrated computing, twice as many bits can transmit more than double the precision of signal.

This raises the question, what if economy-wide computing became integrated? Instead of each investor processing their data in parallel, what if all computing were done on a common processor? This idea of futuristic cloud computing both is a speculation about future technology, but also a way of breaking down the difference between the social planner and decentralized problem into the technological differences between integrated and parallel computing, and the payoff externalities internalized by the planner. This formulation of the problem gives the planner and the individual the same computing constraints and focuses only on the payoff externalities.

---

<sup>14</sup>Note that in both equilibrium and planner problem it might be that only one asset is learned about, when  $\bar{x}_1 \gg \bar{x}_2$ . Consider only the set of parameters that this does not happen.

$$\begin{aligned} \max_{\{\bar{K}_{jt}\}} \quad & \sum_j \bar{x}_j \left( (\phi\Gamma(\bar{K}_{jt}))^{\frac{\phi}{1-\phi}} - (\phi\Gamma(\bar{K}_{jt}))^{\frac{1}{1-\phi}} \right) \\ \text{s.t.} \quad & \sum_j \ln(1 + \Sigma\bar{K}_{jt}) = \bar{B}_t \end{aligned}$$

and, of course  $\bar{K}_{jt} \geq 0$ . The first order condition of the planner is the same as before, except that the Lagrange multiplier is multiplied by the derivative of  $\ln(1 + \Sigma\bar{K}_{jt})$ , which is  $\Sigma/(1 + \Sigma\bar{K}_{jt})$  or alternatively  $(\Sigma + \bar{K}_{jt})^{-1}$ . Working through the same steps as before for the two assets, we can express the social optimum simply as

$$\frac{F(\bar{K}_{2t})(\Sigma^{-1} + \bar{K}_{2t})}{F(\bar{K}_{1t})(\Sigma^{-1} + \bar{K}_{1t})} = \frac{\bar{x}_1}{\bar{x}_2}. \quad (28)$$

Notice that this solution is as if the marginal social value of data is more increasing, or less decreasing than before. In other words, integrated computing created more increasing returns to processing the same type of data at the aggregate level.

Does this mean that the social planner has increasing returns and will only want to process data on large firms? Probably not. It depends on the parameter values. But this does suggest that a future shift to more integrated computing methods would make more concentrated computing more desirable and bring the social optimum and decentralized equilibrium closer.