Big Data in Finance and the Growth of Large Firms

Juliane Begenau
Stanford GSB & NBER

Maryam Farboodi
Princeton

Laura Veldkamp
NYU Stern & NBER

February 5, 2018*

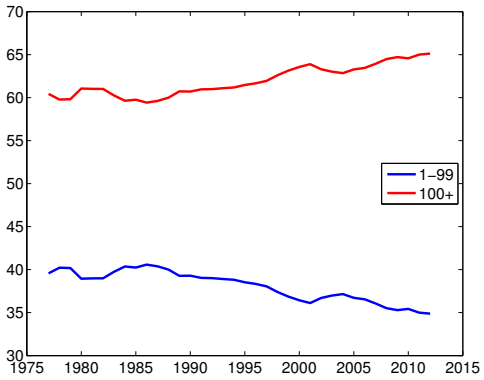
Abstract

One of the most important trends in modern macroeconomics is the shift from small firms to large firms. At the same time, financial markets have been transformed by advances in information technology. The goal of this project is to explore the hypothesis that the use of big data in financial markets has lowered the cost of capital for large firms, relative to small ones, enabling large firms to grow larger. As faster processors crunch ever more data – macro announcements, earnings statements, competitors' performance metrics, export demand, etc. – large firms become more valuable targets for this data analysis. Large firms, with more economic activity and a longer firm history offer more data to process. Once processed, that data can better forecast firm value, reduce the risk of equity investment, and thus reduce the firm's cost of capital. As big data technology improves, large firms attract a more than proportional share of the data processing, enabling large firms to invest cheaply and grow larger.

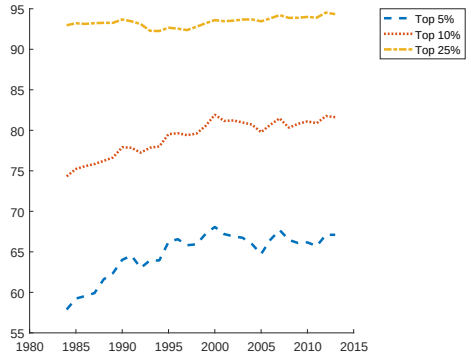
*This paper was prepared for the Carnegie-Rochester-NYU conference. We thank the conference committee for their support of this work. We also thank Nic Kozeniauskas for his valuable assistance with the data and Adam Lee for his outstanding research assistance.

One of the main question in macroeconomics today is why small firms are being replaced with larger ones. Over the last three decades, the percentage of employment at firms with less than 100 employees has fallen from 40% to 35% (Figure 1a); the annual rate of new startups has decreased from 13% to less than 8%, and the share of employment at young firms (less than 5 years) has decreased from 18% to 8% (Davis and Haltiwanger, 2015). While small firms have struggled, large firms (more than 1000 employees) have thrived: The share of the U.S. labor force they employ has risen from one quarter in the 1980s, to about a third today. At the same time, the revenue share of the top 5% of firms increased from 57% to 67% (Figure 1b).

Figure 1: Large Firms Growing Relatively Larger. The left panel uses the Business Dynamics Statistics data published by the Census Bureau (from Kozeniauskas, 2017). It contains all firms with employees in the private non-farm sector in the United States. The right panel uses Compustat/CRSP data. Top x% means the share of all firm revenue earned by the x% highest-revenue firms.



(a) Employment % of large/small firms



(b) Revenue % at firms in top sales percentile

One important difference between large and small firms is their cost of capital (Cooley and Quadrini, 2001). Hennessy and Whited (2007) document that larger firms, with larger revenues, more stable revenue streams, and more collateralizable equipment, are less risky creditors and thus pay lower risk premia. But this explanation for the trend in firm size is challenged by the fact that while small firms are more volatile, the volatility gap between small firms and large firms cash flows has not grown.[1]

If neither volatility nor covariance with market risk has diverged, how could risk premia

[1]Evidence on the volatility gap between large and small firms is in Appendix A.4. Other hypotheses are that the productivity of large firms has increased or that potential entrepreneurs instead work for large firms. This could be because of globalization, or the skill-biased nature of technological change as in Kozeniauskas (2017). These explanations are not exclusive and may each explain some of the change in the distribution of firm size.

and thus the cost of capital diverge? What introduces a wedge between unconditional variance or covariance and risk is information. Even if the payoff variance is constant, better information can make payoffs more predictable and therefore less uncertain. Given this new data, the conditional payoff variance and covariance fall. More predictable payoffs lower risk and lower the cost of capital. The strong link between information and the cost of capital is supported empirically by Fang and Peress (2009), who find that media coverage lowers the expected return on stocks that are more widely covered. This line of reasoning points to an information-related trend in financial markets that has affected the abundance of information about large firms relative to small firms. What is this big trend in financial information? It is the big data revolution.

The goal of this project is to explore the hypothesis that the use of big data in financial markets has lowered the cost of capital for large firms relative to small ones, enabling large firms to grow larger. In modern financial markets, information technology is pervasive and transformative. Faster and faster processors crunch ever more data: macro announcements, earnings statements, competitors' performance metrics, export market demand, anything and everything that might possibly forecast future returns. This data informs the expectations of modern investors and reduces their uncertainty about investment outcomes. More data processing lowers uncertainty, which reduces risk premia and the cost of capital, making investments more attractive.

To explore and quantify these trends in modern computing and finance, we use a noisy rational expectations model where investors choose how to allocate digital bits of information processing power among various firm risks, and then use that processed information to solve a portfolio problem. The key insight of the model is that the investment-stimulating effect of big data is not spread evenly across firms. Large firms are more valuable targets for data analysis because more economic activity and a longer firm history generates more data to process. All the computing power in the world cannot inform an investor about a small firm that has a short history with few disclosures. As big data technology improves, large firms attract a more than proportional share of the data processing. Because data resolves risk, the gap in the risk premia between large and small firms widens. Such an asset pricing pattern enables large firms to invest cheaply and grow larger.

The data side of the model builds on theory designed to explain human information processing (Kacperczyk et al., 2016), and embeds it into a standard model of corporate finance and investment decisions (Gomes, 2001). In this type of model, deviations from Modigliani-Miller imply that the cost of capital matters for firms' investment decisions. In

our model, the only friction affecting the cost of capital works through the information channel. The big data allocation model can be reduced to a sequence of required returns for each firm that depends on the data-processing ability and firm size. These required returns can then be plugged into a standard firm investment model. To keep things as simple as possible, we study the big-data effect on firms' investment decision based on a simulated sample of firms – two, in our case – in the spirit of Hennessy and Whited (2007).

The key link between data and real investment is the price of newly-issued equity. Assets in this economy are priced according to a conditional CAPM, where the conditional variance and covariance are those of a fictitious investor who has the average precision of all investors' information. The more data the average investor processes about an asset's payoff, the lower is the asset's conditional variance and covariance with the market. A researcher who estimated a traditional, unconditional CAPM would attribute these changes to a relative decline in the excess returns (alphas) on small firms. Thus, the widening spread in data analysis implies that the alphas of small firm stocks have fallen relative to larger firms. These asset pricing moments are new testable model predictions that can be used to evaluate and refine big data investment theories.

This model serves both to exposit a new mechanism and as a framework for measurement. Obviously, there are other forces that affect firm size. We do not build in many other contributing factors. Instead, we opt to keep our model stylized, which allows a transparent analysis of the new role that big data plays. Our question is simply how much of the change in the size distribution is this big data mechanism capable of explaining? We use data in combination with the model to understand how changes in the amount of data processed over time affect asset prices of large and small public firms, and how these trends reconcile with the size trends in the full sample of firms. An additional challenge is measuring the amount of data. Using information metrics from computer science, we can map the growth of CPU speeds to signal precisions in our model. By calibrating the model parameters to match the size of risk premia, price informativeness, initial firm size and volatility, we can determine whether the effect of big data on firms' cost of capital is trivial or if it is a potentially substantial contributor to the missing small firm puzzle.

Contribution to the existing literature Our model combines features from a few disparate literatures. The topic of changes in the firm size distribution is a topic taken up in many recent papers, including Davis and Haltiwanger (2015), Kozeniauskas (2017), and Akcigit and Kerr (2017). In addition, a number of papers analyze how size affects the cost

of capital, e.g. Cooley and Quadrini (2001), Hennessy and Whited (2007), and Begenau and Salomao (2017). We explore a very different force that affects firm size and quantify its effect.

Another strand of literature explores long run data or information trends in finance: Asriyan and Vanasco (2014), Biais et al. (2015) and Glode et al. (2012) model growth in fundamental analysis or an increase in its speed. The idea of long-run growth in information processing is supported by the rise in price informativeness documented by Bai et al. (2016).

Over time, it has gotten easier and easier to process large amounts of data. As in Farboodi et al. (2017), this growing amount of data reduces the uncertainty of investing in a given firm. But the new idea that this project adds to the existing work on data and information frictions, is this: Intensive data crunching works well to reduce uncertainty about large firms with long histories and abundant data. For smaller firms, who tend also to be younger firms, data may be scarce. Big data technology only reduces uncertainty if abundant data exists to process. Thus as big data technology has improved, the investment uncertainty gap between large and small firms has widened, their costs of financing have diverged, and big firms have grown ever bigger.

1 Model

We develop a model whose purpose is to understand how the growth in big data technologies in finance affects firm size and gauge the size of that effect. The model builds on the information choice model in Kacperczyk et al. (2016) and Kacperczyk et al. (2015).

1.1 Setup

This is a repeated, static model. Each period has the following sequence of events. First, firms choose entry and firm size. Second, investors choose how to allocate their data processing across different assets. Third, all investors choose their portfolios of risky and riskless assets. At the end of the period, asset payoffs and utility are realized. The next period, new investors arrive and the same sequence repeats. What changes between periods is that firms accumulate capital and the ability to process big data grows over time.

Firm Decisions We assume that firms are equity financed. Each firm i has a profitable 1-period investment opportunity and wants to issue new equity to raise capital for that investment. For every share of capital invested, the firm can produce a stochastic payoff

$f_{i,t}$. Thus total firm output depends on the scale of the investment, which is the number of shares $\bar{x}_{i,t}$, and the output per share $f_{i,t}$:

$$y_{i,t} = \bar{x}_{i,t} f_{i,t}. \tag{1}$$

The owner of the firm chooses how many shares $\bar{x}_{i,t}$ to issue. The owner's objective is to maximize the revenue raised from the sale of the firm, net of the setup or investment cost

$$\tilde{\phi}(\bar{x}_{i,t}, \bar{x}_{i,t-1}) = \phi_0 \mathbf{1}_{(|\Delta \bar{x}_{i,t}|>0)} + \phi_1 |\Delta \bar{x}_{i,t}| + \phi_2 (\Delta \bar{x}_{i,t})^2, \tag{2}$$

where $\Delta \bar{x}_{i,t} = \bar{x}_{i,t} - \bar{x}_{i,t-1}$ and $\mathbf{1}_{|\Delta \bar{x}_{i,t}|>0}$ is an indicator function taking the value of one if $|\Delta \bar{x}_{i,t}|$ is strictly positive. This cost function represents the idea that issuing new equity (or buying equity back) has a fixed cost ϕ_0 and a marginal cost that is increasing in the number of new shares issued. Each share sells at price $p_{i,t}$, which is determined by the investment market equilibrium. The owner's objective is thus

$$Ev_{i,t} = E[\bar{x}_{i,t} p_{i,t} - \tilde{\phi}(\bar{x}_{i,t}, \bar{x}_{i,t-1}) | \mathcal{I}_{t-1}], \tag{3}$$

which is the expected net revenue from the sale of firm i.

The firm makes its choice conditional on the same prior information that all the investors have and understanding the equilibrium behavior of investors in the asset market. But the firm does not condition on $p_{i,t}$. In other words, it does not take prices as given. Rather, the firm chooses $\bar{x}_{i,t}$, taking into account its impact on the equilibrium price.

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_t at the end of period t. Risky assets $i \in \{1, ..., n\}$ have random payoffs $f_{i,t} \sim N(\mu, \Sigma)$, where Σ is a diagonal "prior" variance matrix.[2] We define the $n \times 1$ vector $f_t = [f_{1,t}, f_{2,t}, \ldots, f_{n,t}]'$.

Each asset has a stochastic supply given by $\bar{x}_{i,t} + x_{i,t}$, where noise $x_{i,t}$ is normally distributed, with mean zero, variance σ_x, and no correlation with other noises: $x_t \sim \mathcal{N}(0, \sigma_x I)$. As in any (standard) noisy rational expectations equilibrium model, the supply is random to prevent the price from fully revealing the information of informed investors.

[2]We can allow assets to be correlated. To solve a correlated asset problem simply requires constructing portfolios of assets (risk factors) that are independent from each other, choosing how much to invest and learn about these risk factors, and then projecting the solution back on the original asset space. See Kacperczyk et al. (2016) for such a solution.

Portfolio Choice Problem There is a continuum of atomless investors. Each investor is endowed with begining-of-period wealth, W_t.[3] They have mean-variance preferences over end-of-period wealth, with a risk-aversion coefficient, ρ. Let $\hat{E}_{j,t}$ and $\hat{V}_{j,t}$ denote investor j's period t expectations and variances conditioned on all interim information, which includes prices and signals. Thus, investor j chooses how many shares of each asset to hold, $q_{j,t}$ to maximize period t interim expected utility, $\hat{U}_{j,t}$:

$$\hat{U}_{j,t} = \rho \hat{E}_{j,t}[\hat{W}_{j,t}] - \frac{\rho^2}{2} \hat{V}_{j,t}[\hat{W}_{j,t}] \tag{4}$$

subject to the budget constraint:

$$\hat{W}_{j,t} = r_t W_t + q'_{j,t}(f_t - p_t r_t), \tag{5}$$

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

Prices Equilibrium prices are determined by market clearing:

$$\int q_{j,t} dj = \bar{x}_t + x_t, \tag{6}$$

where the left-hand side of the equation is the vector of aggregate demand and the right-hand side is the vector of aggregate supply of the assets.

Information sets, updating, and data allocation At the start of each period, each investor j chooses the amount of data that she will receive at the interim stage, before she invests. A piece of data is a signal about the risky asset payoff. A time-t signal, indexed by l, about asset i is $\eta_{l,i,t} = f_{i,t} + e_{l,i,t}$, where the data error $e_{l,i,t}$ is independent across pieces of data l, across investors, across assets i and over time. Signal noise is normally distributed and unbiased: $e_{l,i,t} \sim N(0, \sigma_e/\delta)$. By Bayes' law, choosing to observe $\mathcal{M}$ signals, each with signal noise variance σ_e/δ, is equivalent to observing one signal with signal noise variance $\sigma_e/(\mathcal{M}\delta)$, or equivalently, precision $\mathcal{M}\delta/\sigma_e$. The discreteness in signals complicates the analysis, without adding insight. But if we have a constraint that allows an investor to process $\bar{\mathcal{M}}/\delta$ pieces of data, each with precision δ/σ_e, and then we take the limit $\delta \to 0$, we get a quasi-continuous choice problem. The choice of how many pieces of data to process

[3]Since there are no wealth effects in the preferences, the assumption of identical initial wealth is without loss of generality.

about each asset becomes equivalent to choosing $K_{i,j,t}$, the precision of investor j's signal about asset i in period t. Investor j's vector of data-equivalent signals about each asset is $\eta_{j,t} = f_t + \varepsilon_{j,t}$, where the vector of signal noise is distributed as $\varepsilon_{j,t} \sim \mathcal{N}(0, \Sigma_{\eta,j,t})$.The variance matrix $\Sigma_{\eta,j,t}$ is diagonal with the ith diagonal element $K_{i,j,t}^{-1}$. Investors combine signal realizations with priors and information extracted from asset prices to update their beliefs using Bayes' law.

Signal precision choices $\{K_{i,j,t}\}$ maximize start-of-period expected utility, $U_{j,t}$, of the fund's terminal wealth $\hat{W}_{j,t}$. The objective is $-E_{j,t}[\ln \hat{E}_{j,t}[\exp(-\rho \hat{W}_{j,t})]]$, which is equivalent to maximizing

$$U_{j,t} = E_{j,t}[\hat{U}_{j,t}] = E_{j,t}\left[\rho \hat{E}_{j,t}[\hat{W}_{j,t}] - \frac{\rho^2}{2}\hat{V}_{j,t}[\hat{W}_{j,t}]\right], \tag{7}$$

subject to the the budget constraint (5) and three constraints in the information choices.[4]

The first constraint is the *information capacity constraint*. It states that the sum of the signal precisions must not exceed the information capacity:

$$\sum_{i=1}^{n} K_{i,j,t} \leq K_t \quad \text{for each } j, t. \tag{8}$$

In Bayesian updating with normal variables, observing one signal with precision $K_{i,j,t}$ or two signals, each with precision $K_{i,j,t}/2$, is equivalent. Therefore, one interpretation of the capacity constraint is that it allows the manager to observe N signal draws, each with precision $K_{i,j,t}/N$, for large N. The investment manager then chooses how many of those N signals will be about each shock.[5]

The second constraint is the *data availability constraint*. It states that the amount of data processed about the future earnings of firm i cannot exceed the total data generated by the firm. Since data is a by-product of economic activity, data availability depends on the economic activity of the firm in the previous period. In other words, data availability in time t is a function of firm size in $t-1$.

$$K_{i,j,t} \leq \hat{K}(x_{i,t-1}) \quad \text{for all } i, j, t. \tag{9}$$

[4]See Veldkamp (2011) for a discussion of the use of expected mean-variance utility in information choice problems.

[5]The results are not sensitive to the exact nature of the information capacity constraint. We could instead specify a cost function of data processing $c(K_{i,j,t})$. The problem we solve is the dual of this cost function approach. For any cost function, there exists a constraint value K_t such that the cost function problem and the constrained problem yield identical solutions.

This limit on data availability is a new feature of the model. It is also what links firm size/age to the expected cost of capital. We assume that the data availability constraint takes a simple, exponential form: $\hat{K}(x_{i,t-1}) = \alpha \exp(\beta x_{i,t-1})$.

The third constraint is the *no-forgetting constraint*, which ensures that the chosen precisions are non-negative:

$$K_{i,j,t} \geq 0 \quad \text{for all } i, j, t. \tag{10}$$

It prevents the manager from erasing any prior information to make room to gather new information about another asset.

1.2 Equilibrium

To solve the model, we begin by working through the mechanics of Bayesian updating. There are three types of information that are aggregated in posterior beliefs: prior beliefs, price information, and (private) signals. We conjecture and later verify that a transformation of prices p_t generates an unbiased signal about the risky payoffs, $\eta_{p,t} = f_t + \epsilon_{p,t}$, where $\epsilon_{p,t} \sim N(0, \Sigma_{p,t})$, for some diagonal variance matrix $\Sigma_{p,t}$. Then, by Bayes' law, the posterior beliefs about f_t are normally distributed: $f_t \sim N(\hat{E}_{j,t}[f_t], \hat{\Sigma}_{j,t})$, where the posterior mean and precision are given by:

$$\hat{E}_{j,t}[f_t] = \hat{\Sigma}_{j,t}(\Sigma^{-1}\mu + \Sigma_{\eta,j,t}^{-1}\eta_{j,t} + \Sigma_{p,t}^{-1}\eta_{p,t}) \tag{11}$$

$$\hat{\Sigma}_{j,t}^{-1} = \Sigma^{-1} + \Sigma_{p,t}^{-1} + \Sigma_{\eta,j,t}^{-1}. \tag{12}$$

Next, we solve the model in four steps.

Step 1: Solve for the optimal portfolios, given information sets and issuance.

Substituting the budget constraint (5) into the objective function (4) and taking the first-order condition with respect to $q_{j,t}$ reveals that optimal holdings are increasing in the investor's risk tolerance, precision of beliefs, and expected return:

$$q_{j,t}^* = \frac{1}{\rho}\hat{\Sigma}_{j,t}^{-1}(\hat{E}_{j,t}[f_t] - p_t r_t). \tag{13}$$

Step 2: Clear the asset market.

Substitute each agent j's optimal portfolio (13) into the market-clearing condition (6). Collecting terms and simplifying reveals that equilibrium asset prices are linear in payoff risk shocks and in supply shocks:

Lemma 1. $p_t = \frac{1}{r_t}\left(A_t + B_t(f_t - \mu) + C_t x_t\right)$

A detailed derivation of coefficients A_t, B_t, and C_t, expected utility, and the proofs of this lemma and all further propositions are in the Appendix.

In this model, agents learn from prices because prices are informative about the asset payoffs f_t. Next, we deduce what information is implied by Lemma 1. Price information is the signal about f_t contained in prices. The transformation of the price vector p_t that yields an unbiased signal about f_t is $\mu + \eta_{p,t} \equiv B_t^{-1}(p_t r_t - A_t$. Note that applying the formula for $\eta_{p,t}$ to Lemma 1 reveals that $\eta_{p,t} = f_t + \varepsilon_{p,t}$, where the signal noise in prices is $\varepsilon_{p,t} = B_t^{-1} C_t x_t$. Since we assumed that $x_t \sim N(0, \sigma_x I)$, the price noise is distributed $\varepsilon_{p,t} \sim N(0, \Sigma_{p,t})$, where $\Sigma_{p,t} \equiv \sigma_x B_t^{-1} C_t C_t' B_t^{-1'}$. Substituting in the coefficients B_t and C_t from the proof of Lemma 1 shows that public signal precision $\Sigma_{p,t}^{-1}$ is a diagonal matrix with ith diagonal element $\sigma_{p,i,t}^{-1} = \frac{\bar{K}_{i,t}^2}{\rho^2 \sigma_x}$, where $\bar{K}_{i,t} \equiv \int K_{i,j,t} dj$ is the average capacity allocated to asset i.

This market-clearing asset price reveals the firm's cost of capital. We define the cost of capital as follows.

Definition 1. *The cost of capital for firm i is the difference between the (unconditional) expected payout per share the firm will make to investors, minus the (unconditional) expected price per share that the investor will pay to the firm:* $E_t[f_{i,t}] - E_t[p_{i,t}]$.

Because x_t is a mean-zero random variable and the payoff f_t has mean μ, the unconditional expected price is $E_t[p_{i,t}] = A_{i,t}/r$. Therefore, the expected cost of capital for firm i is $\mu - A_{i,t}/r$.

Step 3: Compute ex-ante expected utility.

Substitute optimal risky asset holdings from equation (13) into the first-period objective function (7) to get: $U_{j,t} = \rho r_t W_t + \frac{1}{2} E_t[(\hat{E}_{j,t}[f_t] - p_t r_t)' \hat{\Sigma}_{j,t}^{-1} (\hat{E}_{j,t}[f_t] - p_t r_t)]$. Note that the expected excess return $(\hat{E}_{j,t}[f_t] - p_t r_t)$ depends on signals and prices, both of which are unknown at the start of the period. Because asset prices are linear functions of normally distributed shocks, $\hat{E}_{j,t}[f_t] - p_t r_t$, is normally distributed as well. Thus, $(\hat{E}_{j,t}[f_t] - p_t r_t)\hat{\Sigma}_{j,t}^{-1}(\hat{E}_{j,t}[f_t] - p_t r_t)$ is a non-central χ^2-distributed variable. Computing its mean yields:

$$U_{j,t} = \rho r_t W_t + \frac{1}{2}\operatorname{tr}(\hat{\Sigma}_{j,t}^{-1} V_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]) + \frac{1}{2} E_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]' \hat{\Sigma}_{j,t}^{-1} E_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]. \tag{14}$$

Step 4: Solve for information choices.

Note that in expected utility (14), the choice variables $K_{i,j,t}$ enter only through the posterior variance $\hat{\Sigma}_{j,t}$ and through $V_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t] = V_{j,t}[f - p_t r_t] - \hat{\Sigma}_{j,t}$. Since there is a continuum of investors, and since $V_{j,t}[f - p_t r_t]$ and $E_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]$ depend only on parameters and on aggregate information choices, each investor takes them as given.

Since $\hat{\Sigma}_{j,t}^{-1}$ and $V_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]$ are both diagonal matrices and $E_{j,t}[\hat{E}_{j,t}[f_t] - p_t r_t]$ is a vector, the last two terms of (14) are weighted sums of the diagonal elements of $\hat{\Sigma}_{j,t}^{-1}$. Thus, (14) can be rewritten as $U_{j,t} = r_t W_t + \sum_i \lambda_{i,t} \hat{\Sigma}_{j,t}^{-1}(i,i) - n/2$, for positive coefficients $\lambda_{i,t}$. Since $r_t W_t$ is a constant (in each period t) and $\hat{\Sigma}_{j,t}^{-1}(i,i) = \Sigma^{-1}(i,i) + \Sigma_{p,t}^{-1}(i,i) + K_{i,j,t}$, the information choice problem is:

$$\max_{K_{1,j,t},\ldots,K_{n,j,t} \geq 0} \sum_{i=1}^{n} \lambda_{i,t} K_{i,j,t} + \text{constant} \tag{15}$$

$$s.t. \quad \sum_{i=1}^{n} K_{i,j,t} \leq K_t \tag{16}$$

$$K_{i,j,t} \leq \alpha \exp(\beta x_{i,t-1}) \quad \forall i\,,\ \forall j \tag{17}$$

$$\text{where} \quad \lambda_{i,t} = \bar{\sigma}_{i,t}[1 + (\rho^2 \sigma_x + \bar{K}_{i,t})\bar{\sigma}_{i,t}] + \rho^2 \bar{x}_{i,t}^2 \bar{\sigma}_{i,t}^2, \tag{18}$$

where $\bar{\sigma}_{i,t}^{-1} = \int \hat{\Sigma}_{j,t}^{-1}(i,i) dj$ is the average precision of posterior beliefs about firm i. Its inverse, average variance $\bar{\sigma}_{i,t}$ is decreasing in $\bar{K}_{i,t}$. Equation (18) is derived in the Appendix.

This is not a concave objective, so a first-order approach will not find an optimal data choice. To maximize a weighted sum (15) subject to an unweighted sum (16), the investor optimally assigns all available data, as per (17), to the asset(s) with the highest weight. If there is a unique $i_t^* = \text{argmax}_i\, \lambda_{i,t}$, then the solution is to set $K_{i_t^*,j,t} = \min(K_t, \alpha \exp(\beta x_{i,t-1}))$.

In many cases, after all data processing capacity is allocated, there will be multiple assets with identical $\lambda_{i,t}$ weights. That is because $\lambda_{i,t}$ is decreasing in the average investor's signal precision. When there exist asset factor risks i, l s.t. $\lambda_{i,t} = \lambda_{l,t}$, then investors are indifferent about which assets' data to process. The next result shows that this indifference is not a knife-edge case. It arises whenever the aggregate amount of data processing capacity is sufficiently high.

Lemma 2. *If $\bar{x}_{i,t}$ is sufficiently large $\forall i$ and $\sum_i \sum_j K_{i,j,t} \geq \underline{K}$, then there exist risks l and l' such that $\lambda_{l,t} = \lambda_{l',t}$.*

This is the big data analog to Grossman and Stiglitz (1980)'s strategic substitutability in information acquisition. The more other investors know about an asset, the more informative

prices are and the less valuable it is for other investors to process data about the same asset. If one asset has the highest marginal utility for signal precision, but capacity is high, then many investors will learn about that asset, causing its marginal utility to fall and equalize with the next most valuable asset data. With more capacity, the highest two $\lambda_{i,t}$'s will be driven down until they equate with the next $\lambda_{i,t}$, and so forth. This type of equilibrium is called a "waterfilling" solution (see, Cover and Thomas (1991)). The equilibrium uniquely pins down which assets are being learned about in equilibrium, and how much is learned about them, but not which investor learns about which asset.

Step 5: Solve for firm equity issuance. How a firm chooses $\bar{x}_t$ depends on how issuance affects the asset price. Supply $\bar{x}_t$ enters the asset price in only one place in the equilibrium pricing formula, through A_t. From Appendix equation (33), we see that

$$A_t = \mu - \rho\bar{\Sigma}_t\bar{x}_t. \tag{19}$$

$\bar{x}_t$ has a direct effect on the second term. But also an indirect effect through information choices that show up in $\bar{\Sigma}_t$.

The firm's choice of $\bar{x}_t$ satisfies its first order condition:

$$E[p_t|\mathcal{I}_{t-1}] - \bar{x}_t\left(\rho\bar{\Sigma}_t - \rho\bar{x}_t\frac{\partial\bar{\Sigma}_t}{\partial\bar{x}_t}\right) - \tilde{\phi}_1'(\bar{x}_t, \bar{x}_{t-1}) = 0 \tag{20}$$

The first term is the benefit of more issuance. When a firm issues an additional share, it gets expected revenue $E[p_t|\mathcal{I}_{t-1}]$ for that share. The second term tells us that issuance has a positive and negative effect on the share price. The negative effect on the price is that more issuance raises the equity premium ($\rho\bar{\Sigma}_t$ term). The positive price effect is that more issuance makes data on the firm more valuable to investors. When investors process more data on the firm, it lowers their investment risk, and on average, raises the price they are willing to pay ($\partial\bar{\Sigma}_t/\partial\bar{x}_t$ term). This is the part of the firm investment decision that the rise of big data will affect.

The third term, the capital adjustment cost ($\tilde{\phi}_1'(\bar{x}_t, \bar{x}_{t-1})$), reveals why firms grow in size over time. Firms have to pay to adjust relative to their previous size. Since firms' starting size is small they want to grow, but rapid growth is costly. So, they grow gradually. Each time a firm starts larger, choosing a higher $\bar{x}_t$ becomes less costly because the size of the change, and thus the adjustment cost is smaller.

2 Parameter Choice

In order to quantify the potential effect of big data on firm size, we need a calibrated model. Our calibration strategy is to estimate our equilibrium price equation on recent asset price and dividend data. By choosing model parameters that match the pricing coefficients, we ensure that we have the right average price, average dividend, volatility and dividend-price covariance at the simulation end point. What we do not calibrate to is the evolution of these moments over time. The time path of price and price coefficients are over-identifying moments that we can use to evaluate model performance.

First, we describe the data used for model calibration. Next, we describe moments of the data and the model that we match to identify model parameters. Most of these moments come from estimating a version of our price equation (lemma 1).

Data We use two datasets that both come from CRSP. The first is the standard S&P 500 market capitalization index based on the US stock market's 500 largest companies.[6] The dataset consists of the value-weighted price level of the index p_t, and the value-weighted return $(p_t + d_t)/p_{t-1}$, where d_t is dividends. Both are reported at a monthly frequency for the period 1999.12-2015.12.

Given returns and prices, we impute dividends per share as

$$d_t = \left(\frac{p_t + d_t}{p_{t-1}} - \frac{p_t}{p_{t-1}} \right) p_{t-1}.$$

Both the price series and the dividend series are seasonally adjusted and exponentially detrended. As prices are given in index form, they must be scaled to dividends in a meaningful way. The annualized dividend per share is computed for each series by summing dividends in 12 month windows. Then, in the same 12-month window, prices are adjusted to match this yearly dividend-price ratio.

Finally, because the price variable described above is really an index, and this index is an average of prices, the volatility of the average will likely underestimate the true volatility of representative stock prices. In order to find an estimate for price volatility at the asset level, we construct a quarterly time series of the average S&P constituent stock price for the

[6] As a robustness check, we redo the calibration using a broader index: a composite of the NYSE, AMEX and Nasdaq. This is a market capitalization index based on a larger cross-section of the market - consisting of over 8000 companies (as of 2015). The results are similar. Moment estimates are within about 20% of each other. This is close enough for the simulations to differ imperceptibly. Results are available upon request.

period 2000-2015. Compustat gives us the S&P constituent tickets for each quarter. From CRSP, we extract each company's stock price for that quarter.

Moments Using the price data and implied dividend series, we estimate the linear price equation in lemma 1. We let $C_t x_t$ be the regression residuals. We can then map these A_t, B_t and C_t estimates into the underlying model parameters σ_x, σ (where we assume that $\Sigma_{ii} = \sigma_i = \sigma$ for each i), and μ , using the model solutions (33), (34), and (35), as well as

$$Var[p_t] = B_t^2 \Sigma + C_t^2 \sigma_x. \tag{21}$$

Of course, in the model, A_t, B_t and C_t take on different values, depending on how much data is being processed at date t. So we need to choose a theoretical date t at which to calibrate, which amounts to choosing a data processing level K_t. Therefore, we use the empirical price coefficient estimates, along with the 2015 CPU speed estimate, to tell us the model coefficients in 2015. We use solutions of the model to map these estimated coefficients back into model parameters.[7]

Table 1: Parameters

μ	σ	σ_x^{-1}	r_t	ϕ_0	ϕ_1	ϕ_2	ρ	α	β
15	0.55	0.5	0.10	0.598	0.091	0.0004	1.01	0.249	0.0002

Each period in the model is 1 year. The first three parameters in Table 1 are calibrated to match the model and data values of the three equations above. This is an exactly identified system. The riskless rate is set to match a 1% annual net return.

To calibrate firm investment costs, we use parameter estimates from Hennessy and Whited (2007). Using annual data from 1988-2001, they estimate the cost of external investment funding as $\Lambda(x) = \phi_0 + \phi_1 * \tilde{x} + \phi_2 \tilde{x}^2$, where $\tilde{x}$ is the amount raised. This amount raised corresponds to the change in issuance $\Delta \bar{x}_t$ in our model. Their parameter estimates for ϕ_0, ϕ_1, and ϕ_2 are reported in Table 1.

The next parameter is risk aversion ρ. Risk aversion clearly matters for the level of the risky asset price. But it is tough to identify. The reason for the difficulty is that if we change risk aversion, and then re-calibrate the mean, persistence and variance parameters to match price coefficients and variance at the new risk aversion level, the predictions of the

[7]The current calibration does not hit all these targets accurately. We are continuing to work on improving the fit. A future version will report metrics of model fit with the data targets.

model are remarkably stable. Roughly, doubling variance and halving risk aversion mostly just redefines units of risk. Therefore, we use the risk aversion $\rho = 1.01$ in what follows and explore other values to show that the results do not depend on this choice. This ρ implies a relative risk aversion that is 0.65, not particularly high. In the appendix, we show an example of an alternative parameterization with even lower risk aversion, show how the other parameters change, and show that it yields similar results. We explore variations in other parameters as well.

The data availability parameters α and β are chosen to give our mechanism a shot at meaningful results. We choose parameters so that the constraint binds for small firms only, for about the first decade. This pins both parameters to a narrow range. If this constraint did not bind, there would be little difference between small and large firm outcomes. If the availability constraint was binding for all firms, then there would be no effect of big data growth because there would be insufficient data to process with the growing processing power.

Figure 2: The evolution of processing performance over the period 1978–2007
Hennessy and Patterson (2011)

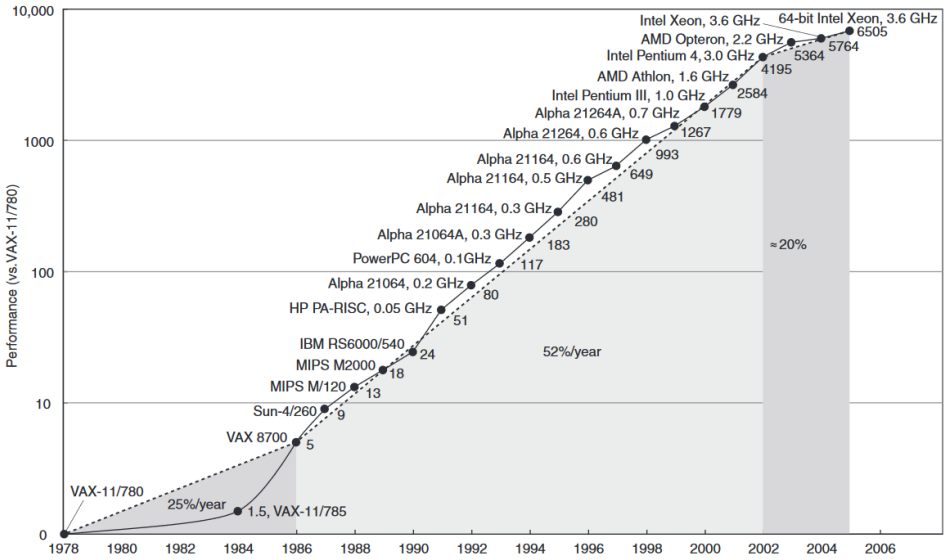


FIGURE 1.16 Growth in processor performance since the mid-1980s. This chart plots performance relative to the VAX 11/780 as measured by the SPECint benchmarks (see Section 1.8). Prior to the mid-1980s, processor performance growth was largely technology-driven and averaged about 25% per year. The increase in growth to about 52% since then is attributable to more advanced architectural and organizational ideas. By 2002, this growth led to a difference in performance of about a factor of seven. Performance for floating-point-oriented calculations has increased even faster. Since 2002, the limits of power, available instruction-level parallelism, and long memory latency have slowed uniprocessor performance recently, to about 20% per year. Copyright © 2009 Elsevier, Inc. All rights reserved.

What changes exogenously at each date is the total information capacity K_t. We normalize $K_t = 1$ in 1980 and then grow K_t at the average rate of growth of CPU speed, as illustrated in Figure 2. We simulate the model in this fashion from 1980-2030.

3 Quantitative Results

Our main results use the calibrated model to understand how the growth of big data affects the evolution of large and small firms and how large that effect might be. We start by exploring how the rise in big data availability changes how data is allocated. Then, we explore how changes in data the investors observe affect the firm's cost of capital. Finally, we turn to the question of how much the change in data and the cost of capital affect the evolution of firms that start out small and firms that start out large.

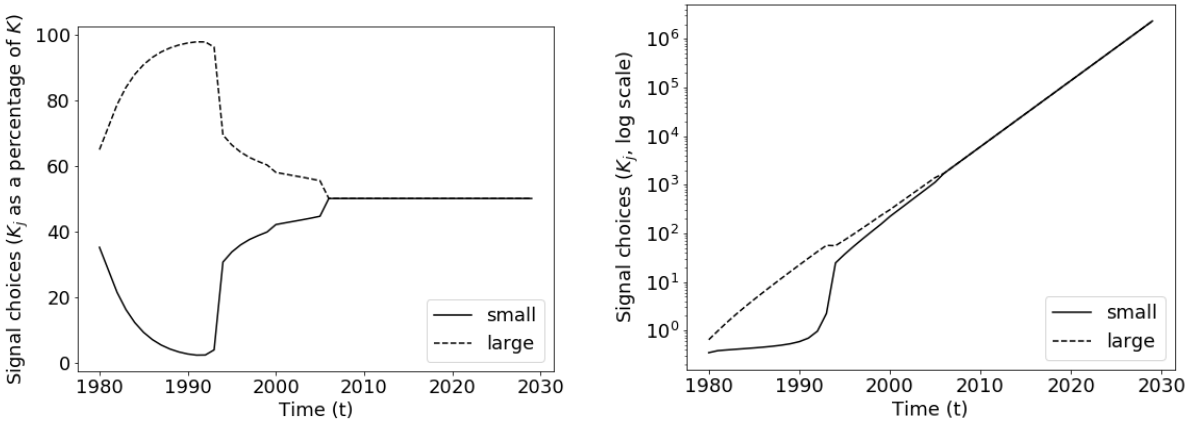
In presenting our results, we try to balance realism with simplicity, which illuminates the mechanism. If we put in a large number of firms, it is, of course, more realistic. But this would also make it harder to see what the trade-offs are. Instead, we characterize the firm distribution with one representative large firm and one representative small firm. The two firms are identical, except that the large firm starts off with a larger size $\bar{x}_0 = 10,000$. The small firm starts off with $\bar{x}_0 = 2000$. Starting in 1980, we simulate our economy with the parameters listed in Table 1 with one period per year until 2030.

3.1 Data Allocation Choices

The reason that data choice is related to firm size is that small firms are young firms and young firms do not have a long historical data series to process. Data comes from having an observable body of economic transactions. A long history with a large amount of economic activity generates this data. A small firm, which is one that has only recently entered, cannot offer investors the data they need to accurately assess risk and return.

But the question is, how does the rise of investors' ability to process big data interact with this size effect? Since investors are constrained in how much data they can process about young, small firms, the increase in data processing ability results in more data being processed about the large firm. We can see this in Figure 3 where the share of data processed on the large firm rises and the share devoted to the small firm falls (left panel). In the right panel, we see that investors are not processing fewer bits of data about the small firm. In fact, as the firm grows, little by little, more data is available. As more data is available, more small firm data is processed and data precision rises.

Figure 3: Investors' Data Choices The left panel shows the share of the total data processed for each firm. The right panel shows the number of bits processed about each firm.



Eventually, the small firm gets large enough and produces a long enough data history that it outgrows its data availability constraint. The availability constraint was pushing data choices for the two firms apart, creating the visual bump in Figure 3. As the constraint relaxes, the bump gives way to a slow, steady convergence. But, even once the data availability constraint stops binding, investors still process more data on the larger firm. A secondary effect of firm size is that data has more value when it is applied to a larger fraction of an investor's portfolio. An investor can use a data set to guide his investment of one percent of the value of his portfolio. But he gains a lot more when he uses that data to guide investment of fifty percent of his portfolio. Big assets constitute more of the value share of the average investor's portfolio. Therefore, information about big assets is more valuable.

Mathematically, we can see firm size $\bar{x}_{i,t}$ enter in the marginal value of information $\lambda_{i,t}$ in (18). Of course, this firm size is the firm's final size that period. But the final size is linked to the firm's initial size through the adjustment cost (2). Firms that are initially larger will have a larger final size because size adjustment is costly. This larger final size is what makes $\lambda_{i,t}$, the marginal value of data, higher.

In the limit, the small firm keeps growing faster than the large firm and eventually catches up. When the two firms approach the same size, the data processing on both converges to an equal, but growing amount of data processing.

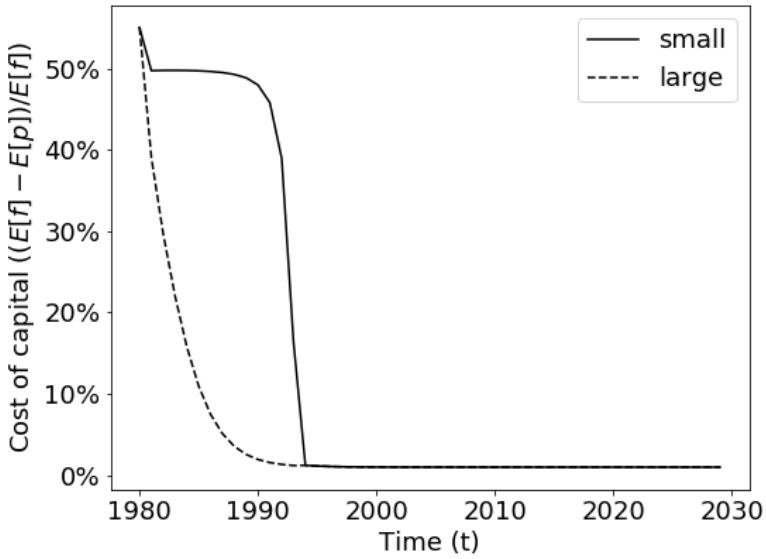
3.2 Capital costs

The main effect of data is to systematically reduce a firm's average cost of capital. Recall that the capital cost is the expected payoff minus the expected price of the asset (Definition 1). Data does not change the firm's payoff, but it does change how a share of the firm is priced. The systematic difference between expected price and payoff is the investor's compensation for risk. Investors are compensated for the fact that firm payoffs are unknown, and therefore buying a share requires bearing risk. The role of data is to help the investor predict that firm payoff. In doing so, data reduces the compensation for risk. Just like a larger data set lowers the variance of an econometric estimate, more data in the model reduces the conditional variance of estimated firm payoffs. An investor who has a more accurate estimate is less uncertain and bears less risk from holding the asset. The representative investor is willing to pay more, on average, for a firm that they have good data on. Of course, the data might reveal problems at the firm that lower the investor's valuation of it. But on average, more data is neither to reveal positive nor negative news. What data does on average improve is the precision and resolution of risk. Resolving the investors' risk reduces the compensation the firm needs to pay the investor for bearing that risk, which reduces the firm's cost of capital.

Figure 4 shows how the large firm, with its more abundant data, has a higher price and lower cost of capital. But this effect is not always smooth. It is not true that more data reduces the cost of capital evenly and proportionately. There is a second force at play here. The second force is that firm size matters. Large firms require investors to hold more of their risk. In CAPM-speak, they have a higher beta. Large firm returns are more correlated with the market return simply because they are more of the market. To induce investors to hold a small amount of a risk is cheap. They can easily diversify it. But to induce investors to hold lots of a risk, the compensation per unit of risk must rise. Thus, because they have more equity outstanding and are more highly correlated with market risk, a large firm with the same volatility and conditional variance as a small firm, would face a higher cost of capital.

As firm size and data evolve together, initially, data dominates. The cost of capital for the large firm falls, from around 50% of earnings per share to close to 1%, because more processing power is reducing the risk of investing in that firm. The small firm cannot initially benefit much from higher processing power because it is a young firm and has little data available to process. As the small firm grows older, the data availability constraint loosens, investors can learn from the firm's track record, risk falls and the cost of capital comes back down. Where the two lines merge is where the small firm finally out-grows

Figure 4: Cost of Capital for a New Firm The sold line represents the cost of capital per share, $E_t[f_{i,t}] - E_t[p_{i,t}]$, normalized by average earnings per share, $E_t[f_{i,t}]$, of the small firm ($\bar{x}_0 = 2000$). The dashed line is the (normalized) cost of capital of the large firm ($\bar{x}_0 = 10000$). Simulations use parameters listed in Table 1.



its data availability constraint. From this point on, the only constraint on processing data on either firm is the total data processing power K. The large and small firms evolve similarly. The only difference between the two firms, after the inflection point where the data availability constraint ceases to bind, is that the small firm continues to have a slightly smaller accumulated stock of capital. Because the small firm continues to be slightly smaller, it has slightly less equity outstanding, and a slightly lower cost of capital. But, when data is abundant, small and large firms converge gradually over time.

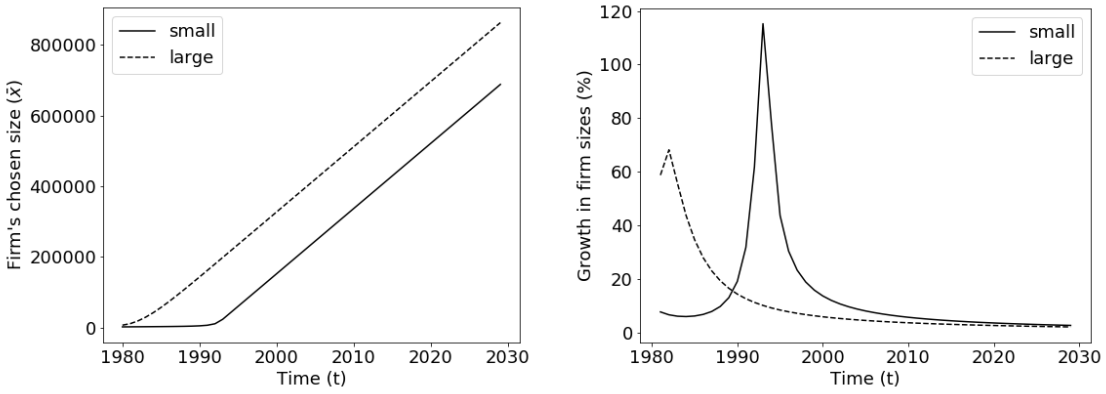
3.3 The Evolution of Firms' Size

In order to understand how big data has changed the size of firms, it is useful to look at how a large firm and a small firm evolve in this economy. Then, we turn off various mechanisms in the model to understand what role is played by each of our key assumptions. Once the various mechanisms are clear, we contrast firm evolution in the 1980's to the evolution of firms in the post-2000 period.

Recall that firms have to pay to adjust, relative to their previous size. Since firms' starting size is small, but rapid growth is costly, firms grow gradually. Figure 5 shows that

both the large and small firms grow. However, the rates at which they grow differ. One reason growth rates differ is that small firms are further from their optimal size. If this were the only force at work, small firms would grow by more each period and that growth rate would gradually decline for both firms, as they approach their optimal size.

Figure 5: The Evolution of Small and Large Firms (level and growth rate) These figures plot firm size $\bar{x}_t$ (left) and growth in firm size, $(\bar{x}_t/\bar{x}_{t-1} - 1) \times 100$ (right), for a small firm, with starting size 2000 and a large firm with starting size 10000. Simulation parameters are those listed in Table 1.



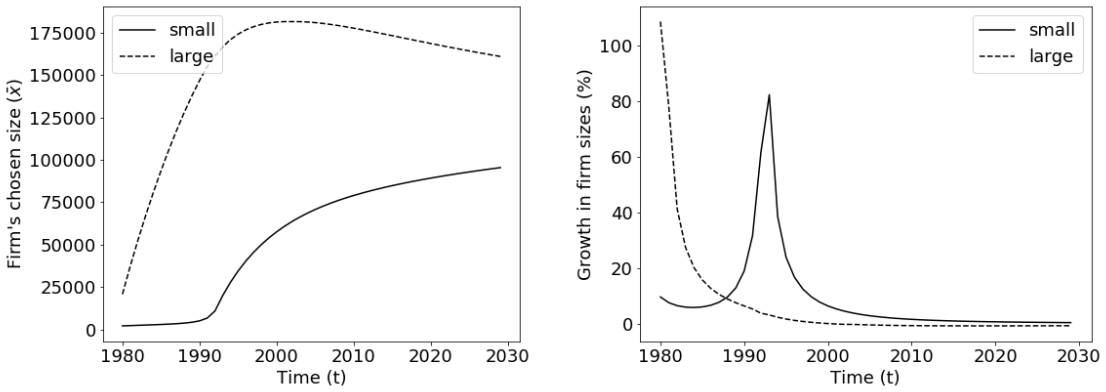
Instead, Figure 5 reveals that small firms sometimes grow faster and sometimes slower than their large firm counterparts. For much of the start of their life, the small firms grow more slowly than the large firms do. These variations in growth rates are due to investors' data processing decisions. This is the force that can contribute to the change in the size of firms.

The level of the size can be interpreted as market capitalization, divided by the expected price. Since the average price ranges from 7 to 15 in this model, these are firms with zero to 12 million dollars of market value outstanding. In other words, these are not very large firms.

The Role of Growing Big Data Plotting firm outcomes over time as in Figure 5 conflates three forces, all changing over time. The first thing changing over time is that firms are accumulating capital and growing bigger. The second change is that firms are accumulating longer data histories, which makes more data for processing available. The third change is that technology enables investors to process more and more of that data over time. We want to understand how each of these contributes to our main results. Therefore, we turn off

features of the model one-by-one, and compare the new results to the main results, in order to understand what role each of these ingredients plays.

Figure 6: Without Improvements in Data Processing, Firm Size Converges These results use the same simulation routine and parameters to plot the same quantities as in Figure 5. The only difference is that these results hold data processing capacity fixed at $K_t = 5\ \forall t$.

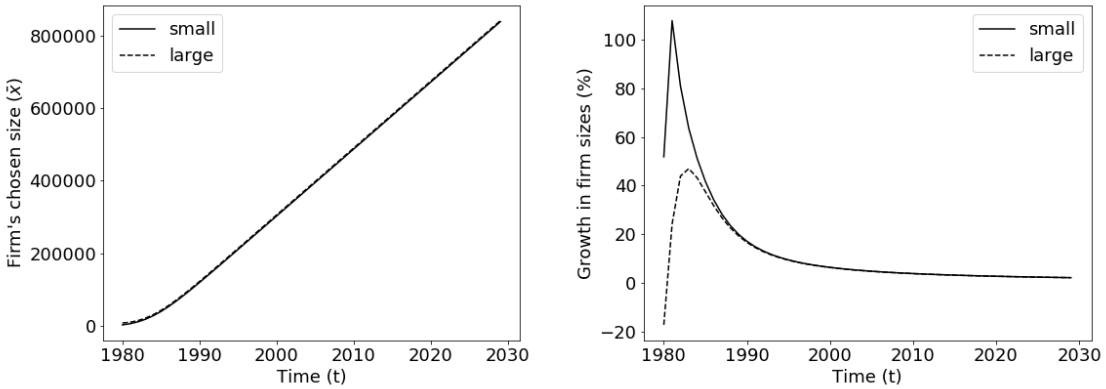


To understand the role that improvements in data processing play, we turn off the growth of big data and compare results. We fix $K_t = 5\ \forall t$. Data processing capacity is frozen at its 1985 level. Firms still have limited data histories and still accumulate capital. Figure 6 shows that this small change in the model has substantial consequences for firm dynamics. Comparing Figures 5 and 6, we can see the role big data plays. In the world with fixed data processing, instead of starting with rapid growth and growing faster as data processing improves, the large firm growth rate starts at the same level as before, but then steadily declines as the firm approaches is stationary optimal size. We learn that improvements in data processing are the sources of firm growth in the model and are central to the continued rapid growth of large firms.

The Role of Limited Data History One might wonder, if large firms attract more data processing, is that alone producing larger big firms? Is the assumption that small firms have a limited data history really important for the results? To answer this question, we now turn off the assumption of limited data history. We maintain the growing data capacity and firm capital accumulation from the original model.

Figure 7 reports results for the model with unlimited firm data histories, but limited processing power, to the full model. Comparing Figures 5 and 7, we can see the difference that data availability makes. In the world where firms have unlimited data histories, small

Figure 7: With Unlimited Data Histories, Small and Large Firms Converge Quickly. These results use the same simulation routine and parameters to plot the same quantities as in Figure 5. The only difference is that these results set the data availability parameters (α, β) to be large enough that the data availability constraint never binds.



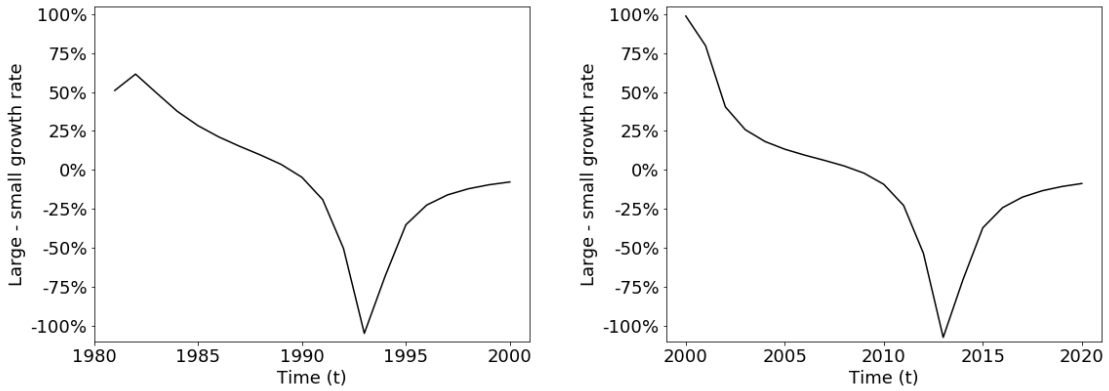
firms quickly catch up to large firms. There is no persistent difference in size. Small firms are far below their optimal size. So they invest rapidly. Investment makes them larger, which increases data processing immediately. Quickly, the initially small and large firms become indistinguishable. Adjustment costs are a friction preventing immediate convergence. But it is really the presence of the data availability constraint that creates the persistant difference between firms with different initial size.

Small and Large Firms in the New Millennium So far, the experiment has been to drop a small firm and a large firm in the economy in 1980 and watch how they evolve. While this is useful to explaining the model's main mechanism, it does not really answer the question of why small firms today struggle more than in the past and why large firms today are larger than the large firms of the past. To answer these questions, we really want to compare small and large firms that enter the economy today to small and large firms that entered in 1980.

To do this small vs. large, today vs. 1980 experiment, we use the same parameters as in Table 1 and use the same starting size for firms. The only difference is that we start with more available processing power. Instead of starting K_t at 1, we start it at the 2000 value, which is about 500.

Each panel of figure 8 shows the growth rate of a large firm, minus the growth rate of a small firm. In the left panel, both firms start in 1980, when data processing capacity was

Figure 8: Large Firms Grow Faster in 2000 than in 1980. Both panels plot a difference in the growth rate of size ($(\bar{x}_t/\bar{x}_{t-1} - 1) \times 100$). The difference is the growth rate of a large firm ($\bar{x}_t$ starting at 10,000) minus the growth rate of a small firm ($\bar{x}_t$ starting at 2000). Both are the result of simulations using parameters in Table 1. The left panel shows the difference in firm growth for firms that start in 1980, with $K_{1980} = 1$. The right panel shows the difference in firm growth for firms that start in 2000, with $K_{2000} \approx 500$.



quite limited. In the right panel, both firms enter in the year 2000, when data is abundant. In both cases, the difference is positive for most of the first decade, meaning that large firms grow faster than small ones. But in 2000, the difference is much more positive. Relative to small firms, large firms grow much more quickly. The difference in 2000 growth rates is nearly twice as large. In both cases, a surviving small firm eventually outgrows its data availability problem, grows quickly, and then converges to the growth rate of the large firm (differences converge to 0).

Of course, in a model with random shocks and exit, many of these small firms would not survive. In a world where large firms gain market share much more rapidly, survival would be all the more challenging. This illustrates how the fact that data processing is more advanced now than in the past may contribute in a substantial way to the puzzle of missing small firms.

For comparison, we examine the growth rates of large and small firms in the U.S. pre-1980 and in the period 1980-2007. We end in 2007 so as to avoid measuring real effects of the financial crisis. For each industry sector and year, we select the top 25% largest firms in Compustat and call those large firms and select the bottom half of the firm size distribution to be our small firms. Within these two sets of firms, we compute the growth rates of various measures of firm size and average them, with an equal weight given to each firm. Then, just as in the model, we subtract the growth rate of large firms from that of small firms. For

most measures, small firms grow more slowly, and that difference grows later in the sample.

Table 2: Large Firm Growth Minus Small Firm Growth from Compustat
For each industry sector and year, large firms are the top 25% largest firms in Compustat; small firms are the bottom half of the firm size distribution. Growth rate is the annual log-difference. Reported figures are equal-weighted averages of growth rates over firms and years.

	prior to 1980	1980 - 2007
Assets	2.1%	8.2%
Investment	14.2%	16.0%
Assets with Intangibles	0.3%	1.1%
Capital Stock	-0.9%	3.7%
Sales	1.4%	2.4%
Market Capitalization	1.1%	8.9%

At times, the magnitudes of the model's growth rates are quite large, compared to the data. Of course, the data is averaged over many years and many firms at different points in their life cycle. This smooths out some of the extremes in the data. If we average the firm growth in our model from 1980-1985, for firms that enter in 1980, we get 39.5% for large firms and 7.3% for small firms, a difference of 32.2%. If we average the firm growth in our model from 2000-2005, for firms that enter in 2000, we get 59.9% for large firms and 7.3% for small firms, a difference of 52.6%.

While it is not unheard of for a small firm to double in size, some of this magnitude undoubtedly reflects some imprecision of our current calibration. A larger adjustment cost, or some labor hiring delay, would help to moderate the extremes of firm size growth. The results also miss many aspects of the firm environment that have changed in the last four decades. The type of firms entering in the last few years are quite different than firms of past years. They have different sources of revenue and assets that might be harder to value. Firm financing has changed, with a shift toward internal financing. Venture capital funding has become more prevalent and displaced equity funding for many firms, early in their life cycle. All of these forces would moderate the large effect we document here.

Our results only show that big data is a force with some potential. There is a logical way in which the growth of big data and the growth of large firms is connected. This channel has the potential to be quantitatively powerful. The role of big data in firms is thus a topic ripe for further exploration.

4 Conclusion

Big data is transforming the modern economy. While many economists have used big data, fewer think about how the use of data by others affects market outcomes. This paper starts to explore the ways in which big data might be incorporated in modern economic and financial theory. One way that big data is used is to help financial market participants make more informed choices about the firms in which they invest. These investment choices affect the prices, cost of capital, and investment decisions of these firms. We set up a very simple model to show how such big data choices might be incorporated and one way in which the growth of big data might affect the real economy. But this is only a modest first step.

One might also consider how firms themselves use data, to refine their products, to broaden their customer market, or to increase the efficiency of their operations. Such data, produced as a by-product of economic activity, might also favor the large firms whose abundant economic activity produces abundant data.

Another step in a big-data agenda would be to consider the sale of data. In many information models, we think of signals that are observed and then embedded in one's knowledge, not easily or credibly transferrable. However data is an asset that can be bought, sold and priced on a market. How does the rise of markets for data change firms choices, investments, evolution and their valuations as firms? If data is a storable, sellable, priced asset, then investment in data should be valued just as if it were investment in a physical asset. Understanding how to price data as an asset might help us to better understand the valuations of new-economy firms and better measure aggregate economic activity.

References

Akcigit, U. and W. Kerr (2017): "Growth through Heterogeneous Innovations," *Journal of Political Economy*, forthcoming.

Asriyan, V. and V. Vanasco (2014): "Informed Intermediation over the Cycle," Stanford University Working Paper.

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

Begenau, J. and J. Salomao (2017): "Firm Financing over the Business Cycle," Stanford Working Paper.

Biais, B., T. Foucault, and S. Moinas (2015): "Equilibrium Fast Trading," *Journal of Financial Economics*, 116, 292–313.

Cooley, T. F. and V. Quadrini (2001): "Financial Markets and Firm Dynamics," *American Economic Review*, 91, 1286–1310.

Cover, T. and J. Thomas (1991): *Elements of information theory*, John Wiley and Sons, New York, New York, first ed.

Davis, S. J. and J. Haltiwanger (2015): "Dynamism Diminished: The Role of Credit Conditions," *in progress.*

Fang, L. and J. Peress (2009): "Media Coverage and the Cross-section of Stock Returns," *The Journal of Finance*, 64, 2023–2052.

Farboodi, M., A. Matray, and L. Veldkamp (2017): "Where Has All the Big Data Gone?" Working Paper, Princeton University.

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

Gomes, J. F. (2001): "Financing Investment," *The American Economic Review*, 91, pp. 1263–1285.

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

Hennessy, C. A. and T. M. Whited (2007): "How Costly Is External Financing? Evidence from a Structural Estimation," *The Journal of Finance*, LXII.

Hennessy, J. and D. Patterson (2011): *Computer Architecture*, Elsevier.

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.

Kozeniauskas, N. (2017): "Technical Change and Declining Entrepreneurship," Working Paper, New York University.

Van Nieuwerburgh, S. and L. Veldkamp (2010): "Information acquisition and under-diversification," *Review of Economic Studies*, 77 (2), 779–805.

Veldkamp, L. (2011): *Information choice in macroeconomics and finance*, Princeton University Press.

A Proofs

A.1 Useful notation, matrices and derivatives

All the following matrices are diagonal with ii entry given by:

1. Average signal precision: $(\bar{\Sigma}_{\eta,t}^{-1})_{ii} = \bar{K}_{i,t}$, where $\bar{K}_{i,t} \equiv \int K_{i,j,t} dj$.
2. Precision of the information prices convey about shock i: $(\Sigma_{p,t}^{-1})_{ii} = \frac{\bar{K}_{i,t}^2}{\rho^2 \sigma_x} = \sigma_{i,p,t}^{-1}$
3. Precision of posterior belief about shock i for an investor j is $\hat{\sigma}_{i,j,t}^{-1}$, which is equivalent to

$$(\hat{\Sigma}_{j,t}^{-1})_{ii} = (\Sigma^{-1} + \Sigma_{\eta,j,t}^{-1} + \Sigma_{p,t}^{-1})_{ii} = \sigma_i^{-1} + K_{i,j,t} + \frac{\bar{K}_{i,t}^2}{\rho^2 \sigma_x} = \hat{\sigma}_{i,j,t}^{-1} \tag{22}$$

4. Average posterior precision of shock i: $\bar{\sigma}_{i,t}^{-1} \equiv \sigma_i^{-1} + \bar{K}_{i,t} + \frac{\bar{K}_{i,t}^2}{\rho^2 \sigma_x}$. The average variance is therefore $(\bar{\Sigma}_t)_{ii} = [\left(\sigma_i^{-1} + \bar{K}_{i,t} + \frac{\bar{K}_{i,t}^2}{\rho^2 \sigma_x}\right)]^{-1} = \bar{\sigma}_{i,t}$.
5. Ex-ante mean and variance of returns: Using Lemma 1 and the coefficients given by (33), (34) and (35), we can write the return as:

$$\begin{aligned} f_t - p_t r_t &= (I - B_t)(f_t - \mu) - C_t x_t + \rho \bar{\Sigma}_t \bar{x}_t \\ &= \bar{\Sigma}_t \left[\Sigma^{-1}(f_t - \mu) + \rho \left(I + \frac{1}{\rho^2 \sigma_x} \bar{\Sigma}_{\eta,t}^{-1'} \right) x_t \right] + \rho \bar{\Sigma}_t \bar{x}_t. \end{aligned}$$

This expression is a constant plus a linear combination of two normal variables, which is also a normal variable. Therefore, we can write

$$f_t - p_t r_t = V_t^{1/2} u_t + w_t, \tag{23}$$

where u_t is a standard normally distributed random variable $u_t \sim N(0, I)$, and w_t is a non-random vector measuring the ex-ante mean of excess returns

$$w_t \equiv \rho \bar{\Sigma}_t \bar{x}_t. \tag{24}$$

and V_t is the ex-ante variance matrix of excess returns:

$$\begin{aligned} V_t &\equiv \bar{\Sigma}_t \left[\Sigma^{-1} + \rho^2 \sigma_x \left(I + \frac{1}{\rho^2 \sigma_x} \bar{\Sigma}_{\eta,t}^{-1'} \right) \left(I + \frac{1}{\rho^2 \sigma_x} \bar{\Sigma}_{\eta,t}^{-1'} \right)' \right] \bar{\Sigma}_t \\ &= \bar{\Sigma}_t \left[\Sigma^{-1} + \rho^2 \sigma_x \left(I + \frac{1}{\rho^2 \sigma_x} (\bar{\Sigma}_{\eta,t}^{-1'} + \bar{\Sigma}_{\eta,t}^{-1}) + \frac{1}{\rho^4 \sigma_x^2} \bar{\Sigma}_{\eta,t}^{-1'} \bar{\Sigma}_{\eta,t}^{-1} \right) \right] \bar{\Sigma}_t \\ &= \bar{\Sigma}_t \left[\Sigma^{-1} + \rho^2 \sigma_x I + (\bar{\Sigma}_{\eta,t}^{-1'} + \bar{\Sigma}_{\eta,t}^{-1}) + \frac{1}{\rho^2 \sigma_x} \bar{\Sigma}_{\eta,t}^{-1'} \bar{\Sigma}_{\eta,t}^{-1} \right] \bar{\Sigma}_t \\ &= \bar{\Sigma}_t \left[\rho^2 \sigma_x I + \bar{\Sigma}_{\eta,t}^{-1'} + \Sigma^{-1} + \bar{\Sigma}_{\eta,t}^{-1} + \Sigma_{p,t}^{-1} \right] \bar{\Sigma}_t \\ &= \bar{\Sigma}_t \left[\rho^2 \sigma_x I + \bar{\Sigma}_{\eta,t}^{-1'} + \bar{\Sigma}_t^{-1} \right] \bar{\Sigma}_t. \end{aligned}$$

The first line uses $E[x_t x_t'] = \sigma_x I$ and $E[(f_t - \mu)(f_t - \mu)'] = \Sigma$, the fourth line uses (36) and the fifth line uses $\bar{\Sigma}_t^{-1} = \Sigma^{-1} + \Sigma_{p,t}^{-1} + \bar{\Sigma}_{\eta,t}^{-1}$.

This variance matrix V_t is a diagonal matrix. Its diagonal elements are:

$$\begin{aligned}(V_t)_{ii} &= (\bar{\Sigma}_t \left[\rho^2\sigma_x I + \bar{\Sigma}_{\eta,t}^{-1} + \bar{\Sigma}_t^{-1}\right] \bar{\Sigma}_t)_{ii} \\ &= \bar{\sigma}_{i,t}[1 + (\rho^2\sigma_x + \bar{K}_{i,t})\bar{\sigma}_{i,t}]. \end{aligned} \tag{25}$$

A.2 Solving the Model

Step 1: Portfolio Choices From the FOC, the optimal portfolio is chosen by investor j is

$$q_{j,t}^* = \frac{1}{\rho}\hat{\Sigma}_{j,t}^{-1}(\hat{E}_{j,t}[f_t] - p_t r_t). \tag{26}$$

where $\hat{E}_{j,t}[f_t]$ and $\hat{\Sigma}_{j,t}$ depend on the skill of the investor.
Next, we compute the portfolio of the average investor.

$$\begin{aligned}\bar{q} \equiv \int q_{j,t}^* dj &= \frac{1}{\rho}\int \hat{\Sigma}_{j,t}^{-1}(\hat{E}_{j,t}[f_t] - p_t r_t)dj \\ &= \frac{1}{\rho}\left(\int \Sigma_{\eta,j,t}^{-1}\eta_{j,t}dj + \Sigma_{p,t}^{-1}\eta_{p,t} + \bar{\Sigma}_t^{-1}(\mu - p_t r_t)\right) \\ &= \frac{1}{\rho}\left(\bar{\Sigma}_{\eta,t}^{-1}f_t + \Sigma_{p,t}^{-1}\eta_{p,t} + \bar{\Sigma}_t^{-1}(\mu - p_t r_t)\right), \end{aligned} \tag{27}$$

where the fourth equality uses the fact that average noise of private signals is zero.

Step 2: Clearing the asset market and computing expected excess return Lemma 1 describes the solution to the market-clearing problem and derives the coefficients A_t, B_t, and C_t in the pricing equation. The equilibrium price, along with the random signal realizations determines the interim expected return $(\hat{E}_{j,t}[f_t] - p_t r_t)$. But at the start of the period, the equilibrium price and one's realized signals are not known. To compute beginning-of-period utility, we need to know the ex-ante expectation and variance of this interim expected return.

The interim expected excess return can be written as: $\hat{E}_{j,t}[f_t] - p_t r_t = \hat{E}_{j,t}[f_t] - f_t + f_t - p_t r_t$ and therefore its variance is:

$$V_t[\hat{E}_{j,t}[f_t] - p_t r_t] = V_t[\hat{E}_{j,t}[f_t] - f_t] + V_t[f_t - p_t r_t] + 2\,\mathrm{Cov}_t[\hat{E}_{j,t}[f_t] - f_t, f_t - p_t r_t]. \tag{28}$$

Combining (11) with the definitions $\eta_{j,t} = f_t + \varepsilon_{j,t}$ and $\eta_{p,t} = f_t + \varepsilon_{p,t}$, we can compute expectation errors:

$$\begin{aligned}\hat{E}_{j,t}[f_t] - f_t &= \hat{\Sigma}_{j,t}\left[(\Sigma^{-1}\mu + (\Sigma_{\eta,j,t}^{-1} + \Sigma_{p,t}^{-1})f_t + \Sigma_{\eta,j,t}^{-1}\varepsilon_{j,t} + \Sigma_{p,t}^{-1}\varepsilon_{p,t}\right] - f_t \\ &= \hat{\Sigma}_{j,t}\left[-\Sigma^{-1}(f_t - \mu) + \Sigma_{\eta,j,t}^{-1}\varepsilon_{j,t} + \Sigma_{p,t}^{-1}\varepsilon_{p,t}\right]\end{aligned}$$

Computing the expectation, we obtain $E_t[\hat{E}_{j,t}[f_t] - f_t] = \hat{\Sigma}_{j,t}\hat{\Sigma}_{j,t}^{-1}\mu - \mu = 0$ and its variance is $V_t[\hat{E}_{j,t}[f_t] - f_t] = \hat{\Sigma}_{j,t}\left[\Sigma^{-1} + \Sigma_{\eta,j,t}^{-1} + \Sigma_{p,t}^{-1}\right]\hat{\Sigma}_{j,t}' = \hat{\Sigma}_{j,t}$.

From (23) we know that $V_t[f_t - p_t r_t] = V_t$. To compute the covariance term, we can rearrange the definition of $\eta_{p,t}$ to get $p_t r_t = B_t\eta_{p,t} + A_t - B_t\mu$ and $\eta_{p,t} = f_t + \varepsilon_{p,t}$ to write

$$\begin{aligned}f_t - p_t r_t &= (I - B_t)f_t - A_t - B_t\varepsilon_{p,t} + B_t\mu &(29)\\ &= \rho\bar{\Sigma}_t\bar{x}_t + \bar{\Sigma}_t\Sigma^{-1}(f_t - \mu) - (I - \bar{\Sigma}_t\Sigma^{-1})\varepsilon_{p,t} &(30)\end{aligned}$$

where the second line comes from substituting the coefficients A_t and B_t from Lemma 1. Since the constant $\rho\bar{\Sigma}_t\bar{x}_t$ does not affect the covariance, we can write

$$\begin{aligned}\mathrm{Cov}_t[\hat{E}_{j,t}[f_t]-f_t, f_t-p_tr_t] &= \mathrm{Cov}[-\hat{\Sigma}_{j,t}\Sigma^{-1}(f_t-\mu)+\hat{\Sigma}_{j,t}\Sigma_{p,t}^{-1}\varepsilon_{p,t}, \bar{\Sigma}_t\Sigma^{-1}(f_t-\mu)-(I-\bar{\Sigma}_t\Sigma^{-1})\varepsilon_{p,t}] \\ &= -\hat{\Sigma}_{j,t}\Sigma^{-1}\Sigma\Sigma^{-1}\bar{\Sigma}_t-\hat{\Sigma}_{j,t}\Sigma_{p,t}^{-1}\Sigma_{p,t}(I-\Sigma^{-1}\bar{\Sigma}_t)] \\ &= -\hat{\Sigma}_{j,t}\Sigma^{-1}\bar{\Sigma}_t-\hat{\Sigma}_{j,t}(I-\Sigma^{-1}\bar{\Sigma}_t) \;=\; -\hat{\Sigma}_{j,t}\end{aligned}$$

Substituting the three variance and covariance terms into (28), we find that the variance of excess return is $V_t[\hat{E}_{j,t}[f_t]-p_tr_t]=\hat{\Sigma}_{j,t}+V_t-2\hat{\Sigma}_{j,t}=V_t-\hat{\Sigma}_{j,t}$. Note that this is a diagonal matrix. Substituting the expressions (25) and (22) for the diagonal elements of V_t and $\hat{\Sigma}_{j,t}$ we have

$$(V_t[\hat{E}_{j,t}[f_t]-p_tr_t])_{ii}=(V_t-\hat{\Sigma}_{j,t})_{ii}=(\bar{\sigma}_{i,t}-\hat{\sigma}_{i,j,t})+(\rho^2\sigma_x+\bar{K}_{i,t})\bar{\sigma}_{i,t}^2$$

In summary, the excess return is normally distributed as $\hat{E}_{j,t}[f_t]-p_tr_t\sim\mathcal{N}(w_t, V_t-\hat{\Sigma}_{j,t})$.

Step 3: Compute ex-ante expected utility Ex-ante expected utility for investor j is $U_{j,t}=E_t\left[\rho\hat{E}_{j,t}[\hat{W}_{j,t}]-\frac{\rho^2}{2}\hat{V}_{j,t}[\hat{W}_{j,t}]\right]$. In period 2, the investor has chosen his portfolio and the price is in his information set, therefore the only payoff-relevant, random variable is f_t. We substitute the budget constraint in the optimal portfolio choice from (26) and take expectation and variance conditioning on $\hat{E}_{j,t}[f_t]$ and $\hat{\Sigma}_{j,t}$ to obtain $U_{j,t}=\rho r_tW_t+\frac{1}{2}E_t[(\hat{E}_{j,t}[f_t]-p_tr_t)'\hat{\Sigma}_{j,t}^{-1}(\hat{E}_{j,t}[f_t]-p_tr_t)]$.

Define $m_t\equiv\hat{\Sigma}_{j,t}^{-1/2}(\hat{E}_{j,t}[f_t]-p_tr_t)$ and note that $m_t\sim\mathcal{N}(\hat{\Sigma}_{j,t}^{-1/2}w_t, \hat{\Sigma}_{j,t}^{-1/2}V_t\hat{\Sigma}_{j,t}^{-1/2'}-I)$. The second term in the $U_{i,t}$ is equal to $E[m_t'm_t]$, which is the mean of a non-central Chi-square. Using the formula, if $m_t\sim N(E[m_t], Var[m_t])$, then $E[m_t'm_t]=\mathrm{tr}(Var[m_t])+E[m_t]'E[m_t]$, we get

$$U_{j,t}=\rho r_tW_t+\frac{1}{2}\mathrm{tr}(\hat{\Sigma}_{j,t}^{-1/2}V\hat{\Sigma}_{j,t}^{-1/2'}-I)+\frac{1}{2}w_t'\hat{\Sigma}_{j,t}^{-1}w_t=\rho r_tW_t+\frac{1}{2}\mathrm{tr}(\hat{\Sigma}_{j,t}^{-1}V)-\mathrm{tr}(I)+\frac{1}{2}w_t'\hat{\Sigma}_{j,t}^{-1}w_t.$$

Finally, we substitute the expressions for $\hat{\Sigma}_{j,t}^{-1}$ and w_t from (22) and (24):

$$\begin{aligned}U_{j,t} &= \rho r_tW_t-\frac{N}{2}+\frac{1}{2}\sum_{i=1}^{N}\left(\sigma_i^{-1}+K_{i,j,t}+\frac{\bar{K}_{i,t}^2}{\rho^2\sigma_x}\right)(V_t)_{ii}+\frac{\rho^2}{2}\sum_{i=1}^{N}\bar{x}_{i,t}^2\bar{\sigma}_{i,t}^2\left(\sigma_i^{-1}+K_{i,j,t}+\frac{\bar{K}_{i,t}^2}{\rho^2\sigma_x}\right) \\ &= \frac{1}{2}\sum_{i=1}^{N}K_{i,j,t}[(V_t)_{ii}+\rho^2\bar{x}_{i,t}^2\bar{\sigma}_{i,t}^2]+\rho r_tW_t-\frac{N}{2}+\frac{1}{2}\sum_{i=1}^{N}\left(\sigma_i^{-1}+\frac{\bar{K}_{i,t}^2}{\rho^2\sigma_x}\right)[(V_t)_{ii}+\rho^2\bar{x}_{i,t}^2\bar{\sigma}_{i,t}^2] \\ &= \frac{1}{2}\sum_{i=1}^{N}K_{i,j,t}\lambda_{i,t}+\text{constant} \qquad (31)\end{aligned}$$

$$\lambda_{i,t}=\bar{\sigma}_{i,t}[1+(\rho^2\sigma_x+\bar{K}_{i,t})\bar{\sigma}_{i,t}]+\rho^2\bar{x}_{i,t}^2\bar{\sigma}_{i,t}^2 \qquad (32)$$

where the weights $\lambda_{i,t}$ are given by the variance of expected excess return $(V_t)_{ii}$ from (25) plus a term that depends on the supply of the risk.

Step 4: Information choices The attention allocation problem maximizes ex-ante utility in (31) subject to the information capacity, data availability and no-forgetting constraints (16), (17) and (10). Observe that $\lambda_{i,t}$ depends only on parameters and on aggregate average precisions. Since each investor has zero mass within a continuum of investors, he takes $\lambda_{i,t}$ as given. Since the constant is irrelevant, the optimal choice maximizes a weighted sum of attention allocations, where the weights are given by $\lambda_{i,t}$

(equation (18)), subject to a constraint on an un-weighted sum. This is not a concave objective, so a first-order approach will not deliver a solution. A simple variational argument reveals that allocating all capacity to the risk(s) with the highest $\lambda_{i,t}$ achieves the maximum utility. For a formal proof of this result, see Van Nieuwerburgh and Veldkamp (2010). Thus, the solution is given by: $K_{i,j,t} = K_t$ if $\lambda_{i,t} = \max_k \lambda_{k,t}$, and $K_{i,j,t} = 0$, otherwise. There may be multiple risks i that achieve the same maximum value of $\lambda_{i,t}$. In that case, the manager is indifferent about how to allocate attention between those risks. We focus on symmetric equilibria.

A.3 Proofs

Proof of Lemma 1

Proof. Following Admati (1985), we know that the equilibrium price takes the following form $p_t r_t = A_t + B_t(f_t - \mu) + C_t x_t$ where

$$A_t = \mu - \rho\bar{\Sigma}_t\bar{x}_t \tag{33}$$
$$B_t = I - \bar{\Sigma}_t\Sigma^{-1} \tag{34}$$
$$C_t = -\rho\bar{\Sigma}_t\left(I + \frac{1}{\rho^2\sigma_x}\bar{\Sigma}_{\eta,t}^{-1'}\right) \tag{35}$$

and therefore the price is given by $p_t r_t = \mu + \bar{\Sigma}_t\left[(\bar{\Sigma}_t^{-1} - \Sigma^{-1})(f_t - \mu) - \rho(\bar{x}_t + x_t) - \frac{1}{\rho\sigma_x}\bar{\Sigma}_{\eta,t}^{-1'}x_t\right]$. Furthermore, the precision of the public signal is

$$\Sigma_{p,t}^{-1} \equiv \left(\sigma_x B_t^{-1}C_tC_t'B_t^{-1'}\right)^{-1} = \frac{1}{\rho^2\sigma_x}\bar{\Sigma}_{\eta,t}^{-1'}\bar{\Sigma}_{\eta,t}^{-1} \tag{36}$$

□

Proof of Lemma 2 See Kacperczyk et. al (2016).

A.4 Firm Volatility Data

The introduction of our paper claims that differential trends in the volatility of large and small firms' earnings is not a plausible explanation for the different trends in the cost of capital. To support this claim, we explore whether the volatility of large and small firms has diverged. We find some fluctuations, but no consistent trend in the difference.

Our volatility measure is based on the annual growth rate in earnings calculated at the firm level from quarterly CRSP/Computat data from 1960 - 2016. Earnings are constructed by multiplying basic earnings per share, excluding extraordinary items (EPS) by the number of shares used to calculate EPS. We measure the volatility of earnings growth as the rolling standard deviation over the past 20 quarters. The firms are split by size in a number of ways: firstly, the firms in the sample are split by whether or not they are (at that time) a member of the S&P500 index. Secondly, we split firms by whether or not they were in the top half of the earnings distribution in each quarter. Lastly, we consider only firms in the bottom and top quartiles of the earnings distribution. In the plots below, the dashed lines are the median volatility, whilst the solid line is the trend extracted from this series using a HP filter, with $\lambda = 1600$.

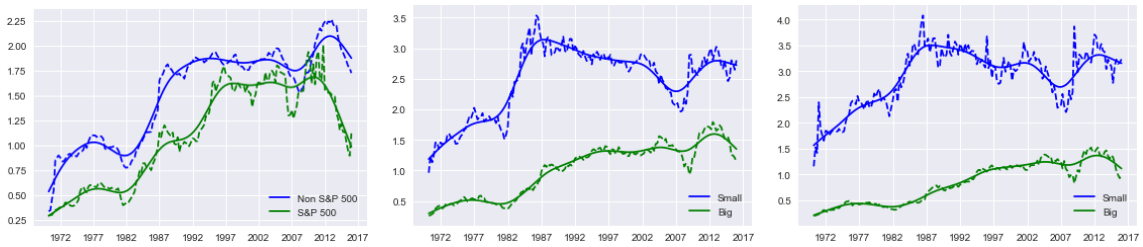


Figure 9: 20 month rolling standard deviation of growth in earnings: Large and small firms. Left panel: Median volatility by whether a firm is a member of the S&P500 index. Middle panel: Median volatility by whether a firm has earnings above or below the median each quarter. Right panel: Median volatility by whether a firm's earnings are in the top or bottom quartile each quarter

Figure 9 plots volatility for large and small firms, over time. Whether the firms are cut at the median, the top and bottom quartiles, or by membership in the S&P 500, in every case, there are fluctuations in volatility, and there are long-run increases in volatility. But there is no consistent long-run trend in the gap between the different sized firms.