Machine-Readable Data and Financial Experts in Asset Management

Junli Zhao*

January 2021

Click here for the latest version.

Abstract

Should financial experts (e.g., buy-side asset managers and analysts) fear the rise of algorithms? As machine-readable (clean and structured) data are essential for the development and functioning of algorithms, I study this question by investigating whether financial experts benefit from more machine-readable data in information production in asset management. I first develop a model in which an institutional investor's performance and asset holdings depend on two inputs: the amount of machine-readable data and the number of financial experts, and derive how changes induced by an increase in the amount of machine-readable data depend on the relation between the two inputs. Exploiting an exogenous regulatory shock that makes corporate filings more machine-readable, I find that institutions with more financial experts experience larger performance improvement than institutions with fewer financial experts, consistent with financial experts benefiting from more machine-readable data. This result helps evaluate the disruption brought by modern algorithms.

Keywords: Information Technology; Skilled Labor; Information Acquisition.

*Email: junli.zhao@hec.edu. I thank Jean-Edouard Colliard, François Derrien, Olivier Dessaint, Thierry Foucault, Johan Hombert, Joel Peress, Christophe Spaenjers, and participants at HEC Paris brownbag for helpful comments. All errors are my own.

1 Introduction

Computer algorithms are transforming the financial industry, with a potentially large impact on its labor force, in particular financial experts. On the one hand, these algorithms boost the productivity of financial experts by automating routine but complex tasks, enabling financial experts to generate more value. For example, Goldman Sachs' proprietary software system SecDB helps its financial experts evaluate the impact of the trades they propose.[1] On the other hand, algorithms have the potential to displace financial experts. For instance, robo-analysts are able to generate recommendations faster and better than human analysts (Coleman et al., 2020). As computers now manage about 35% of US public equities (vs. 24% for human asset managers),[2] understanding how computer algorithms may disrupt financial institutions and the finance labor market is important.

As clean and well-structured (machine-readable) datasets are vital for the development and functioning of computer algorithms, identifying the relation between machine-readable data and financial experts helps answer the previous question. Focusing on the production of information in asset management, this paper aims at investigating whether machine-readable data help financial experts generate more precise information (complementarity) or make them less essential in its production (substitution).

I provide evidence that machine-readable data complement financial experts. I first develop a model in which a financial institution's performance and holdings depend on two inputs: the amount of machine-readable data it has and the number of financial experts it employs. The model suggests that the relation between the two inputs can be inferred using a shock that increases the amount of machine-readable data. Such a shock helps all institutions generate better information. However, institutions with more financial experts benefit more (less) if the two inputs are complements (substitutes). The model derives predictions that link the impact of a data shock on institutions' excess returns and asset holdings to the degree of complementarity. I then test which relation holds empirically using the SEC's eXtensible Business Reporting Language (XBRL) mandate in 2009 as an exogenous shock. All the results are consistent with the hypothesis that machine-readable data and financial experts are complements rather than substitutes.

[1] See, *Understanding SecDB: Goldman Sachs's Most Valued Trading Weapon*, the Wall Street Journal, Sept. 7, 2016

[2] The rest is owned by other investors, such as individuals and companies that are not asset managers. See, *March of the machines*, the Economist, June 11, 2019.

The exogenous regulatory shock, the SEC's eXtensible Business Reporting Language (XBRL) mandate, is an important ingredient for the empirical testing. This mandate requires firms to provide a machine-readable version of their corporate filings (10-K, 10-Q, etc.) using the XBRL format. Data items in the XBRL files are tagged with standard taxonomies. This feature makes it much easier for computers to extract information such as numbers in footnotes or numbers scattered in long paragraphs of texts, increasing the amount of data that are ready for large-scale computer-based analysis. Despite different organization of data, the XBRL filings contain the same information as traditional text filings. The XBRL mandate was implemented through a three-year phase-in period from 2009 to 2011, during which large, medium, and small firms complied successively. The staggered implementation provides a set-up to employ a (triple) difference-in-differences method.

To derive predictions that identify the relation between the two inputs, I develop a model in which each institutional investor trades one risk-free asset and several risky assets. Each institution can condition its demand on market prices and a private signal about the payoffs of the risky assets. The precision of an institution's signal is increasing in the amount of machine-readable data it has and the number of financial experts it employs. The amount of machine-readable data of an institution is increasing in the number of computer scientists it employs and decreasing in the cost of data processing. I then consider comparative statics after an exogenous decrease in the cost of data processing, which affects the institutions in two ways.

First, the shock has a *direct effect*: it increases the amount of machine-readable data directly, which in turn improves the investors' signal precision. How this improvement affects institutions with different number of financial experts depends on whether the two inputs are complements or substitutes. If they are complements, more machine-readable data improves the productivity of financial experts, the shock benefits institutions with a large number of financial experts more. Given an increase in the amount of machine-readable data with the same magnitude, the signal precision of institutions with more financial experts increases relative to institutions with fewer financial experts. If instead the two inputs are substitutes, an increase in the amount of machine-readable data decreases the marginal value of financial experts. As a result, the signal precision of institutions with more financial experts decreases relative to institutions with fewer financial experts.

Second, the shock has a *market price effect*: after the shock all institutions produce more

precise information, and hence market prices become more informative. When updating their beliefs about the asset payoffs, institutions with less precise information put a higher weight on information contained in the market prices than other institutions. They thus benefit more from the shock, resulting in a *leveling effect* that reduces their information gap relative to other institutions.

More precise information decreases uncertainty, and thus increases performance and decreases dispersion in holdings (standard deviation of holdings on the same stock across institutions). Stock ownership may be reallocated. Combining the two effects described above gives the following predictions on equilibrium performance, stock ownership, and dispersion in asset holdings, depending on whether the two inputs are complements or substitutes:

(i) The performance of institutions with more financial experts on the treated stock *does not decrease* (*decreases*) relative to that of institutions with fewer financial experts, given the same number of computer scientists;

(ii) The fraction of stock shares owned by institutions with more financial experts *does not decrease* (*decreases*) relative to that of institutions with fewer financial experts, given the same number of computer scientists;

(iii) Dispersion of holdings across finance-intensive institutions *does not increase* (*increases*) relative to that across base type institutions.

An empirical investigation thus requires information on institutional investors' labor force, for which I use their foreign high skilled labor (H-1B visa application) data. The numbers of IT- and finance-related workers, scaled by the institution's assets, provide a proxy for its labor inputs. The scaled employment is labeled as high (low) if it falls in the top (bottom) tercile of its distribution. With labels in both dimensions, I classify institutions into the four types: (1) *base* type, i.e., low employment of both types of workers; (2) *IT-intensive* type, high employment of IT workers but low employment of finance workers; (3) *finance-intensive* type, high employment of finance workers but low employment of IT workers; and (4) *bi-intensive* type, high employment of both types of workers. The classification is then verified by a manual check on some of the largest institutions in the data. Consistent with expectation, institutions well-known for their quantitative investment strategies are classified as IT-intensive while those classified as finance-intensive are more often associated

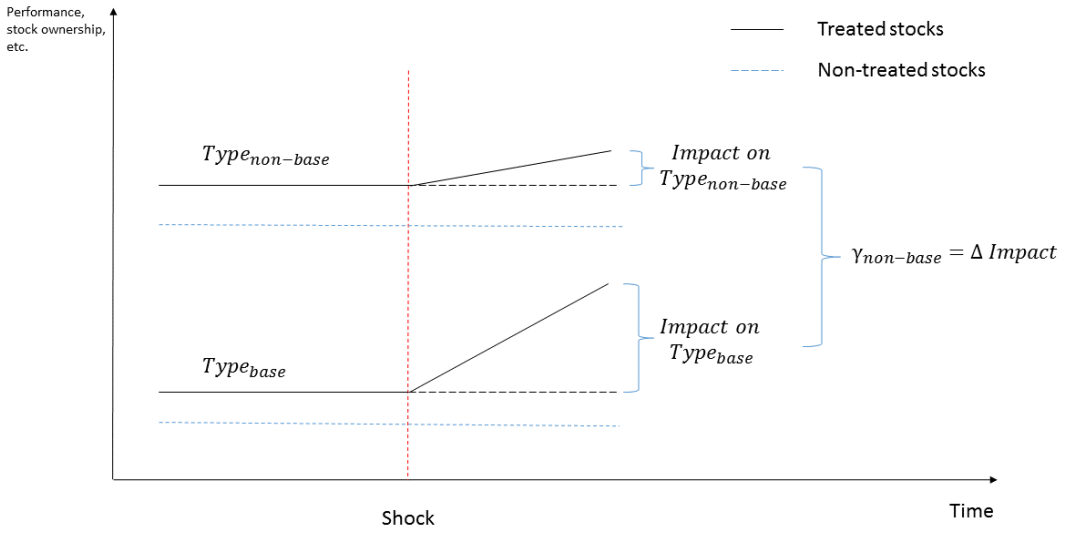

Figure 1: Illustration of differential impacts of the shock on $Type_{non-base}$ (IT-intensive, finance-intensive or bi-intensive) institutions and the base type institutions. The dependent variable is performance, stock ownership or dispersion of holdings.

with fundamental analysis.

To test which relation holds empirically, I compare differences across institutional investors in their performance and holding on the treated and control stocks, before and after the shock (Figure 1). In the analysis, I only consider two quarters before and after each event.

The predictions on performance are tested using the institutions' excess returns. For an institution, its excess return on stock j is computed as the product of its excess holding on asset j, relative to the market average holding, and the return of the stock. The annualized return of finance-intensive institutions increases by about 4 basis points on one treated stock relative to the base types after the XBRL shock, which is in line with financial experts benefiting from more machine-readable data. Similarly, the annualized return of bi-intensive type institutions increases by about 0.25 basis points on one treated stock relative to IT-intensive type institutions after the XBRL shock. As a byproduct, I find that the annualized return difference between the IT-intensive types and the base types decreases by about 3 basis point per stock after the shock. This result implies that IT-intensive institutions lose part of their informational advantage after the shock, which is consistent with that computer scientists help aggregate data and the information production function is concave in the amount of machine-readable data. Using changes in excess holdings as a proxy for trading, I

also run the tests with returns from trading and obtain similar results.

For each stock in each quarter, I compute the fraction of total shares outstanding held by institutions of each type. Comparing to base type institutions, the fraction of total shares outstanding of the treated stocks owned by IT-intensive institutions decreases by 0.6 percentage point. This again is consistent with the hypothesis that the informational advantage of IT-intensive institutions is smaller after the shock. The shock has no significant impact on the fraction owned by base-type institutions and finance-intensive institutions at face value. One explanation for this result is that finance-intensive institutions are smaller than other institutions in the sample and thus changes in their fraction is mechanically smaller. To address this issue, I divide the fraction by the total assets of each type and obtain a measure of ownership per dollar for each stock and each type of institution. I find that, relative to the non-treated stocks, the ownership per dollar of treated stocks increases for base type institutions and decreases for IT-intensive institutions after the implementation of the XBRL mandate. The increase in ownership per dollar is higher for finance-intensive institutions. These results are also consistent with complementarity.

The tests on the dispersion of excess holdings provide additional evidence for complementarity. Following the XBRL shock, the standard deviation of excess holding on the treated stocks decreases by 100 basis points among base type institutions, compared to the non-treated stocks. In addition, compared to the base type, the effect is larger for the finance-intensive type (about 250 basis points).

As IT-intensive institutions are larger and finance-intensive institutions are smaller than the base type institutions, one concern is that the results are driven by institution size. As a robustness check, I repeat the tests using a matched sample. I use coarsened exact matching on assets under management and turnover. Based on their values in the two variables, institutions are put into two-dimensional bins. Institutions within the same bin are matched. I construct a sample which only keeps institutions matched with the base type institutions. Differences in size are much smaller in this sample. Tests using this sample give results similar to the baseline analysis.

The rest of the paper is organized as follows. In Section 2, I lay out the theoretical framework and empirical predictions. Section 3 describes the shock and data in details. Main results are presented in Section 4. Section 5 concludes.

Related literature. This paper contributes to a surging literature on the impact of IT,

especially robots and artificial intelligence (AI), on labor. See, e.g., David (2015), Acemoglu and Restrepo (2018), Brynjolfsson et al. (2018) and Webb (2019). Based on regional data, Akerman et al. (2015) find that broadband internet improves the productivity of high skilled workers. Several papers show how "machines" can complement human experts by reducing behavioral biases and providing new insights, in e.g., earning forecast (van Binsbergen et al., 2020), real asset valuation (Aubry et al., 2019), or the selection of corporate directors (Erel et al., 2018).

Different from the previous literature, which mostly relies on simulation, this paper, using actual investment behavior of institutional investors, provides evidence that high skilled financial workers benefit from modern information technologies. Coleman et al. (2020) show that computers can generate investment recommendations faster and more accurate than human analysts. Grennan and Michaely (2020) find that sell-side analysts are more likely to shift their coverage or even leave the profession when their portfolio stocks are more exposed to AI analysis. Complementing their research, this paper suggests that access to better in-house technologies can help analysts mitigate the negative impact of AI on them.

Secondly, this paper is related to a fast growing literature studying the effect of information technologies on financial institutions and markets. Modern information technologies have enlarged institutions' portfolio strategies (Abis, 2017). By easing access to public information and providing alternative datasets (Grennan and Michaely, 2019), modern IT technologies make price efficiency higher (Gao and Huang, 2020), investors disagree less (Chang et al., 2020), and corporations invest more (Goldstein et al., 2020). They also potentially contribute to the trend of rising price informativeness ((Bai et al., 2016), Farboodi et al. (2020)). Several studies investigate how improved technologies affect investors' information production. Farboodi and Veldkamp (2020) argue that technological progress in data processing can lead investors to rely more on data about others' demand than fundamentals but both types of data continue to be processed. Dugast and Foucault (2018) highlight the possibility that, with abundant data, investors might choose to rely more on raw signals than waiting for processed signals, which may reduce price informativeness. By looking into the information production function and different types of workers, this paper provides evidence that information technologies complement financial experts, especially by generating non-traditional information. Focusing on a different dichotomy, Abis and Veldkamp (2020)

study how the productivity of data aggregation and data analysis skills in asset management evolved in the past decade. The results in this paper lend support to their assumption that the two skills are complements.

Thirdly, this paper is related to the large literature on information acquisition (e.g., Grossman and Stiglitz (1980), Verrecchia (1982); see Veldkamp (2011) for a survey). The precision of investors' signals plays an important role in this literature. Kacperczyk et al. (2016) develop a theory on how investors choose the precision of signals when they face an information capacity constraint. Van Nieuwerburgh and Veldkamp (2010) relate such information choice to mutual fund under-diversification. My paper puts a structure on signal precision that links precision and labor inputs, and proposes a measure that can sort financial institutions by their information capacity.

Lastly, this paper is related to the literature on the impacts of the XBRL mandate and the empirical literature that exploited the H-1B visa program. Some studies suggest that the adoption of XBRL lowers information-aggregation costs. For example, using the XBRL Voluntary Filing Program before the mandate, Efendi et al. (2016) find that the market reaction is larger when XBRL reports are filed, indicating XBRL files have higher information value than HTML files. Other findings indicate that XBRL reduces event return volatility (Kim et al., 2012), analyst forecast dispersion (Liu et al., 2014), stock price synchronicity (Dong et al., 2014), increases breadth of ownership (Kim et al., 2019), and quantitative footnote disclosure (Blankespoor, 2019). Bhattacharya et al. (2018) find that small institutions benefit more than large institutions from the mandate. This paper further shows that institutions with fewer resources in IT benefit more. Several studies have exploited H-1B data to study the effect of high skilled labor on start-up success (Dimmock et al., 2019), the impact of skilled immigrant labor on innovation (Kerr and Lincoln, 2010), employment structure (Kerr et al., 2015), native wages and employment (Peri et al., 2015) and the cross-section of equity return (Sharifkhani, 2018).

2 Theoretical Framework and Empirical Predictions

In this section, I first solve a simplified multi-asset noisy rational expectation equilibrium model (Admati (1985), Kacperczyk et al. (2016)). The multi-asset setting matches the empirical methodology, where at a given time some assets are affected by the XBRL shock while

other assets act as control. The comparative statics of the model give empirical predictions on how a shock on data processing cost affects the institutions' performance and holding dispersion.

There are multiple risky assets, indexed by j, with payoff $f = \mu + z$, where $z \sim N(0, \Sigma)$. Σ is the prior variance and a diagonal matrix. The supply of risky assets is $\bar{x} + x$, where $x \sim N(0, \sigma_x^2 I)$. I is the identity matrix. There is a riskless asset with return r.

A measure 1 of investors, indexed by i, have mean-variance utility: $U_i = E[W_i] - \frac{\rho}{2} Var[W_i]$, where $W_i = (W_{0i} - q_i' p)r + q_i' f$ is the end-of-period wealth of investor i and W_{0i} is her initial wealth. q_i is the vector of asset quantities investor i chooses to hold.

Before trading, investors receive a signal about the assets' payoffs, the precision of which depends on the amount of machine-readable data it has and the number of financial experts it employs. The signals η_{si} are independent and identically distributed across investors, with $\eta_{si} = f + \epsilon_{si}$, where $\epsilon_{si} \sim N(0, K_i^{-1})$. K_i is also diagonal. Its jth diagonal element $(K_i)_{jj}$ is the precision of investor i's signal on asset j. I assume $(K_i)_{jj} = K(D_{ij}, n_{fi})$. D_{ij} is the amount of machine-readable data of asset j that investor i has. n_{fi} is the number of finance specialists employed by investor i. The information production function is increasing in D_{ij} and n_{fi}, i.e., $\frac{\partial K}{\partial D} > 0$ and $\frac{\partial K}{\partial n_f} > 0$. Computer scientists can help aggregate machine-readable data, for example, by using programs to parse traditional text corporate filings. For a given number of computer scientists, the amount of machine-readable data also depends on the difficulty or cost of processing data. To take the two observations into consideration, I assume that $D_{ij} = D(c_j, n_{pi})$ with $\frac{\partial D}{\partial c} < 0$ and $\frac{\partial D}{\partial n_p} > 0$. c_j is the cost of processing data of asset j. n_{pi} is the number of IT specialists ("programmers") employed by investor i.

The goal of this paper is to identify the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$: a positive sign means that machine-readable data and financial experts are complements, whereas a negative sign indicates substitutability.

Investors are characterized by their inputs (n_{pi}, n_{fi}). In the model, the inputs are exogenously given. To be closer to the empirical setting, I focus on a case where n_{pi} and n_{fi} are either low or high, $n_{pi} \in \{\underline{n}_p, \overline{n}_p\}$ and $n_{fi} \in \{\underline{n}_f, \overline{n}_f\}$. The model thus features four types of investors: base type with $(\underline{n}_p, \underline{n}_f)$, IT-intensive type with $(\overline{n}_p, \underline{n}_f)$, finance-intensive type with $(\underline{n}_p, \overline{n}_f)$ and bi-intensive type with $(\overline{n}_p, \overline{n}_f)$. The measure of type (n_p, n_f) investors is denoted by $\mu(n_p, n_f)$ and constant throughout the model.

Conjecturing that the prices provide an unbiased signal about f, $\eta_p = f + \epsilon_p$, where

$\epsilon_p \sim N(0, \Sigma_p)$, one can show the following result.

Lemma 1. *There is an equilibrium such that the price p and investors' equilibrium holding q_i are given by*

$$p = \frac{1}{r}(A + Bf + Cx)$$
$$q_i = \frac{1}{\rho}\hat{\Sigma}_i^{-1}(\hat{\mu}_i - pr)$$

where

$$\hat{\mu}_i = E[f|\eta_{si}, p] = \hat{\Sigma}_i(\Sigma^{-1}\mu + \Sigma_p^{-1}\eta_p + K_i\eta_{si})$$
$$\hat{\Sigma}_i = var[f|\eta_{si}, p] = (\Sigma^{-1} + \Sigma_p^{-1} + K_i)^{-1}$$

A, B, C and Σ_p are given in Appendix.

The result shows that the equilibrium holding of investor i depends on the posterior variance of her belief $\hat{\Sigma}_i$ and the expected payoff $\hat{\mu}_i - pr$.

As signal precision is not directly observed in the data, I rely on expected returns, stock ownership, and holding dispersion, which can be measured empirically, to derive empirical predictions. One commonly used measure of expected return is excess return (see, e.g., Kacperczyk et al. (2016)). For an investor i, the excess return on asset j is defined as the unconditional expectation of the product of excess holding relative to the market, $q_{ij} - \bar{q}_j$, and excess payoff of one unit of the asset, $f_j - p_j r$, namely,

$$E[R_{ij}(n_{pi}, n_{fi})] = E[(q_{ij} - \bar{q}_j)(f_j - p_j r)]$$

Here q_{ij} is the quantity of asset j held by investor i. $\bar{q}_j = \int_i q_{ij}\, di$ is the average holding on asset j across investors and equals $\bar{x} + x$ due to market clearing. The excess holding filters out holding due to market information and adjusts for risk. It is more sensitive to private information than the gross holdings q_{ij}. Averaging across investors with the same labor composition (n_p, n_f), a simple calculation gives that

$$E[R_j(n_p, n_f)] = (\rho\bar{x}_j^2\bar{\sigma}_j^2 + \frac{1}{\rho}v_j)(K(D(c_j, n_p), n_f) - \bar{K}_j) \tag{1}$$

Here $\bar{K}_j = \int (K_i)_{jj}\, di$ is the average signal precision on asset j. $\bar{\sigma}_j = (\bar{\Sigma}^{-1})_{jj} = \int (\hat{\Sigma}_i^{-1})_{jj}\, di = (\Sigma^{-1})_{jj} + (\Sigma_p^{-1})_{jj} + \bar{K}_j$ is the average posterior variance of payoff on asset j. $v_j = \bar{\sigma}_j^2(\rho^2\sigma_x^2 + \bar{K}_j) + \bar{\sigma}_j$ measures the unconditional variance of the excess payoff $f_j - p_j r$. Given market variables ($x_j, \bar{\sigma}_j, \bar{K}_j$ and v_j), investors' excess return is increasing in their signal precision K_j.

To derive predictions that are informative about $\frac{\partial^2 K}{\partial D \partial n_f}$, I consider how excess returns change after an exogenous decrease in c_j. A smaller c_j implies more machine-readable data and thus better information on asset j for each investor. Depending on labor composition and the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$, the impact ($\frac{\partial E[R_j(n_p, n_f)]}{\partial c_j}$) is heterogeneous across different types of investors, as shown by the following results.

Proposition 1. *An exogenous decrease in c_j does not affect excess returns on asset j' for all $j' \neq j$. For excess return on asset j,*

(i) If $\frac{\partial^2 K}{\partial D \partial n_f} < 0$, $\frac{\partial E[R_j(n_p, \overline{n}_f)]}{\partial c_j} < \frac{\partial E[R_j(n_p, \underline{n}_f)]}{\partial c_j}$. That is substitution implies that investors with more financial experts benefit less from the shock than investors with fewer financial experts.
(ii) $\frac{\partial E[R_j(n_p, \overline{n}_f)]}{\partial c_j} > \frac{\partial E[R_j(n_p, \underline{n}_f)]}{\partial c_j}$ only if $\frac{\partial^2 K}{\partial D \partial n_f} > 0$. A situation where investors with more financial experts benefit more implies complementarity.

Note that part (i) is a sufficient condition while part (ii) gives a necessary condition. These results are intuitive. Other assets than j are not affected because asset returns are not correlated. The heterogeneous effects of an increase in α_j on excess returns come from two channels: (a) *market price channel*. As all investors produce more precise information about asset j, its market price also incorporates more information. As investors with fewer financial experts produce less precise information than other investors, holding the number of computer scientists the same, they put more weight on information in the market price. Therefore, investors with fewer financial experts benefit more from a more informative market price. The impact of a higher α_j is the largest for them through this channel. (b) *direct channel*. The direct effect on the precision of private signal depends on investors' labor composition and the sign of $\frac{\partial^2 K}{\partial D \partial n_f}$, i.e., the relation between the two inputs. If the two inputs are substitutes, an increase in the amount of machine-readable data decreases the marginal productivity of financial experts. As a result, the signal precision of investors with more financial experts decreases relative to investors with fewer financial experts. If instead they are complements, more machine-readable data benefit investors with more financial experts even more, resulting larger informational advantage for investors with

more financial experts.

Note that the two channels may have similar or opposite impacts on investors' informational advantage, depending on the relation of the two inputs. In the case of substitution, both effects reduce the informational advantage of investors with more financial experts over investors with fewer. In the case of complementarity, the market price channel works against the direct effect. Overall, complementarity is a necessary condition for larger informational difference between investors with more and fewer financial experts.

The exogenous shock on α_j also changes the unconditional holding of investors. For investors with (n_p, n_f), their average unconditional holding on asset j is

$$E[q_j|(n_p, n_f)] = \hat{\sigma}_j(n_p, n_f)^{-1}\bar{\sigma}_j\bar{x}_j \tag{2}$$

where $\hat{\sigma}_j(n_p, n_f)^{-1} = (\Sigma^{-1})_{jj} + (\Sigma_p^{-1})_{jj} + K(\alpha_j n_p, n_f)$. The market clearing condition implies that the unconditional ownership of stock j by a type (n_p, n_f) investor is given by

$$\begin{aligned} E[Ownership|(n_p, n_f)] =& \frac{\mu(n_p, n_f)E[q_j|(n_p, n_f)]}{\bar{x}_j} \cdot \frac{1}{\mu(n_p, n_f)} \\ =& \hat{\sigma}_j(n_p, n_f)^{-1}\bar{\sigma}_j \end{aligned}$$

Similar to Proposition 1, the following result on stock ownership is immediate.

Proposition 2. *Under Assumption 1, an exogenous increase in α_j does not affect unconditional ownership of asset j' for all $j' \neq j$. For asset j,*

(i) If $\frac{\partial^2 K}{\partial D \partial n_f} < 0$, investors with more financial experts reduce their ownership of asset j relative to investors with fewer financial experts after the shock.
(ii) If investors with more financial experts increase their ownership of asset j relative to investors with fewer financial experts after the shock only if $\frac{\partial^2 K}{\partial D \partial n_f} > 0$.

The intuition behind this result is that higher signal precision decreases ex ante uncertainty and hence affects the unconditional ownership. The mechanism is similar to its effect on excess returns.

For holding dispersion, I use the standard deviation of excess holding for each type of investors. Specifically, for type (n_p, n_f) investors, their holding dispersion on asset j,

$\delta_j(n_p, n_f)$, is defined as

$$\begin{aligned}\delta_j(n_p, n_f) \equiv & \sqrt{Var[q_{ij} - \bar{q}_j | i \in (n_p, n_f)]} \\ = & \frac{1}{\rho}\sqrt{v_j(K(D(c_j, n_p), n_f) - \bar{K}_j)^2 + K(D(c_j, n_p), n_f)}\end{aligned} \tag{3}$$

$Var[\cdot]$ is the variance operator. $i \in (n_p, n_f)$ means that the variance is conditional on investor i with type (n_p, n_f).

Proposition 3. *After an increase in α_j, $\delta_{j'}$ doesn't change for all $j' \neq j$. Moreover, if σ_x and $K(D(c_j, \overline{n}_p), \overline{n}_f)$ are large enough, if n_p and n_f are substitutes, $\delta_j(\underline{n}_p, \overline{n}_f)$ decreases less than $\delta_j(\underline{n}_p, \underline{n}_f)$. If instead, $\delta_j(\underline{n}_p, \overline{n}_f)$ decreases no less than $\delta_j(\underline{n}_p, \underline{n}_f)$, then n_p and n_f are complements.*

Holding dispersion depends both on the distance to the average precision and investors' own precision. Note that as investors are collectively more informed on asset j, v_j decreases. When the supply of the asset is noisy enough (σ_x is large enough), the first term in the square root dominates and $\delta_j(\underline{n}_p, \underline{n}_f)$ decreases. Conversely, a smaller δ_j confirms that the first term dominates. For the IT-intensive type, their information improves less than the base type, and so does the change in distance to the average precision. As a result, $\delta_1(\overline{n}_p, \underline{n}_f)$ decreases less than $\delta_1(\underline{n}_p, \underline{n}_f)$. For the finance-intensive type, whether the change in distance is larger or smaller than the base type depends on the relation between n_p and n_f.

2.1 Empirical predictions

The previous results suggest that we can infer the relation between the two inputs by exploiting a shock that decreases the cost of data processing on some stocks but not on others, using the following regression,

$$\begin{aligned}y_{ijt} = & \gamma_\theta Type_{i,\theta} \times Treated_j \times Post_t + \beta\, Treated_j \times Post_t + Type_{i,\theta} \times Treated_j \\ & + Type_{i,\theta} \times Post_t + Treated_j + Type_{i,\theta} + Post_t + Controls\end{aligned} \tag{4}$$

$Type_\theta$ is an indicator for either IT-, finance- or bi-intensive type. $Treated_j$ is an indicator for the treated stocks, $Post_t$ indicates whether the treatment has been given. γ_θ captures the differential impact between $Type_\theta$ and the base type investors. β captures the impact of the shock on treated stocks relative to non-treated stocks for the base type investors. When

running this regression with excess returns as the dependent variable, it yields γ_θ^r and β^r. Proposition 1 implies the following tests.

Result 1. *In the case of substitution, $\gamma^r_{finance} < 0$: the performance gap between base type investors and finance-intensive investors becomes smaller.*

Result 2. *$\gamma^r_{finance} \geq 0$ implies complementarity, i.e., if the performance gap between base type investors and finance-intensive investors becomes larger, then machine-readable data complement financial experts.*

Results 1 and 2 also hold when using stock ownership as the dependent variable, in which case regression (4) yields the estimates γ_θ^s and β^s.

Running regression (4) using holding dispersion as the dependent variable gives γ_θ^δ. Proposition 3 provides the following predictions.

Result 3. *$\gamma^\delta_{finance} \leq 0$ implies complementarity. In the case of substitution, $\gamma^\delta_{finance} > 0$. If holding dispersion decreases more for finance-intensive investors than for the base type investors, the two inputs are complements. If the two inputs are substitutes, holding dispersion decreases less for finance-intensive investors.*

Result 3 provides another way to determine how machine-readable data interact with financial experts.

3 Empirical Setup

3.1 The IT-Augmenting Shock and Sample Construction

I use the implementation of the SEC's XBRL mandate between 2009 and 2011 as an IT-augmenting shock. XBRL (eXtensible Business Reporting Language) is a programming language that facilitates communication of large volumes of business information using standard taxonomies and tagging. When preparing a financial statement in XBRL format, companies identify and tag each element in the statement with the standard taxonomy developed by the SEC. The tags are linked to their descriptive information, such as name, year, units, detailed definition and also relationship with other items. These features allow users of such reports to quickly search and locate the item and related information they are interested in. Cross-company or cross-time comparison is also much simplified due to the

standard taxonomies. Unlike XBRL, doing such analysis is difficult and costly with the static files (HTML or plain text) in the EDGAR system. Although HTML files are also organized using tags, those tags are mostly location-based, have little relationship to the content, and can vary across files. To get a sense of the difficulty, think about how to extract numbers from a footnote. Such numbers are usually scattered in text. If an analysis requires such information from various companies, users of plain text or HTML files must either deploy a number of human analysts to search for the numbers manually, or develop sophisticated textual analysis programs that are based on location and context to extract information. Both methods require significant efforts and costs, and are prone to error.

Due to a special data structure, manipulation of XBRL files requires some programming skills. For this reason, the mandate simplifies the task and increases the productivity of computer scientists more than financial experts. The mandate makes data extraction easier for all investors rather than just benefit top computer scientists. This feature makes the mandate more likely to reduce rather than to increase the information gap between institutional investors with different level of IT resources.

The identification strategy of this paper exploits the staggered implementation of the XBRL mandate: firms with public float larger than 700 million USD are required to comply from June 15, 2009; firms with public float between 50 million USD and 700 million USD must report in the XBRL format from June 15, 2010. For the rest, the mandate came into effect in June 2011. In principle, firms may voluntarily choose to disclose in the XBRL format before they are obliged to. As long as the early compliance is not meant to benefit a special type of investors, it does not pose a threat to the identification of this paper.

In each quarter, a stock's XBRL status is inferred from its file format in the SEC's EDGAR system. It is labeled as an XBRL stock (treated) if it has filed 10-K, 10-Q or 8-K in the XBRL format in that quarter. Figure 2 plots the time series of the number of XBRL stocks. As shown in the graph, there are three major dates when firms comply with the XBRL mandate: 2009 Q3, 2010 Q3 and 2011 Q3. In the analysis, I focus only on two quarters before and after each event (one cohort). I then stack observations from all cohorts together and align them along the period relative to the XBRL event. This construction avoids using the same observation as both treated and control. The sample period is thus from 2009 Q1 to 2011 Q4.

[Insert Figure 2]

3.2 Labor Inputs

Since data on institutional investors' entire labor force is not available, I use their foreign skilled labor as a proxy. Information on foreign skilled labor is obtained using the Labor Condition Application (LCA) data from the U.S. Department of Labor. This dataset contains each employee's job title, brief job description, and proposed contract duration. Such information allows me to construct a panel of institutions' desired number of foreign high skilled workers in both IT and finance related positions at each point in time. The LCA is a prerequisite for the H-1B visa application. The H-1B visa is a temporary program that permits foreign skilled -workers in specialty -occupations to work in the U.S. These occupations require theoretical and practical application of highly specialized knowledge like engineering or accounting and attainment of a bachelor's or higher degree in the specific specialty (or its equivalent). An H-1B visa permits the holder to work in the U.S. for three years and can be renewed for a maximum of six years. The application of the H-1B visa has to be sponsored by an employer.

I classify jobs into IT- and finance-related positions based on job codes and job titles. For observations after mid-2009, I only use the Standard Occupational Classification (SOC) from the Department of Labor. IT jobs are positions with the SOC code starts with 15-11 (Computer Occupations) or 11-3021 (Computer and Information Systems Managers). Finance jobs are positions with the SOC code starts with 13-20 (Financial Specialists) or 11-3031 (Financial Managers). For observations before mid-2009 the job code classification is based on the Occupational Title (OT) codes from U.S. Citizenship and Immigration Services. For IT-related jobs, I include jobs with the OT code starts with 03 (Computer-Related Occupations), and 199 (Miscellaneous Professional, Technical, and Managerial Occupations) if the job title mentions one of the following words: developer, software, system, program, and information. For finance-related jobs, I include jobs for which the OT code starts with 50 (Occupations in Economics), 186 (Finance, Insurance, and Real Estate Managers and Officials), and 199 (Miscellaneous Professional, Technical, and Managerial Occupations) if the job tile mentions one of the following words: analyst, research, financial and investment. Using information on contract duration and assuming no separation, I calculate the total number of both types of jobs for each institution in each quarter. As these numbers proxy cumulative hiring over the past three years (the maximum and the most frequent contract duration), they are more

likely to reflect current labor composition than only just new recruitment.

An underlying assumption of this measure is that institutional investors on average do not particularly prefer native skilled workers for either IT or finance positions. With this assumption, the distribution of foreign skilled workers can proxy for the distribution of skilled workers in the entire workforce. If it is violated, for example, if native financial experts are preferred, then many finance-intensive institutions would be classified as base type. This goes against finding a significant difference between base type institutions and the other types. There may be another concern that the measure based on the LCA data can be problematic since the visa may not be granted. Most foreign graduates benefit from the Optional Practical Training (OPT) program, which allows them to work for at least one year without holding other visas. It is a common practice that firms hire new employees relying on the OPT program and apply for the H-1B visa in advance before the OPT program expires. Even if the visa is not granted, the firm has time to search for another employee to refill the position. Therefore the LCA data likely reflects the firm's desired number of positions.

To classify institutions into the four types, I merge the previous panel with institutional investor data from Thomson Reuters 13-F dataset, using the names of the institutions. I keep only institutions with at least one H-1B visa application between 2007 and 2013. I exclude banks and insurance companies (Thomson Reuters type code 1 and 2). For each cohort, I then sort institutions based on their IT and finance intensity, which is the ratio between their number of IT or finance positions and their assets under management (AUM), both measured at one quarter before the event. If an institution falls in the top tercile of IT intensity distribution and in the bottom tercile of finance intensity distribution, I classify it as IT-intensive type. The opposite is classified as finance-intensive. A base type institution falls in the bottom tercile in both dimensions while a bi-intensive type assumes top terciles. The summary statistics of jobs in an institution is given in the first two rows of Table 1. The summary statistics of investor size by their type is described in row 3 until row 6. In the sample, there are about 420 institutions. IT-intensive and base type institutions are relatively larger than finance-intensive and bi-intensive institutions.

[Insert Table 1]

One way to verify that the classification is correct is to check whether investors' type roughly corresponds to their investment philosophy. An institution can probably rely more

on quantitative methods if it has many computer scientists and more on discretionary or fundamental methods if its team is mainly composed of financial analysts. Table 2 reports the twenty largest institutions for IT-intensive, finance-intensive, and bi-intensive types. Consistent with this intuition, the measure classifies institutions which are well-known for their quantitative approach, such as D. E. Shaw & Co., Renaissance Technologies, or Two Sigma Investments, as IT-intensive institutions. Institutions labeled as finance-intensive seem to rely more on fundamental analysis. For example, Tremblant Capital states on its website that its managers "conduct deep fundamental research to uncover investments that are trading at a material dislocation from fair value." Sandler Capital Management believes that "in-depth fundamental research and deep industry knowledge are the primary contributing factors to successful investing." Sirios Capital Management identifies itself as "a fundamentally-driven investment firm ... and its investment process is driven by fundamental research on a company-by-company basis." Many large asset management firms, such as Merrill Lynch, UBS Securities and Bridgewater Associates are classified as the bi-intensive type. These facts lend us confidence that the measure can capture an institution's advantage in both dimensions.

Table 3 provides further comparison on the different types of institutions. In Table 3, I compare each type of institutions to the base type using the following regression:

$$y_{it} = \alpha + \sum_{\theta} \beta_{\theta} Type_{i\theta t} + \alpha_t + Controls_{it}$$

The dependent variable in column (1), (2), and (3) are institutional turnover, average market capitalization of portfolio stocks (weighted by portfolio weight), and log number of portfolio stocks, respectively. The result in column (1) shows that comparing to base-type institutions, IT-intensive institutions have higher turnover whereas finance-intensive institutions have lower turnover. This is also consistent with the intuition that IT-intensive institutions may exploit high frequency information, e.g. order flow, more often and have a shorter investment horizon. By contrast, finance-intensive institutions may rely more on fundamental approaches and have a longer investment horizon. Column (2) shows that on average finance-intensive institutions hold stocks with a smaller market capitalization. Column (3) suggests that these institutions hold similar number of stocks once controlled for their size.

3.3 Other Variables

I report how I construct other variables such as excess returns, stock ownership, dispersion in holding, and control variables.

Following Kacperczyk et al. (2016), I compute the excess return of investor i on stock j at time t as the product of its excess holding, $w_{ijt} - \bar{w}_{ijt}$, and the return on the stock, r_{jt},

$$R_{ijt}^{holding} = (w_{ijt} - \bar{w}_{jt})r_{jt}$$

where w_{ijt} is the weight of stock j in the portfolio of investor i in quarter t. $\bar{w}_{jt}$ is the average of w_{ijt} across all investors. r_{jt} is stock j's cumulative return in quarter t. Stock return data is from the CRSP. Even though each investor should hold almost every stock to gain from diversification, few of them do in reality, either due to fixed costs or information capacity constraint. In the analysis, I only consider non-zero weights. The returns are measured in basis points.

I also measure excess returns using institutions' trading behavior as follows,

$$R_{ijt}^{Trading} = [(w_{ij,t} - \bar{w}_{ij,t}) - (w_{ij,t-1} - \bar{w}_{ij,t-1})]r_{jt}$$

Here, $(w_{ij,t} - \bar{w}_{ij,t}) - (w_{ij,t-1} - \bar{w}_{ij,t-1})$, the change in investor i's excess holding on stock j between $t-1$ and t, proxies her trading. To avoid mechanical change due to movement in sample averages or trading in other assets, I only consider observations with a change in the number of shares held.

I construct two measures of stock ownership. The first one is the fraction of total shares outstanding owned by a type of institution. Formally,

$$Fraction_{j\theta t} = \frac{\sum_i (Shares_{ijt} \cdot Type_{i,\theta})}{Total\ Shares_{jt}}$$

$Share_{ijt}$ is the number of shares of stock j held by institution i in quarter t. $Type_{i,\theta}$ indicates whether institution i is of type θ. Because types with larger institutions mechanically have higher ownership fractions, their changes may also be mechanically larger. To address this concern, I consider a second measure, which is the fraction scaled by the total assets of that

type.

$$FracScaled_{j\theta t} = \frac{Fraction_{j\theta t}}{\sum_i (Assets_{it} \cdot Type_{i,\theta})} \times 10^6$$

The measure is multiplied by the constant 10^6 simply to avoid too many leading zeros in the estimates.

The holding dispersion across all investors δ_{jt} is the standard deviation of excess return, $w_{ijt} - \bar{w}_{ijt}$, for each stock j in each quarter t. $\delta_{j\theta t}$ is calculated similarly conditional on the type of investors θ. The dispersion is measured in percentage points.

To control for an institution's other characteristics, I include their turnover and assets under management. Following Ben-David et al. (2010), an institution's turnover is the ratio between its total trading value for a given quarter and its assets under management.

4 Empirical Results

4.1 Substitutes or Complements ?

I run the main tests based on excess returns, ownership and dispersion of excess holdings. For excess returns,

$$\begin{aligned} y = & \sum_\theta \gamma^r_\theta \, Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^r XBRL_{jc} \times Post_q + Controls_{ijqc} \\ & + \text{Stock-Type-Cohort FE} + \text{Type-Period-Cohort FE} \ (+\text{Institution FE}) \end{aligned} \tag{5}$$

The dependent variable y is R_{ijqc}, $Fraction_{j\theta qc}$ or $\delta_{j\theta qc}$. Period q is measured as the number of quarters relative to the event time in each cohort. Indicating the period q and cohort c together is equivalent to indicating the calendar quarter t. $Type_{i\theta c}$ is the type indicator for investor i in cohort c. The sum is over all the types except the base type. $XBRL_{jc}$ is equal to one if stock j complies to the XBRL mandate in cohort c and zero otherwise. $XBRL_{jc}$, $Type_{i\theta c}$ and $Type_{i\theta c} \times XBRL_{jc}$ are absorbed by Stock-Type-Cohort fixed effects. $Post_q$ and $Type_{i\theta c} \times Post_q$ are absorbed by Type-Period-Cohort fixed effects. I also report results that control for Stock-Period-Cohort fixed effects, which absorb $XBRL_{jc} \times Post_q$. In the regression on the excess returns, I control for institution fixed effects. Control variables such as assets under management, institution turnover and stock market values are included. In the regressions

on stock ownership and holding dispersion, I control for stock market capitalization, shares owned by institutional investors, book-to-market ratio and leverage ratio. I also control for the average asset under management of the institutions that owns the stock for each type. γ_θ measures the difference between type θ investors and the base type investors on the XBRL stocks, before and after the shock, relative to the unaffected stocks.

4.1.1 Evidence on Excess Returns

The results of regression (5) are reported in Table 4. Columns (1) and (2) report the results using excess holding returns as the dependent variable. In column (1), the excess returns on the treated stocks increase more for finance-intensive institutions. Comparing to control stocks, the annualized excess return of finance-intensive institutions on one treated stock is about 3 basis points higher than that of base-type institutions. This result is consistent with Result 1, in favor of complementarity, and rejects Result 2 or substitution. The point estimate for Bi-intensive institutions is higher for IT-intensive institutions, also consistent with complementarity instead of substitution, even though the difference is not statistically significant.

The performance gap between the base type institutions and IT-intensive institutions on the treated stocks decreases by about 0.6 basis point. For an IT-intensive institution that holds the average number of stocks (148), this decrease implies that the annualized performance gap decreases by 0.3 percentage point if all stocks are treated. This negative coefficient implies that IT-intensive institutions lose part of their informational advantage on the treated stocks after the shock. One possible explanation for this result is that IT-intensive institutions were able to process more data on the stocks since they have more computer scientists, which gives them an edge over institutions with fewer computer scientists. The mandate decreased the cost of data processing so that their advantage coming from more machine-readable data is diminished.

In column (2), I control for stock-period-cohort fixed effects, which absorb stock fixed effects and $XBRL_{jc} \times Post_q$. The results are qualitatively similar.

The XBRL mandate may also help investors make more informed trades. Columns (3) and (4) of Table 4 report the results with trading returns. The results are similar to excess holding returns and suggest complementarity.

[Insert Table 4]

[Insert Figure 3]

Figure 3 provides information on the time evolution of the performance gap between the IT-intensive or finance-intensive institutions and the base type institutions on the treated stocks, relative to the performance gap on the control stocks. It plots the estimates of $Type \times XBRL \times Dummy_q$. These estimates capture the differential impact between the IT-intensive (or finance-intensive) type and the base type investors on the treated stocks compared to the non-treated stocks. We expect the performance gap to be smaller (larger) for the IT-intensive (finance-intensive) type after the event. Before the implementation, there is no significant difference, suggesting that the parallel trend assumption is not violated. The two performance gaps diverge after the treatment. Figure 4 plots the similar estimates for trading returns.

[Insert Figure 4]

4.1.2 Evidence on Stock Ownership

Table 5 reports the results on stock ownership.

[Insert Table 5]

The dependent variable in the first two columns is the fraction of total share outstanding owned by each type of investors. In column (1), compared to the base type investors, the ownership on the treated stocks by IT-intensive and bi-intensive investors decreases by about 0.007 and 0.002 percentage point, or 23% and 6% of the sample average, respectively, after the implementation of the XBRL mandate. The coefficient for the base type is not statistically significant. In column (2), the result is similar even after controlling for stock-period-cohort fixed effects. In both columns, the coefficient on the finance-intensive type is negative though insignificant. One potential explanation for this finding is that finance-intensive institutions have less assets than base type institutions in the sample. Even if they overweight the treated stocks more than the base type institutions, their ownership of treated stocks might decrease. To mitigate the effect of institution size, in column (3) and column (4), I divide the fraction of ownership by the total assets of that type. In column (3), scaled ownership

by base type investors increases on the treated stocks comparing to the control group after the implementation of XBRL. In column (3) and column (4), scaled ownership by finance-intensive investors increased further than the base type investors. For IT-intensive and bi-intensive institutions, the results are similar to columns (1) and (2).

Figure 5 plots the time evolution of the differences in scaled ownership between the IT-intensive or finance-intensive institutions and the base type institutions on the treated stocks, relative to the control stocks.

[Insert Figure 5]

4.1.3 Evidence on Stock Holding Dispersion

The results on dispersion of excess holdings are shown in Table 6. In column (1), the negative coefficient before $XBRL \times Post$, even though not significant, shows that the XBRL mandate decreases dispersion on the treated stocks among the base type institutions, consistent with the XBRL mandate being an IT-augmenting shock. In addition, the coefficient before the interaction with the IT-intensive (finance-intensive) indicator is positive (negative). These coefficients are not statistically significant. One reason may be that I control for stock-type-cohort fixed effects. These fixed effects make sure that the comparison is among the same stock-type pair in each cohort but it also restricts the comparison to at most four observations: two before and two after the event. In column (3) and (4), I control for treated-type-cohort fixed effects, which is still reasonably conservative. After the implementation of the XBRL mandate, the dispersion on the treated stocks among the base type institutions decreased by 0.1 percentage point (p-value <0.1), or 20% of the sample mean. For finance-intensive institutions, the dispersion decreased even more by 0.15 percentage point (p-value<0.05). The dispersion among IT-intensive type and bi-intensive type institutions barely decreased.

[Insert Table 6]

Overall, the evidence on excess returns, stock ownership, and dispersion of excess holdings are more consistent with computer scientists and financial experts being complementary. I now provide additional results using a matched sample.

4.2 Robustness Checks

One concern for the results is that the IT-intensive type investors are much larger than the base type investors while the finance-intensive type investors are smaller. The results may be driven by investor characteristics related to their size. It is not surprising that IT intensity is correlated with size. Given fixed set-up costs of an IT system, larger institutions find it more economical to have more IT resources than small institutions. However, the comparison between the bi-intensive type and the base type suggests that the size effect is not likely to explain the results. In the sample, the by-intensive type investors are also smaller than the base type investors. Yet, the estimation result is very different from the results from the finance intensive type and the base type.

To make the base type and other types more comparable in terms of size and investment horizon, I match the base type with other types using the Coarsened Exact Matching method in each cohort, based on their assets under management and turnover, both measured at one quarter before an event. Table 7 gives the summary statistics of investor size by their type in the matched sample. Base$_i t$, Base$_f in$ and Base$_b i$ refers to the base type investors matched with IT-intensive, finance-intensive and bi-intensive type investors respectively.

I run similar regressions as in (5) on the three types, IT-intensive, finance-intensive and bi-intensive separately, relative to the base type. The results are reported in Table 8 and Table 9. The results are also consistent with computer scientists and financial advisors being complements.

[Insert Table 8]

[Insert Table 9]

5 Conclusion

The rapid progress of information technologies brings deep changes to the financial industry. The ability of computers to process massive amounts of numbers, images, and natural languages demonstrates their potential not only for enhancing productivity but also replacing human labor. Using information on highly skilled labor in the finance industry, this paper suggests that, at least until recently, financial institutions exploited modern technologies mostly to assist their financial experts.

Understanding this complementarity helps job seekers or employers to make better career or recruitment decisions, and also helps regulators and policymakers to evaluate the impact of new technologies on the finance industry. Whether a similar complementarity can be found in other areas of the financial industry and how the relation might be changed by new technologies remains an open question. Using more recent changes or technology shocks, the methodology used in this paper can also shed light on these issues. In addition, regulators can apply this method to identify winners and losers from other regulations or technologies.

References

Abis, S. (2017). Man vs. machine: Quantitative and discretionary equity management. *Unpublished working paper. Columbia Business School*.

Abis, S. and Veldkamp, L. (2020). The changing economics of knowledge production. *Available at SSRN 3570130*.

Acemoglu, D. and Restrepo, P. (2018). Artificial intelligence, automation and work. Technical report, National Bureau of Economic Research.

Admati, A. R. (1985). A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica: Journal of the Econometric Society*, pages 629–657.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The skill complementarity of broadband internet. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Aubry, M., Kräussl, R., Manso, G., and Spaenjers, C. (2019). Machine learning, human experts, and the valuation of real assets. *HEC Paris Research Paper No. FIN-2019-1332*.

Bai, J., Philippon, T., and Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654.

Ben-David, I., Glushkov, D., and Moussawi, R. (2010). Do arbitrageurs really avoid high idiosyncratic risk stocks? *Available at SSRN 1572955*.

Bhattacharya, N., Cho, Y. J., and Kim, J. B. (2018). Leveling the playing field between large and small institutions: evidence from the sec's xbrl mandate. *The Accounting Review*, 93(5):51–71.

Blankespoor, E. (2019). The impact of information processing costs on firm disclosure choice: Evidence from the xbrl mandate. *Journal of Accounting Research*, 57(4):919–967.

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? In *AEA Papers and Proceedings*, volume 108, pages 43–47.

Chang, Y.-C., Hsiao, P.-J., Ljungqvist, A., and Tseng, K. (2020). Testing disagreement models.

Coleman, B., Merkley, K. J., and Pacelli, J. (2020). Man versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *Available at SSRN 3514879*.

David, H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30.

Dimmock, S. G., Huang, J., and Weisbenner, S. J. (2019). Give me your tired, your poor, your high-skilled labor: H-1b lottery outcomes and entrepreneurial success. Technical report, National Bureau of Economic Research.

Dong, Y., Li, O. Z., Lin, Y., and Ni, C. (2014). Does information processing cost affect firm-specific information acquisition?-evidence from xbrl adoption. *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming*.

Dugast, J. and Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial Economics*, 130(2):367–391.

Efendi, J., Park, J. D., and Subramaniam, C. (2016). Does the xbrl reporting format provide incremental information value? a study using xbrl disclosures during the voluntary filing program. *Abacus*, 52(2):259–285.

Erel, I., Stern, L. H., Tan, C., and Weisbach, M. S. (2018). Selecting directors using machine learning. Technical report, National Bureau of Economic Research.

Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2020). Where has all the data gone? Working Paper 26927, National Bureau of Economic Research.

Farboodi, M. and Veldkamp, L. (2020). Long-run growth of financial data technology. *American Economic Review*, 110(8):2485–2523.

Gao, M. and Huang, J. (2020). Informing the market: The effect of modern information technologies on information production. *The Review of Financial Studies*, 33(4):1367–1411.

Goldstein, I., Yang, S., and Zuo, L. (2020). The real effects of modern information technologies. Technical report, National Bureau of Economic Research.

Grennan, J. and Michaely, R. (2019). Fintechs and the market for financial analysis. *Michael J. Brennan Irish Finance Working Paper Series Research Paper*, (18-11):19–10.

Grennan, J. and Michaely, R. (2020). Artificial intelligence and high-skilled work: Evidence from analysts. *Available at SSRN*.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, 70(3):393–408.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

Kerr, S. P., Kerr, W. R., and Lincoln, W. F. (2015). Skilled immigration and the employment structures of us firms. *Journal of Labor Economics*, 33(S1):S147–S186.

Kerr, W. R. and Lincoln, W. F. (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics*, 28(3):473–508.

Kim, J.-B., Li, B., and Liu, Z. (2019). Information-processing costs and breadth of ownership. *Contemporary Accounting Research*, 36(4):2408–2436.

Kim, J. W., Lim, J.-H., and No, W. G. (2012). The effect of first wave mandatory xbrl reporting across the financial information environment. *Journal of Information Systems*, 26(1):127–153.

Liu, C., Wang, T., and Yao, L. J. (2014). Xbrl's impact on analyst forecast behavior: An empirical study. *Journal of accounting and public policy*, 33(1):69–82.

Peri, G., Shih, K., and Sparber, C. (2015). Foreign and native skilled workers: What can we learn from h-1b lotteries? Technical report, National Bureau of Economic Research.

Sharifkhani, A. (2018). Immigration policy and equity returns: Evidence from the h-1b visa program. In *31st Australasian Finance and Banking Conference*.

van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2020). Man vs. machine learning: The term structure of earnings expectations and conditional biases. Technical report, National Bureau of Economic Research.

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

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

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica: Journal of the Econometric Society*, pages 1415–1430.

Webb, M. (2019). The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*.

Appendix

Proof of Poposition 1

Proof. (a)Derivation of excess return

The derivation of $E[R]$ follows Kacperczyk et al. (2016) closely.

Conjecture that that the prices provide an unbiased signal about f, $\eta_p = f + \epsilon_p$, where $\epsilon_p \sim N(0, \Sigma_p)$. In a linear equilibrium where p is a linear function of investors' signals and asset supply,

$$p = \frac{1}{r}(A + Bf + Cx)$$

with A, B and C to be determined.

Based on private signals and prices, investors update their belief,

$$\hat{\mu}_i = E[f|\eta_{si}, p] = \hat{\Sigma}_i(\Sigma^{-1}\mu + \Sigma_p^{-1}\eta_p + K_i\eta_{si})$$
$$\hat{\Sigma}_i = var[f|\eta_{si}, p] = (\Sigma^{-1} + \Sigma_p^{-1} + K_i)^{-1}$$

First order conditions then give their holding conditional on signal realizations:

$$q_i = \frac{1}{\rho}\hat{\Sigma}_i^{-1}(\hat{\mu}_i - pr)$$

Market clearing gives

$$\begin{aligned} A =& \bar{\Sigma}(\Sigma^{-1}\mu - \rho\bar{x}) \\ B =& I - \bar{\Sigma}\Sigma^{-1} \\ C =& -\rho\bar{\Sigma}(I + \frac{1}{\sigma_x^2\rho^2}\bar{K}) \\ \Sigma_p^{-1} =& (\sigma_x^2 B^{-1}CC'B^{-1'})^{-1} = \frac{1}{\sigma_x^2\rho^2}\bar{K}\bar{K}^T, \end{aligned}$$

$\bar{K} = \int K_i di$ and $\bar{\Sigma}^{-1} = \int \hat{\Sigma}_i^{-1} di = \Sigma^{-1} + \Sigma_p^{-1} + \bar{K}$.

Using the expression of p, we can write $f - pr$ as

$$\begin{aligned} f - pr =& (I - B)f - Cx - A \\ =& \bar{\Sigma}[\Sigma^{-1}z + \rho(I + \frac{1}{\sigma_x^2\rho^2}\bar{K}^{-1})x] + \rho\bar{\Sigma}\bar{x} \\ =& V^{\frac{1}{2}}u + w \end{aligned}$$

where $V = \bar{\Sigma}[\rho^2\sigma_x^2 I + \bar{K} + \bar{\Sigma}^{-1}]\bar{\Sigma}$, $u \sim N(0,1)$, and $w = \rho\bar{\Sigma}\bar{x}$.

The market average holding is then given by

$$\begin{aligned} \bar{q} =& \frac{1}{\rho}\int \hat{\Sigma}_i^{-1}(\hat{\mu}_i - pr)di \\ =& \frac{1}{\rho}[\bar{K}z + \Sigma_p(z + \epsilon_p) + \bar{\Sigma}^{-1}(\mu - pr)] \end{aligned}$$

Investor i's excess holding

$$\begin{aligned} q_i - \bar{q} =& \frac{1}{\rho}[K_i\epsilon_{si} + (\hat{\Sigma}_i^{-1} - \bar{\Sigma}^{-1})(u + z - pr)] \\ =& \frac{1}{\rho}[K_i\epsilon_{si} + \Delta_i(V^{\frac{1}{2}}u + w)] \end{aligned}$$

where $\Delta_i = \bar{\Sigma}_i^{-1} - \bar{\Sigma}^{-1} = K_i - \bar{K}$. The last equality follows from $u + z - pr = f - pr$ and the expression of $f - pr$.

It is straightforward to check that all the matrices above are diagonal. With the expression of $q_i - \bar{q}$ and $f - pr$, investor's excess return can be easily calculated.

$$E_i[R] = E[(q_i - \bar{q})^T(f - pr)] = \rho(\bar{x}^T\bar{\Sigma}\Delta_i\bar{\Sigma}\bar{x}) + \frac{1}{\rho}Tr(\Delta_i V)$$

where $Tr(\cdot)$ is the trace operator.

Since all matrices are diagonal, the excess return of asset j for an investor with (n_p, n_f) is given by

$$E[R_j(n_p, n_f)] = \rho\bar{x}_j^2\bar{\sigma}_j^2(K_j(n_p, n_f) - \bar{K}_j) + \frac{1}{\rho}(K_j(n_p, n_f) - \bar{K}_j)v_j$$

where $\bar{x}_j$ is the jth element of $\bar{x}$. $\bar{\sigma}_j = (\bar{\Sigma})_{jj}$ and $v_j = (V)_{jj}$ are the jth diagonal element of $\bar{\Sigma}$ and V respectively.

(b) Results related to γ_2

The difference of the excess return on asset j relative to the base type investors is

$$\begin{aligned}\Delta R_j(n_p, n_f) =& E[R_j(n_p, n_f)] - E[R_j(\underline{n}_p, \underline{n}_f)] \\ =& \rho \bar{x}_j^2 \bar{\sigma}_j^2 (K_j(n_p, n_f) - K_j(\underline{n}_p, \underline{n}_f)) + \frac{1}{\rho}(K_j(n_p, n_f) - K_j(\underline{n}_p, \underline{n}_f))v_j\end{aligned}$$

The change of the difference after the shock is

$$\gamma_j(n_p, n_f) = \Delta R_j'(n_p, n_f) - \Delta R_j(n_p, n_f)$$

Since the shock is only on asset 1, $K_2(n_p, n_f)$ is not affected. So are $\bar{\sigma}_2$ and v_2. Thus ΔR_2 does not change and $\gamma_2(n_p, n_f) = 0$.

(c) Results related to $\gamma_1(\overline{n}_p, \underline{n}_f)$

For asset 1,

$$\gamma_1(n_p, n_f) = \Delta R_1(\alpha n_p, \alpha n_f) - \Delta R_1(n_p, n_f)$$

$\alpha > 1$. To determine the sign of $\gamma_1(\overline{n}_p, \underline{n}_f)$, it is useful to consider

$$\begin{aligned}\frac{\partial(K(\alpha\overline{n}_p, \underline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} =& \frac{\partial}{\partial\alpha}\int_{\alpha\underline{n}_p}^{\alpha\overline{n}_p} \frac{\partial K(n_p, \underline{n}_f)}{\partial n_p} dn_p \\ =& \overline{n}_p \frac{\partial K(\alpha\overline{n}_p, \underline{n}_f)}{\partial n_p} - \underline{n}_p \frac{\partial K(\alpha\underline{n}_p, \underline{n}_f)}{\partial n_p}\end{aligned}$$

Since $\alpha > 1$, it has the same sign as

$$\begin{aligned}\alpha\frac{\partial(K(\alpha\overline{n}_p, \underline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} =& \alpha\overline{n}_p \frac{\partial K(\alpha\overline{n}_p, \underline{n}_f)}{\partial n_p} - \alpha\underline{n}_p \frac{\partial K(\alpha\underline{n}_p, \underline{n}_f)}{\partial n_p} \\ & < 0\end{aligned}$$

The last inequality follows from $\frac{\partial}{\partial n_p}(n_p \frac{\partial K}{\partial n_p}) < 0$ (Assumption 1) and $\overline{n}_p > \underline{n}_p$.

After the shock, $\bar{K}_1$ increases. So does the precision of price signal σ_{p1}^{-1}. Since $\bar{\sigma}_1^{-1} =$

$\sigma_1^{-1} + \bar{K}_1 + \sigma_{p1}^{-1}$, $\bar{\sigma}_1$ decreases. It remains to check how v_1 changes. We only need to consider

$$\begin{aligned}
\frac{\partial v_j}{\partial \bar{K}_j} =& \frac{\partial}{\partial \bar{K}}\left(\frac{1}{\sigma^{-1}+\sigma_p^{-1}+\bar{K}} + \frac{\rho^2\sigma_x^2+\bar{K}}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2}\right) \\
=& -\frac{1}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2} + \frac{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^2 - 2(\rho^2\sigma_x^2+\bar{K})(\sigma^{-1}+\sigma_p^{-1}+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^4} \\
=& -\frac{2(\rho^2\sigma_x^2+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^3} \\
<& 0
\end{aligned}$$

The subscript j is omitted on the RHS.

Combining the results above, it is immediate that $\gamma_1(\overline{n}_p, \underline{n}f) < 0$.

(d) Results related to $\gamma_1(\underline{n}_p, \overline{n}_f)$

Consider

$$\begin{aligned}
\frac{\partial(K(\alpha\underline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} =& \frac{\partial}{\partial\alpha}\int_{\underline{n}_f}^{\overline{n}_f} \frac{\partial K(\alpha\underline{n}_p, n_f)}{\partial n_f} dn_f \\
=& \underline{n}_p \int_{\underline{n}_f}^{\overline{n}_f} \frac{\partial^2 K(\alpha\underline{n}_p, n_f)}{\partial n_f \partial n_p} dn_f
\end{aligned}$$

If $\frac{\partial^2 K(n_p, n_f)}{\partial n_f \partial n_p} \leq 0$, i.e., n_p and n_f are substitutes, then $\frac{\partial(K(\alpha\underline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} \leq 0$. Combining with the results that $\bar{\sigma}_1$ and v_1 decrease, we conclude that $\gamma_1(\underline{n}_p, \overline{n}_f) < 0$.

If $\frac{\partial^2 K(n_p, n_f)}{\partial n_f \partial n_p} > 0$, i.e., n_p and n_f are complements, then $\frac{\partial(K(\alpha\underline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} > 0$. In this case, the sign of $\gamma_1(\underline{n}_p, \overline{n}_f)$ can not be pinned down in general. However, it is clear that $\frac{\partial^2 K(n_p, n_f)}{\partial n_f \partial n_p} > 0$ is a necessary condition for $\gamma_1(\underline{n}_p, \overline{n}_f) \geq 0$.

(e) Results related to $\gamma_1(\overline{n}_p, \overline{n}_f)$

Consider

$$\begin{aligned}
\frac{\partial(K(\alpha\overline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} =& \frac{\partial(K(\alpha\overline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \overline{n}_f) + K(\alpha\underline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} \\
=& \frac{\partial(K(\alpha\overline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \overline{n}_f))}{\partial\alpha} + \frac{\partial(K(\alpha\underline{n}_p, \overline{n}_f) - K(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha}
\end{aligned}$$

From the results above, we know that the first term is negative. The second term is exactly the same as in part (d) of this proof. Therefore $\gamma_1(\overline{n}_p, \overline{n}_f) < \gamma_1(\underline{n}_p, \overline{n}_f)$.

We can also rewrite the term above as

$$\begin{aligned}\frac{\partial(K(\alpha\overline{n}_p,\overline{n}_f)-K(\alpha\underline{n}_p,\underline{n}_f))}{\partial\alpha}=&\frac{\partial(K(\alpha\overline{n}_p,\overline{n}_f)-K(\alpha\overline{n}_p,\underline{n}_f)+K(\alpha\overline{n}_p,\underline{n}_f)-K(\alpha\underline{n}_p,\underline{n}_f))}{\partial\alpha}\\=&\frac{\partial(K(\alpha\overline{n}_p,\overline{n}_f)-K(\alpha\overline{n}_p,\underline{n}_f))}{\partial\alpha}+\frac{\partial(K(\alpha\overline{n}_p,\underline{n}_f)-K(\alpha\underline{n}_p,\underline{n}_f))}{\partial\alpha}\end{aligned}$$

The second term is exactly the same as in part (c) of this proof. The sign of the first term depends on the sign of $\frac{\partial^2 K(n_p,n_f)}{\partial n_f \partial n_p}$ as in part (d).

Thus $\gamma_1(\overline{n}_p,\overline{n}_f)<\gamma_1(\overline{n}_p,\underline{n}_f)<0$ if n_p and n_f are substitutes, and $\gamma_1(\overline{n}_p,\underline{n}_f)<\gamma_1(\overline{n}_p,\overline{n}_f)<\gamma_1(\underline{n}_p,\overline{n}_f)$ if n_p and n_f are complements. □

Proof of Proposition 3

Proof. To calculate the standard deviation of excess holding on asset j by investors of type (n_p,n_f), it is useful to consider the following holding dispersion. For a given investor i,

$$\begin{aligned}E[(q_{ij}-\bar{q}_j)^2]=&\frac{1}{\rho^2}E[(\Delta_{ij}(v_j^{\frac{1}{2}}u_j+w_j)+K_{ij}\epsilon_{ij})^2]\\=&\frac{1}{\rho^2}E[(K_{ij}-\bar{K}_j)^2(v_j^{\frac{1}{2}}u_j+w_j)^2+2(K_{ij}-\bar{K}_j)(v_j^{\frac{1}{2}}u_j+w_j)K_{ij}\epsilon_{ij}+K_{ij}^2\epsilon_{ij}^2]\\=&\frac{1}{\rho^2}[(v_j+\rho^2\bar{\sigma}_j^2\bar{x}_j^2)(K_{ij}-\bar{K}_j)^2+K_{ij}]\end{aligned}$$

The first equality follows from the expression of $q_i-\bar{q}$. $\Delta_{ij}=K_{ij}-\bar{K}_j$ has been used in the second line. The third equality follows from the fact that $u\sim N(0,1)$, $w_j=\rho\bar{\sigma}_j\bar{x}_j$ and $E[\epsilon_{ij}]=K_{ij}^{-1}$. USing these results,

$$E[(q_{ij}-\bar{q}_j)^2]-E[(q_{ij}-\bar{q}_j)]^2=\frac{1}{\rho^2}[v_j(K_{ij}-\bar{K}_j)^2+K_{ij}]$$

Since investors belonging to the same type have the same K_{ij}, we have (3)

$$\delta_j(n_p,n_f)=\frac{1}{\rho}\sqrt{v_j(K_j(n_p,n_f)-\bar{K}_j)^2+K_j(n_p,n_f)}$$

Noting that $K_j(n_p,n_f)$ increases while v_j decreases after the shock. We conclude that if δ_j decreases then the first term under the square root dominates. In addition, as $\frac{\partial v_j}{\partial K_j}=$

$-\frac{2(\rho^2\sigma_x^2+\bar{K})}{(\sigma^{-1}+\sigma_p^{-1}+\bar{K})^3}$, the magnitude of the change due to an infinitesimal increase in K becomes unbounded when σ_x goes to infinity. Thus when σ_x is large enough, we only need to consider the first term,

$$\delta_j(n_p, n_f) \approx \frac{1}{\rho}\sqrt{v_j}|K_j(n_p, n_f) - \bar{K}_j| \tag{6}$$

For the base type investors, we have immediately that

$$\delta_j(\underline{n}_p, \underline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\underline{n}_p, \underline{n}_f)),$$

since they have the smallest K. $\delta_j(\underline{n}_p, \underline{n}_f)$ is likely to decrease after the shock, due to a smaller v_j and a possibly smaller distance to the average precision, $\bar{K}_j - K(\underline{n}_p, \underline{n}_f)$. The latter comes from the concavity of the production function.

As the dispersion for the untreated asset is not affected, we only need to consider the changes among different type of investors for the treated asset. If $K(\overline{n}_p, \overline{n}_f)$ is large enough such that $K(\overline{n}_p, \underline{n}_f) < \bar{K}_j$ and $K(\underline{n}_p, \overline{n}_f) < K_j$, (6) becomes

$$\delta_j(\overline{n}_p, \underline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\overline{n}_p, \underline{n}_f))$$

and,

$$\delta_j(\underline{n}_p, \overline{n}_f) = \frac{1}{\rho}\sqrt{v_j}(\bar{K}_j - K(\underline{n}_p, \overline{n}_f))$$

The impact of the shock on the difference in dispersion between the IT-intensive type and the base type is then given by,

$$\begin{aligned}\frac{\partial(\delta_j(\overline{n}_p, \underline{n}_f) - \delta_j(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} =& \frac{1}{2\rho\sqrt{v_j}}\frac{\partial v_j}{\partial\alpha}(K(\alpha\underline{n}_p, \underline{n}_f) - K(\alpha\overline{n}_p, \underline{n}_f)) \\ &+ \frac{1}{\rho}\sqrt{v_j}\frac{\partial(K(\alpha\underline{n}_p, \underline{n}_f) - K(\alpha\overline{n}_p, \underline{n}_f))}{\partial\alpha}\end{aligned}$$

Since $\frac{\partial v_j}{\partial K} < 0$, the first term is positive. The second term is also positive, as shown in the proof of Proposition 1. Thus, we conclude that $\frac{\partial(\delta_j(\overline{n}_p, \underline{n}_f) - \delta_j(\alpha\underline{n}_p, \underline{n}_f))}{\partial\alpha} > 0$. For the difference between the finance-intensive type and the base type, we have

$$\frac{\partial(\delta_j(\alpha \underline{n}_p, \overline{n}_f) - \delta_j(\alpha \underline{n}_p, \underline{n}_f))}{\partial \alpha} = \frac{1}{2\rho\sqrt{v_j}} \frac{\partial v_j}{\partial \alpha} (K(\alpha \underline{n}_p, \underline{n}_f) - K(\alpha \underline{n}_p, \overline{n}_f)) + \frac{1}{\rho} \sqrt{v_j} \frac{\partial (K(\alpha \underline{n}_p, \underline{n}_f) - K(\alpha \underline{n}_p, \overline{n}_f))}{\partial \alpha}$$

The first term is again positive. The second term is positive in case of substitution and negative in case of complementarity. Therefore, a negative change in the difference implies complementarity.

□

Table 1: Summary statistics. All continuous variables are winsorized at the top and bottom one percent to mitigate the influence of extreme values

Variables	N	mean	std	p25	median	p75
JobPerInst(IT)	421	10.24	115.51	0.00	0.00	1.00
JobPerInst(finance)	421	8.01	74.91	0.00	1.00	2.00
Assets(base)	227	1981.37	4395.51	130.09	549.92	1800.86
Assets(IT)	79	11920.84	33407.91	526.73	1992.71	8372.48
Assets(finance)	133	447.86	2505.98	29.12	81.70	203.06
Assets(Bi)	38	1381.04	4168.60	29.23	109.77	544.57
MarketCap	496873	5.63	19.83	0.38	1.09	3.10
RetHolding	486935	-1.29	11.02	-2.53	-0.16	0.62
RetTrading	357328	0.50	16.17	-0.54	0.00	0.67
Fraction	96626	0.03	0.05	0.00	0.01	0.05
Fraction_Scaled	96627	1.50	5.46	0.06	0.24	0.80
InstOwn	96609	0.59	0.28	0.36	0.64	0.83
Book-to-market	95146	0.82	0.73	0.36	0.65	1.05
Debt-to-equity	95182	2.29	3.96	0.41	1.02	2.44
Dispersion	74319	0.40	1.84	0.01	0.04	0.19
AveMarketCap	3289	0.34	0.85	0.01	0.05	0.25
NumOfStocks	3122	131.20	259.62	9.00	30.00	120.00
Turnover	2837	0.16	0.12	0.06	0.12	0.23

Table 2: Classification for the twenty largest institutions.

IT-intensive	Finance-intensive	Bi-intensive
LORD, ABBETT & CO. LLC	BARCLAYS BANK PLC	MERRILL LYNCH & CO INC
LAZARD CAPITAL MARKETS LLC	MOORE CAPITAL MANAGEMENT, INC.	UBS SECURITIES LLC
KEYBANK NATIONAL ASSOCIATION	BALYASNY ASSET MANAGEMENT LP	SG AMERICAS SECURITIES, LLC
D. E. SHAW & CO., L.P.	ROYAL BANK OF CANADA	BAIN CAPITAL, LLC
CITIGROUP INC	FIRST QUADRANT L.P.	BLACKSTONE GROUP
GENERAL ELECTRIC COMPANY	MASON CAPITAL MANAGEMENT	SOROS FUND MANAGEMENT, L.L.C.
RCM CAPITAL MANAGEMENT LLC	VISIUM ASSET MANAGEMENT, LP	BRIDGEWATER ASSOCIATES INC.
RENAISSANCE TECHNOLOGIES CORP.	TUDOR INVESTMENT CORPORATION	TPG CAPITAL, L.P.
LOEWS CORPORATION	ROCHDALE INVESTMENT MGMT LLC	WOLVERINE ASSET MGMT, L.L.C.
CREDIT SUISSE SECS (USA) LLC	COBALT CAPITAL MGMT, INC.	DAVIDSON KEMPNER CAP MGMT L.L.
AQR CAPITAL MANAGEMENT, LLC	TFS CAPITAL LLC	ENVESTNET ASSET MGMT, INC.
FISHER INVESTMENTS	TREMBLANT CAPITAL GROUP	INTEL CORPORATION
COHEN & STEERS CAP MGMT, INC.	THIRD POINT LLC	QVT FINANCIAL LP
JACOBS LEVY EQUITY MGMT, INC.	SIRIOS CAPITAL MGMT, L.P.	ATTICUS CAPITAL, L.P.
HIGHBRIDGE CAPITAL MGMT, LLC	U.S. GLOBAL INVESTORS, INC.	CTC LLC
TWO SIGMA INVESTMENTS, LLC	SANDLER CAPITAL MANAGEMENT	NORTHCOAST ASSET MGMT LLC
ACADIAN ASSET MANAGEMENT LLC	CAPSTONE INVT ADVISORS, LLC	NUVEEN LLC
GAMCO INVESTORS, INC.	RHO CAPITAL PARTNERS, INC.	MARINER INVESTMENT GROUP LLC
BRANDES INVT PARTNERS, LP	BRAHMAN CAPITAL CORP.	TRAXIS PTNR LLC
ARROWSTREET CAPITAL, L.P.	CYRUS CAPITAL PARTNERS, L.P.	ELLINGTON MGMT GROUP, L.L.C.

Table 3: Results of regression on institutional characteristics. $y_{it} = \alpha + \sum_{\theta} \beta_\theta Type_{i\theta t} + \alpha_t + Controls_{it}$. In column (1) the dependent variable is institution's asset turnover. The dependent variable in column (2) is the log value of average market capitalization of portfolio stocks of an institution at time t. The average is weighted by portfolio weights. Column (3) uses the log value of number of portfolio stocks as the dependent variable. In all the regressions, I control for institution size. Standard errors are clustered at quarter and institution level.

	(1)	(2)	(3)
	Turnover	AveMarketCap	LogNumStocks
Bi-intensive	0.00972	0.64836	-0.24438
	(0.320)	(1.225)	(-0.421)
IT-intensive	0.13873***	-0.58262	0.70796
	(3.136)	(-1.034)	(1.210)
Finance-intensive	-0.08354*	-0.95361*	-0.41504
	(-2.117)	(-1.910)	(-0.982)
Control	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	2,837	3,289	3,122
R^2	0.06	0.62	0.40

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Results of regression on excess returns(bps). $r_{ijqc} = \sum_{\theta} \gamma_{\theta} \, Type_{i\theta c} \times XBRL_{jc} \times Post_q + \beta^r \; XBRL_{jc} \times Post_q$ + Institution FE + Stock-Type-Cohort FE + Type-Period-Cohort FE + $Controls_{ijqc}$. Period q is measured as the number of quarters relative to the implementation XBRL mandate in each cohort. The dependent variable in the first two columns is excess holding return. The dependent variable in columns (3) and (4) is excess trading return. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$. Controls include asset under management, institution turnover, and stock market capitalization. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)	(4)
	Holding	Holding	Trading	Trading
XBRL × Post	-0.67671		-0.04242	
	(-0.565)		(-0.167)	
XBRL × Post × IT	-0.60411***	-0.76533**	-0.41631**	-0.73453***
	(-3.679)	(-3.098)	(-2.715)	(-4.484)
XBRL × Post × Finance	0.82200**	1.10378*	2.19206***	1.84578**
	(2.574)	(1.996)	(3.566)	(2.729)
XBRL × Post × Bi-intensive	-0.50870*	-0.69297*	-0.04490	-0.31821
	(-2.053)	(-1.847)	(-0.104)	(-0.811)
Controls	Yes	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes	Yes
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes	Yes
Observations	470,871	468,850	346,288	343,846
R^2	0.24	0.51	0.36	0.73

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Results of regression on stock ownership. $Ownership_{j\theta qc} = \sum_{\theta} \gamma_{\theta}^{s} \, Type_{i\theta c} \times XBRL_{jc} \times Post_{q} + \beta^{s} \, XBRL_{jc} \times Post_{q} +$ Stock-Type-Cohort FE + Type-Period-Cohort FE + $Controls_{ijqc}$. The dependent variable in the first two columns is the fraction of total share outstanding owned by each type of investors. In column (3) and column (4), the fraction is divided by the total assets of the type. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$ and Stock FE. Control variables include stock market capitalization, book-to-market ratio, debt-to-equity ration, total institutional ownership and average asset under management of each type. Standard errors are clustered at stock and quarter-cohort level.

	(1)	(2)	(3)	(4)
XBRL × Post	0.00125		0.22367**	
	(1.518)		(2.882)	
XBRL × Post × IT	-0.00657***	-0.00630***	-0.22843**	-0.20367**
	(-6.207)	(-5.900)	(-3.102)	(-2.612)
XBRL × Post × Finance	-0.00108	-0.00134	0.88590***	0.86258**
	(-1.267)	(-1.524)	(3.180)	(2.975)
XBRL × Post × Bi-intensive	-0.00173*	-0.00154	-0.32760***	-0.28570***
	(-1.891)	(-1.621)	(-3.690)	(-3.208)
Controls	Yes	Yes	Yes	Yes
Stock FE	Yes	No	Yes	No
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes	Yes
Observations	87,642	84,972	87,662	84,817
R^2	0.95	0.97	0.77	0.84

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Results of regression on stock holding dispersion, $\delta_{j\theta qc} = \sum_{\theta} \gamma_{\theta} \ Type_{i\theta c} \times XBRL_{jc} \times Post_{qc} + \beta^{\delta} \ XBRL_{jc} \times Post_{qc}$ + Stock-Type-Cohort FE + Type-Period-Cohort FE + $Controls_{ijqc}$. In columns (2) and (4), I control for Stock-Period-Cohort FE, which absorbs $XBRL \times post$ and Stock FE. Control variables include stock market capitalization, book-to-market ratio, debt-to-equity ration, total institutional ownership and average asset under management of each type. In columns (3) and (4), Stock-Type-Cohort FE is replaced by Treated-Type-Cohort FE. Standard errors are clustered at stock and quarter-cohort level.

	(1)	(2)	(3)	(4)
XBRL × Post	-0.09160		-0.09992*	
	(-1.424)		(-2.013)	
XBRL × Post × IT	0.09408	0.09853	0.10746*	0.11536*
	(1.486)	(1.528)	(2.099)	(2.072)
XBRL × Post × Finance	-0.09361	-0.05572	-0.15142**	-0.14356**
	(-0.691)	(-0.409)	(-2.473)	(-2.249)
XBRL × Post × Bi-intensive	0.01195	-0.00466	0.12168*	0.13588**
	(0.117)	(-0.050)	(2.192)	(2.712)
Controls	Yes	Yes	Yes	Yes
Stock-Cohort FE	Yes	No	Yes	No
Stock-Period-Cohort FE	No	Yes	No	Yes
Type-Period-Cohort FE	Yes	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	No	No
Treated-Type-Cohort FE	No	No	Yes	Yes
Observations	64,599	60,183	66,781	62,813
R^2	0.82	0.87	0.30	0.36

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Assets(millions) by investor type in the matched sample.

	Base_{it}	IT	Base_{fin}	Finance	Base_{bal}	Balanced
count	225.00	73.00	201.00	132.00	199.00	36.00
mean	2071.44	4911.52	858.13	446.33	1330.27	822.18
std	4510.77	8051.18	1027.20	2515.71	2812.11	2323.74
min	1.60	2.09	1.60	0.34	1.60	1.58
median	550.84	1513.23	395.50	75.85	463.63	115.86
max	46805.44	45549.32	6491.78	28686.12	28051.74	13495.52

Table 8: Regressions on excess holding returns (matched sample). $r_{ijtc} = \gamma_\theta \, Type_{i\theta c} \times XBRL_{jc} \times Post_t + XBRL_{jc} \times Post_t + Controls_{ijtc}$ + Institution FE + Stock-Type-Cohort FE + Type-Period-Cohort FE. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)
	IT	Finance	Bi-intensive
IT × XBRL × Post	-0.81285**		
	(-2.871)		
Finance × XBRL × Post		1.28249**	
		(3.083)	
Bi × XBRL × Post			-0.18505
			(-0.417)
Controls	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Stock-Period-Cohort FE	Yes	Yes	Yes
Type-Period-Cohort FE	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes
Observations	307,198	130,744	172,247
R^2	0.49	0.37	0.43

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regressions on excess trading returns (matched sample). $r_{ijqc} = \gamma_\theta \ Type_{i\theta c} \times XBRL_{jc} \times Post_{qc} + XBRL_{jc} \times Post_{qc} + Controls_{ijqc}$ + Institution FE + Stock-Type-Cohort FE + Type-Period-Cohort FE. Standard errors are clustered at institution and quarter-cohort level.

	(1)	(2)	(3)
IT × XBRL × Post	-0.34153***		
	(-3.477)		
Finance × XBRL × Post		1.37825***	
		(4.643)	
Bi × XBRL × Post			-0.29868
			(-1.322)
Controls	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Stock-Period-Cohort FE	Yes	Yes	Yes
Type-Period-Cohort FE	Yes	Yes	Yes
Stock-Type-Cohort FE	Yes	Yes	Yes
Observations	226,084	87,109	113,663
R^2	0.54	0.43	0.47

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Results from regression on analyst coverage data. $Analyst_{it} = \alpha + \sum_{\theta} Type_{i\theta t} + \alpha_t + Controls_{it}$. In columns (1) and (2), the dependent variable is $Coverage_{it}$, which is the average analyst coverage for the stocks held by institution i at time t. Analyst coverage is the number of analysts who issued earning forecasts on a stock within 90 days before a forecasting ending period lies in quarter t. In columns (3) and (4), the dependent variable is the forecasting accuracy $Accuracy_{it}$, which is calculated as the negative of the absolute value of the difference between the actual earnings per share and the median analyst forecast normalized by stock price. Columns (1) and (3) use simple averages. In columns (2) and (4) the averages are weighted using excess holding weights. Standard errors are clustered at institution and quarter level.

	(1)	(2)	(3)	(4)
	Coverage	Coverage(weighted)	Accuracy	Accuracy(weighted)
Bi-intensive	0.52534	-0.07389	-0.00709	0.00019
	(0.772)	(-0.058)	(-1.221)	(0.175)
IT-intensive	-0.10242	2.65686**	-0.01149	-0.00352*
	(-0.165)	(2.051)	(-1.556)	(-1.867)
Finance-intensive	0.06039	-1.61500**	-0.00548	0.00187**
	(0.107)	(-2.159)	(-1.294)	(2.547)
Control	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	16,460	10,339	16,299	10,352
R^2	0.63	0.28	0.13	0.15

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

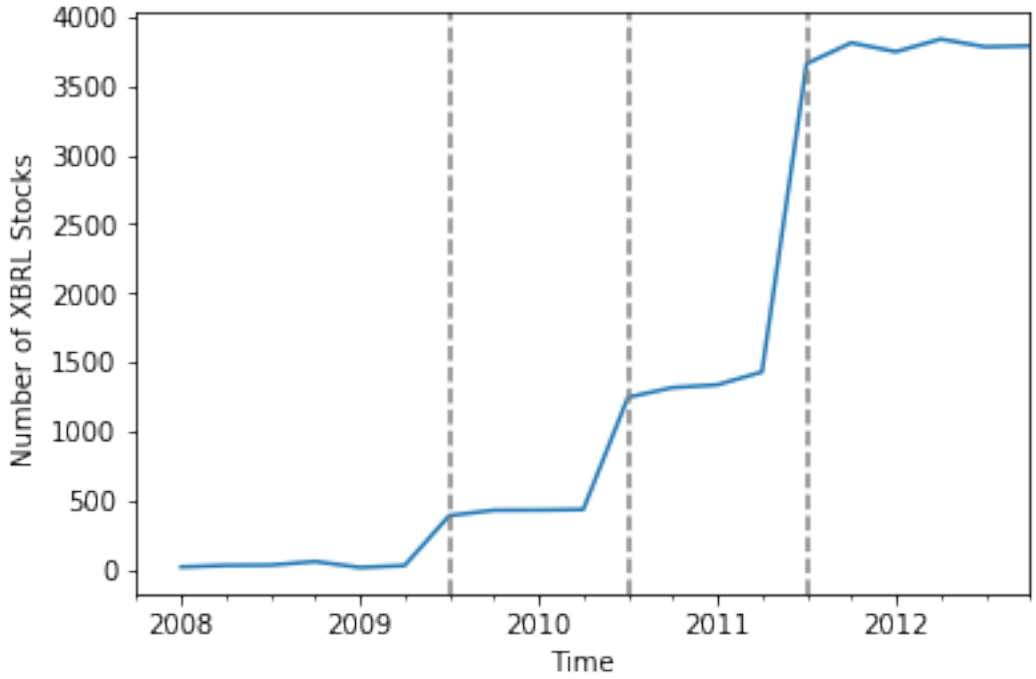

Figure 2: Number of stocks that complied with the XBRL mandate in each quarter.

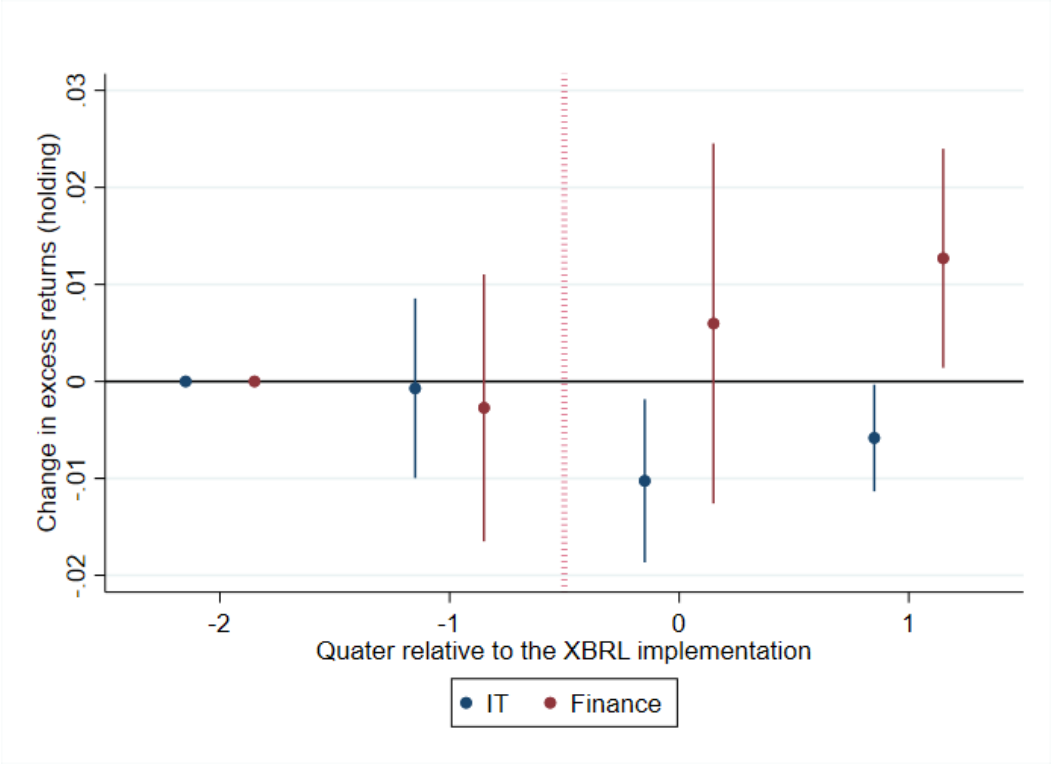

Figure 3: Plot of differential impact for holding returns (in percentage points). It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.

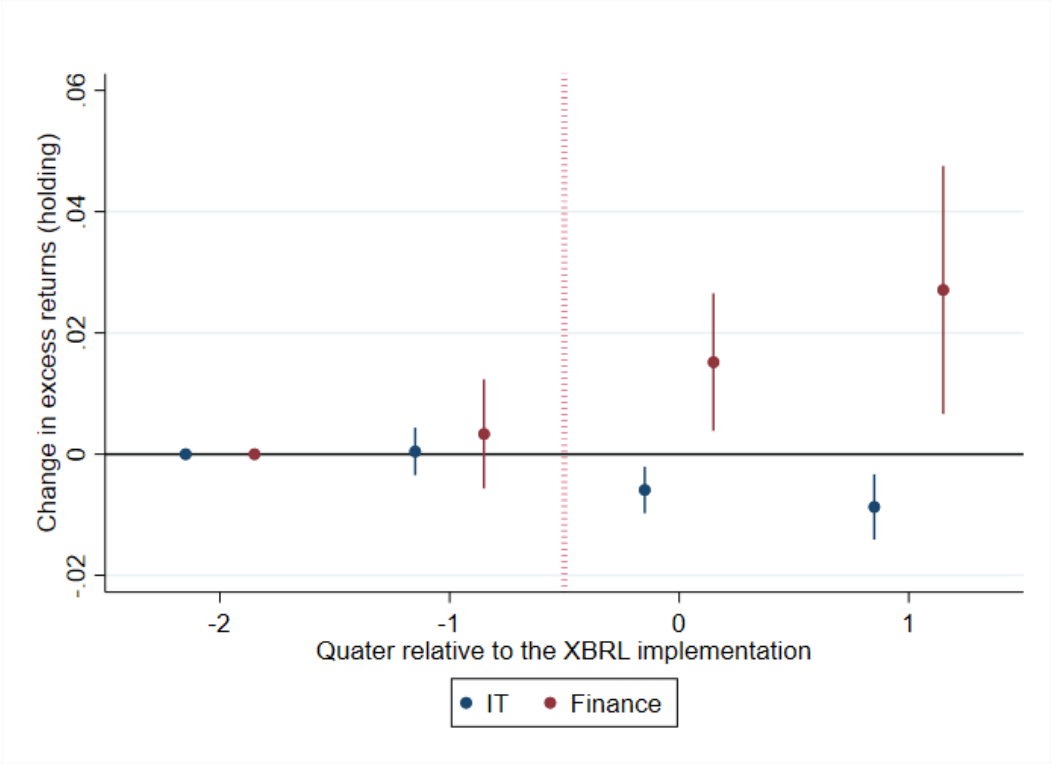

Figure 4: Plot of differential impact for trading returns (in percentage points). It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.

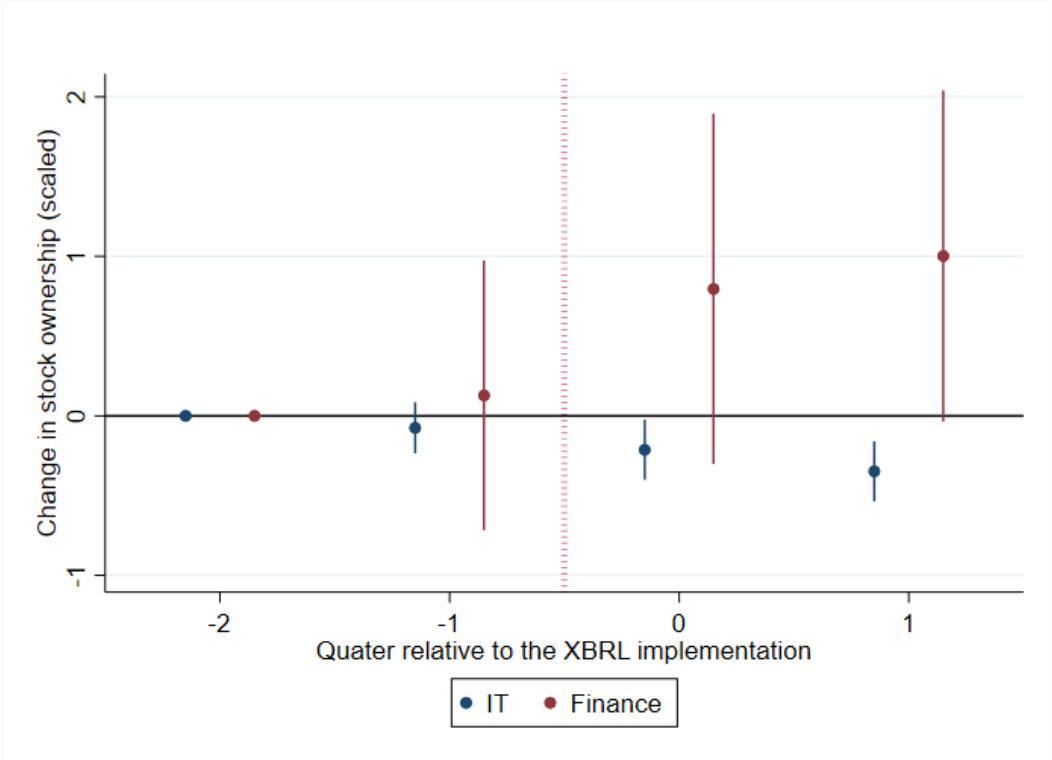

Figure 5: Plot of differential impact for stock ownership. It plots the differential impacts on the IT-intensive (finance-inventive) type and the base type investors on the treated and non-treated stocks, i.e., changes in $Type \times XBRL$ estimates relative to two quarters before the event.