Do Investors Have Data Blind Spots? The Role of Data Vendors in Capital Markets

Sara Easterwood *

August 29, 2024

Abstract

Financial data vendors intermediate the flow of information from firms to investors. I study frictions that arise in the context of this intermediation by focusing on one of the most prominent data vendors in the finance industry – Standard & Poor's ('S&P') Compustat database. Compustat provides subscribers with decades of 10-K and 10-Q data; however, it does not cover every public firm in every period. I show that institutional investment is over 36% below its unconditional mean for firms not covered in Compustat. A quasi-natural experiment confirms a plausibly causal connection between Compustat's data coverage and investor demand: a technology shock at S&P in the 1990s causes a discrete reduction in missing data. This change in data coverage is followed by a significant increase in institutional investment for treated firms relative to control firms. I then show that missing Compustat data is associated with lower informational efficiency of equity prices. These results highlight the role that data vendors play in facilitating the flow of information within financial markets.

*Easterwood, saraeast@vt.edu. I am grateful to my advisors Bradley Paye (chair), Roger Edelen, Gregory Kadlec, and Jeffry Netter for their invaluable guidance and support. I thank Tyler Beason, Elizabeth Bickmore, Felipe Cabezón, Kevin Crotty (discussant), Robert Davidson, Michael Gelman, Raman Kumar, Andrew MacKinlay, Alex Pecora, Vijay Singal, Wei-ling Song, Yessenia Tellez, Ngoc-Khanh Tran, Jin Xu, Penfei Ye, and seminar participants at Louisiana State University, the 2024 SFS Cavalcade North America Conference, and Virginia Tech for helpful comments. I thank Randall Billingsley, Chris Brightman, and Michael Kender for providing incite into the sources and uses of financial data in the professional asset management industry. I thank all members of S&P Global's staff who provided invaluable information regarding the evolution of the Compustat database over time. All errors are my own.

1 Introduction

Over the last several decades, data has become an increasingly important component of the global economy (Farboodi and Veldkamp, 2023). With this evolution, an entire industry of data aggregators has emerged. These data vendors act as information intermediaries in a variety of contexts by collecting and aggregating data on clients' behalf. Their services are often utilized in financial research and practice because they can reduce the implicit and explicit costs associated with acquiring and processing information (Blankespoor et al., 2020).[1] A nascent literature examines the role these intermediaries play in capital markets (D'Souza et al., 2010; Akbas et al., 2018; Schaub, 2018; Kaplan et al., 2021; Bochkay et al., 2022). However, it is not well understood how a data vendor's data coverage ultimately affects information access and investor actions. Specifically, when building and maintaining any database, the vendor must decide what information is collected and when, and when and how that information is updated over time as, e.g., restatements are made, disclosure regulations change, the composition of firms changes, and so forth. This raises an important question: can these data coverage decisions impact real outcomes, such as investment? This paper sheds light on this issue.

There are many data vendors in the finance industry that provide many different types of information. However, establishing a causal connection between any data vendor's data coverage and investor actions presents a formidable identification challenge. Data aggregators have an incentive to cater their coverage to client demand. Thus, when investor demand for a firm's information is high, data vendors are more likely to cover that firm. Indeed, many vendors readily state that client requests influence data coverage decisions, and studies such as D'Souza et al. (2010), Akbas et al. (2018), and Schaub (2018) emphasize the interdependence between investor demand and information dissemination speeds in major financial

[1] Kolanovic and Krishnamachari (2017) estimate that the investment management industry spent approximately $2-3 billion on data and related information technologies in 2017. They also forecast that this expenditure will experience double digit annual growth on the order of 10-20% in the following years.

databases such as Compustat, I/B/E/S, and First Call, respectively. Additionally, more salient firms are more likely to have both higher investor demand and more comprehensive data coverage in financial databases. If this salience is time-varying, it is difficult to resolve this correlated omitted variable issue.

In this study, I evaluate the connection between institutional investment and data coverage in Standard & Poor's Compustat database. Compustat provides subscribers with public firms' financial statement information, and is one of the oldest and most prominent data vendors in the finance industry.[2] It therefore plays a potentially significant role as an information intermediary. Importantly, a technology shock at S&P in the 1990s provides a quasi-natural experiment to evaluate a plausibly causal connection between Compustat's data coverage and investor demand. I use the development and evolution of Compustat as an empirical experimental setting to assess the role that data vendors play in facilitating the flow of information within capital markets.

Standard & Poor's aggregates firms' 10-K and 10-Q filings, and provides subscribers with a standardized dataset (Compustat) consisting of income statement, balance sheet, and statement of cash flow information. As of 2024, Compustat covers tens of thousands of firms over a nearly 75-year historical period. It does not, however, cover every publicly traded firm in all fiscal years or quarters. This missing data is a function of how S&P built and expanded Compustat over time,[3] and financial statement information for the uncovered firms can be obtained from other sources, such as SEC filings. Summary statistics reveal that firms with missing Compustat data have 3.3% lower institutional ownership on average relative to firms that are covered in the database. Given that institutions are the dominant investors in capital markets,[4] and that the unconditional average fraction of institutions that own shares of a firm is 4%, this is an economically very large effect. However, this statistic fails to control for other

[2]S&P began developing Compustat in 1962, roughly 20-30 years before similar financial data vendors such as Bloomberg, Worldscope and FactSet. Compustat marketing materials state that tens of thousands of hedge funds, money managers, analysts, researchers, and corporations utilize the database.

[3]I obtain information regarding the evolution of Compustat's data coverage from many email exchanges with S&P Global support staff, as well as from S&P's Compustat data guide (Standard & Poor's, 2003).

[4]See, e.g., Gompers and Metrick (2001); Edelen et al. (2016); Koijen and Yogo (2019), and many others.

firm characteristics known to predict variation in institutional demand (e.g., Falkenstein, 1996; Gompers and Metrick, 2001; Edelen et al., 2022). Using a panel regression analysis that includes controls for important firm characteristics such as size and index inclusion, I show that firms that are not covered in Compustat have 1.5% lower institutional ownership on average relative to covered firms. This continues to be an economically large effect: even controlling for important firm characteristics, institutional investment in firms with missing Compustat data is over 36% below its mean.

Although suggestive, the summary statistics and associated panel regressions do not address all concerns related to correlated omitted variable bias or reverse causality. In an effort to evaluate if there is a plausibly causal connection between Compustat's data coverage and institutional investment, I next introduce a novel, quasi-natural experiment. Standard & Poor's has always maintained two separate internal data collection systems: one for firms in financial services industries, and one for firms in other industries. (S&P typically refers to these non-financial firms as 'industrials'.) For several decades, the financial firms' data collection system's data coverage was not as comprehensive as the industrial firms' data collection system and, until the early 1990s, Compustat provided accounting data for only a subset of banks and financial services institutions. In the early 1990s, S&P significantly enhanced their financial firm data collection system. As a result of this technology shock, they began collecting financial statement data for all financial services firms from the 1993 fiscal-year onward.[5] This data became available in the Compustat North America database between 1993 and 1994, and led to a discrete, precipitous reduction in the fraction of publicly listed firms with missing Compustat data.

I conduct a difference-in-differences analysis to evaluate the impact of this technology shock, and the associated change in data coverage, on investor demand. In this regression setting, treated firms are defined as financial services firms with no data coverage in Compustat prior to 1993. Control firms are defined as financial services firms with complete data

[5]S&P Global support staff provided information regarding this technology shock, and the associated change in data coverage, in email correspondences which took place in 2023 and 2024.

coverage prior to the technology shock, and are matched to the treated sample based on observable firm characteristics (institutional ownership, size, index membership, and trading volume). By construction, the treatment and control groups have equal average levels of institutional ownership at the beginning of the sample period (1988). I show that the treated and control samples continue to have roughly equal average levels of ownership throughout the late-1980s and early-1990s. Then, following the technology shock, treated firms experience a significant increase in ownership relative to control firms. Throughout the mid-to-late 1990s, the increase in institutional ownership for treated firms is nearly three times larger than the increase in ownership for control firms.[6] Collectively, this analysis is consistent with the following causal interpretation: a meaningful fraction of institutional investors use Compustat to access firms' financial statement data. Thus, when a firm is not covered in Compustat, those institutions do not invest in the firm. Once the firm begins to be covered, institutional investment increases.

A primary concern with the difference-in-differences analysis is that the technology shock coincides with other firm-level shocks that may affect investor actions. Importantly, however, in order to provide a plausible alternative explanation, such shocks must effect the subset of treated firms *differently* than the subset of control firms. At least two notable events coincide with this shock: the implementation of the SEC's Electronic Data Gathering, Analysis, and Retrieval ('EDGAR') system, which took place over the period 1993–1996, and the 1994 passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act. It is unlikely that either of these events differentially impacted the treated versus control samples. The EDGAR system's introduction improved access to all firms' public disclosures, including their accounting data (Gao and Huang, 2020; Kim et al., 2024). However, phase-in schedules across the treatment and control groups are similar, and the treatment effect is statistically equal across firms that began filing electronically on EDGAR in each year between 1993 and

[6]To be more specific, treated and control firms have approximately 10 institutional owners on average at the end of 1992. Across the period 1993–1999, treated firms' average ownership increases to over 33 institutions while controls firms' average ownership increases to approximately 18 institutions.

1996. Likewise, the Riegle-Neal Act relaxed several federal regulations related to inter-state banking and branching. Given that both the treatment and control groups are composed exclusively of financial services firms, this reduces concerns that Riegle-Neal provides a plausible alternative explanation. Empirically, I show that, despite the fact that Riegle-Neal's impact varied across individual states (Rice and Strahan, 2010), there is not a discernible difference in the treatment effect across firms located in states with very restrictive versus very open banking laws post-Riegle-Neal.

Compustat provides access to financial statement data. Thus, it is reasonable to conclude that those institutions that utilize Compustat are using accounting data when making investment decisions. There are several potential ways in which this data may be useful. Trading strategies may incorporate accounting information, investment mandates may include restrictions based on financial ratios, and financial statement analysis may help portfolio managers 1) illustrate that they have done their due diligence as fiduciaries, and/or 2) identify variation in risk exposure or mispricing. While it is possible to supplement Compustat with self-collected accounting information, lower average investment in uncovered firms suggests that many institutions do not do so. From an equilibrium perspective, it is only optimal for portfolio managers to self-collect data if the marginal benefits of obtaining the information are greater than the marginal costs associated with collecting that information (Grossman and Stiglitz, 1980). It is possible that, for many institutions, the costs associated with acquiring firms' disclosures and maintaining a database for these uncovered firms exceed the benefits they might accrue from obtaining and trading on the additional accounting data. These 'costs' can include the time and resources spent to acquire the information, develop a standardization method, build and maintain a database, and ensure that the use of the data satisfies any relevant due diligence requirements.

Collectively, this suggests that several empirical patterns should emerge. First, institutions and mutual funds whose strategies and/or investment mandates do not incorporate accounting information (e.g., index funds) should be more likely to invest in firms with no

Compustat data. Likewise, institutions more constrained by agency conflicts and prudent-man regulations should be less inclined to invest in firms with no Compustat data. This is because these institutions are more likely to utilize financial statement analysis to illustrate that they have done their due diligence as fiduciaries, and because their higher burden of due diligence implies that self-collecting data is potentially more costly. Finally, more skilled portfolio managers, who are better able to identify variation in risk exposure or missing pricing, should be more inclined to self-collect data because the potential marginal benefit of obtaining the data is higher for these investors.

Consistent with these hypotheses, I find that funds that are less likely to use accounting information when implementing their trading strategies and following their investment mandates (e.g., index funds, very high turnover funds including high frequency traders, and funds whose trades are most correlated with momentum/contrarian strategies) invest in a significantly larger fraction of firms with missing Compustat data compared to funds that are more likely to use accounting information. Likewise, I find that institutions that are more constrained by agency conflicts and prudent-man regulations (e.g., smaller institutions, banks, insurance companies, and pension funds) are significantly less likely to invest in firms with missing Compustat data relative to less constrained institutions. Finally, I find that more actively managed funds are more likely to invest in firms with missing Compustat data relative to less actively managed funds. To the extent that activeness reflects the portfolio manager's skill and ability to identify variation in risk exposure or mispricing, this is consistent with the notion that investors are more likely to incur the costs associated with acquiring information when the marginal benefit is higher. To the extent that activeness is a function of the portfolio manager's investment constraints, this is also consistent with the notion that agency conflicts influence institutions' dependence on an established data vendor (Compustat) to support their due diligence efforts.

I conclude this study by evaluating the economic consequences of lower institutional investment in firms with missing Compustat data. Merton (1987) links investor attention to

market efficiency, and suggests that 'neglected' firms that face significantly less scrutiny by many market participants will have less informationally efficient equity prices. Several recent studies provide empirical support for this hypothesis (Boone and White, 2015; Ben-Rephael et al., 2017; Kacperczyk et al., 2021; Chen et al., 2022). This suggests that Compustat's data coverage should affect market efficiency because of its impact on investor attention.

I find that earnings surprises, post earnings announcement drift, several measures of price delay proposed by Hou and Moskowitz (2005), and several measures of daily return autocorrelations are significantly larger in magnitude for firms with missing Compustat data. I also find that this effect is mitigated if there are sufficiently many institutions investing in the uncovered firms and/or if there are sufficiently many analysts following the uncovered firms. For example, results indicate that earnings surprises, measured via cumulative abnormal returns around earnings announcements, are 0.3–0.5% larger in magnitude for firms not covered in Compustat. This effect is negated by an (approximately) one standard deviation increase in institutional ownership or analyst coverage. These results are collectively consistent with the notion that limited data coverage by a prominent data vendor reduces the informational efficiency of equity prices via its impact on market participation.

Compustat's data coverage has consistently improved over time, and there are many more alternative sources from which investors can obtain financial statement data post-2010 relative to earlier decades. This suggests that frictions related to the intermediation of *financial statement* information have attenuated over time. However, financial statement data is only one subset of potentially relevant information. The last several decades were accompanied by continuous and exponential improvements in information technologies and data gathering methods. Data vendors such as Glassdoor, the Carbon Disclosure Project, the Privacy Rights Clearinghouse, online retailers (e.g., Amazon), and social media platforms (e.g., Facebook, Twitter) now provide information related to employee satisfaction, pollution, cybersecurity risk, consumer attention, retail investor sentiment, and other firm characteristics. All of this information is plausibly relevant to a significant fraction of investors. As such, the role that

data vendors play as information intermediaries, and their impact on investor actions, will continue to be relevant in studies of capital markets.

Related Literature

This paper relates perhaps most directly to D'Souza et al. (2010), who examine the relation between institutional demand and the speed with which accounting information is disseminated in Compustat. Their analyses suggest that institutions prefer richer information environments, and that Compustat updates financial statement information faster when institutional demand is higher. A number of aspects of my paper are novel relative to D'Souza et al. (2010). I focus on whether or not firms are covered in Compustat, as opposed to Compustat's information dissemination speeds. Importantly, this allows me to utilize a quasi-natural experiment to show that variation in Compustat's data coverage causes variation in institutional ownership. To the best of my knowledge, I am the first to document a plausibly causal connection between a data vendor's data coverage and investor demand.

There are many important financial data vendors and related information technologies, and a growing literature examines their roles in capital markets. Ben-Rephael et al. (2017) use institutions' news searching activity on Bloomberg terminals to develop a measure of institutional investor attention. Akbas et al. (2018) connect the delay with which analysts' earnings forecasts are activated in I/B/E/S to measures of investor demand and market efficiency. Schaub (2018) links information dissemination speeds in the First Call database to price efficiency. Kaplan et al. (2021) show that Thomson Reuters subjectively excludes forecasts from I/B/E/S, and that these discretionary exclusions lead to more accurate consensus forecasts. Bochkay et al. (2022) show that, following a change in Thomson Reuters' methodology for estimating street earnings, analysts' forecasts become less dispersed and more accurate for the subset of treated firms. Bowles et al. (2024) study the timing of anomaly returns around information releases in Compustat, and show that delayed information processing by investors at least partially explains anomaly returns.

Focusing on SEC resources, Gao and Huang (2020), Kim et al. (2024), and Hirshleifer

and Ma (2024) use the staggered implementation of EDGAR to study the impact that information acquisition costs have on measures of information production (Gao and Huang, 2020) and the performance of accounting-based anomalies (Kim et al., 2024; Hirshleifer and Ma, 2024). Kim and Kim (2023) compare firms that file their public disclosures on EDGAR to firms that file on FDICconnect, and link differences in information processing costs across the two platforms to differences in market efficiency. Crane et al. (2023) and Bowles and Reed (2024) use data regarding information requests from EDGAR to link the information acquisition behavior of hedge funds and mutual funds, respectively, to their performance.

This paper also contributes to an emerging literature on the data economy. Goldfarb and Tucker (2019) and Farboodi and Veldkamp (2023) provide more complete reviews of this literature. Related work includes Farboodi and Veldkamp (2020), who develop a theoretical framework to explore the economic consequences of improvements in information processing efficiency. Jones and Tonetti (2020) explore how differences in data property rights determine data's use in the economy and its effect on output, privacy, and welfare. Farboodi et al. (2022) develop a measure of the quantity of data investors have about different groups of assets, and Farboodi and Veldkamp (2024) develop a framework to measure and value data.

Finally, this paper contributes to a growing literature that recognizes the importance of missing and/or incorrect data in prominent economic databases. Chen et al. (2015) use the number of missing variables in Compustat to measure the level of detail in firms' annual 10-K's. Bryzgalova et al. (2024), Chen and McCoy (2024), and Freyberger et al. (2024) discuss alternative econometric methods that researchers can use to address missing data. Chychyla and Kogan (2015), Boritz and No (2020), and Du et al. (2023) compare "as-filed" eXtensible Business Reporting Language ('XBRL') data to accounting data obtained from Compustat, and emphasize that there are often significant discrepancies between the two. Ljungqvist et al. (2009), Chuk et al. (2013), and Karpoff et al. (2017) examine the accuracy and completeness of data obtained from I/B/E/S, First Call, and popular financial misconduct databases, respectively.

2 Financial Statement Information Intermediaries

In research and in practice, it is widely acknowledged that firms' financial statements provide information that is critically important to many investors.[7] There are three basic sources from which investors can obtain this information: financial data vendors, the SEC, or directly from a firm, either by requesting a physical copy of the firm's disclosures or, in some cases, from the firm's website. I describe alternative data vendors and SEC resources in the following subsections.

2.1 Data Vendors

Standard & Poor's is one example of a financial data vendor, and Compustat is their financial statement database. S&P began developing and selling subscriptions to Compustat in the early 1960s, and Compustat covers balance sheet and income statement information dating back to 1950 for some firms. The fact that Standard & Poor's has continued to offer Compustat subscriptions over the past six decades provides a strong signal regarding the database's success in both industry and academia. Studies in the accounting and information systems literatures cite Compustat as one of the most prevalent financial databases (Chychyla and Kogan, 2015) and in S&P's own words, they are "the global standard in providing critical financial information."[8]

Other data vendors that provide (or have provided) financial statement information include Compact Disclosure, Dialog, FactSet, Mead Data Central, Value Line, Worldscope, and Bloomberg. Many of these alternatives either: 1) distribute Compustat data (e.g., FactSet[9]), 2) distribute raw 10-K and 10-Q disclosures instead of pre-standardized databases

[7] For example, the Securities and Exchange Commission states that they require firms to disclose their financial statements so that investors have "the timely, accurate, and complete information they need to make confident and informed decisions about when or where to invest," (https://www.sec.gov/about/what-we-do). Likewise, academic research such as Bushee and Noe (2000), Bushee et al. (2003), and Bird and Karolyi (2016) highlights the importance of financial statement disclosure.

[8] https://www.spglobal.com/en/who-we-are/our-history#fourth

[9] FactSet's IPO prospectus states that they obtain their data from several existing data vendors, including Compustat.

(e.g., Mead Data Central), and/or 3) cover only a fraction of firms, with coverage criteria typically based on firm size (e.g., Value Line[10]). Additionally, similar to Compustat, many of these data vendors readily state that when they began building their databases, they started with the largest firms and expanded their data coverage over time to gradually incorporate all public U.S. companies. For example, Worldscope states that they began collecting data for many large North American firms in the early 1980s, that they added medium sized firms in the mid 1980s, and that they added small firms in the mid 1990s (Thomson Reuters, 2010). Likewise, in a conversation with a member of Bloomberg's Equities Help Desk in November, 2023, they indicated that Bloomberg's coverage was limited to only the largest firms when the database was originally built in the early 1980s, and has since expanded over time to cover a broader and more representative sample.

2.2 SEC Resources

Investors can also obtain financial statement information directly from the SEC. There are reference rooms located in Washington D.C., New York, and Chicago, which provide paper copies of firm's financial statements. Studies such as Blankespoor et al. (2020), Gao and Huang (2020), Kothari et al. (2023), Bowles et al. (2024), and Kim et al. (2024) discuss the many issues associated with using these rooms as a source of information. Not only do investors have to be physically present to obtain the information, but there is also evidence suggesting that paper files are routinely lost and/or stolen (Noble, 1982).

In the mid-1990s, the SEC introduced the EDGAR database, where all public corporations are required to electronically file their public disclosures. While several studies have highlighted the role that EDGAR played in massively reducing information acquisition costs (e.g., Gao and Huang, 2020; Kim et al., 2024; Hirshleifer and Ma, 2024), others have noted that EDGAR is *not* most investors' primary source of public disclosures (e.g., Drake et al., 2015). In 2009, the SEC began to require all public firms to file their financial statements

[10]Value Line's financial statement database covers only 1,650 companies (Kim et al., 2024), which corresponds to less than half of all public U.S. firms.

in eXtensible Business Reporting Language format. The XBRL database was implemented in an effort to facilitate data retrieval and analysis (SEC Release No. 33-9002).

2.3 Implications

Anecdotal evidence suggests that many investors rely on data vendors, and not SEC resources, to obtain firms' financial statement information. For example, in a letter to the SEC, Pricewaterhouse Coopers ('PwC') states that: "Based on our discussions with investors and analysts, we understand that investors acquire the large majority of relevant information, including financial data, from sources not controlled by the reporting entity," and, in fact, "the majority of analytical source material is obtained from data aggregators," (Pricewaterhouse Coopers, LLP, 2006, June 8). PwC highlights the very high costs that analysts and investors face if they are forced to manually process paper filings, and indicate that this contributes to wide spread reliance on third party intermediaries. Likewise, Harris and Morsfiled (2012) point out that, unless both the Financial Accounting Standards Board and the SEC make significant efforts to simplify the underlying taxonomy of financial statements, improving data access via systems such as EDGAR and XBRL is highly unlikely to be sufficient for investors to readily use the data provided.

Whether or not a significant fraction of investors have historically relied specifically on Compustat to obtain firms' financial statement data is ultimately an empirical question. Although there are a number of alternative options, none of them is perfectly efficient or costless. The remainder of this paper thus focuses on the following question: does Compustat data coverage affect investor demand?

3 Sample Construction

I obtain monthly and daily stock return and price information from CRSP. Data regarding annual and quarterly financial statement information is from the Compustat North America database. The CRSP and Compustat samples cover the period 1962–2022, and are merged

using the CCM link table. I match accounting data from the $t-1$ fiscal year to price information from July in year t through June in year $t+1$. I define the sample of public firms as the set of firm-month common stock (shrcd = 10 or 11) observations in the CRSP database that are listed on the NYSE, AMEX, or NASDAQ exchanges (exchcd = 1, 2, or 3) and have non-missing and non-zero price (prc, cfacpr) and shares outstanding (shrout, cfacshr) information.

I obtain data regarding the timing of information releases within Compustat North America from the Compustat Point-In-Time ('PIT') database. The PIT data is available for a limited historical period beginning in December 1986. I also merge this data to the CRSP sample using the CCM link table. In this case, I match each CRSP-firm-month observation to PIT data that 1) is recorded as available in Compustat as of the same month-end (PIT Point Date $\leqslant$ the relevant CRSP month-end date), and 2) within the set of available PIT data, is from the firm's most recent fiscal-year end.

Institutions' stock holdings data are from the Thomson Reuters 13f database (s34 file).[11] This data is available at the quarterly frequency beginning in 1980, and contains long-only equity positions for institutional investors with at least $100 million in total equity under management ('EUM'). I categorize institutional investors using classifications from Brian Bushee's website and from Ralph Koijen's website.[12] The adjusted investor types include: insurance companies; banks; pension funds; mutual fund companies; investment companies/advisors, including hedge funds; and miscellaneous. For robustness, I also obtain data regarding mutual fund holdings from the Thomson Reuters 13f database (s12 file).[13] Finally, I obtain summary data describing analyst coverage and earnings announcements

[11] The SEC requires all institutional investors who manage equity investments exceeding $100 million in any of the past four quarters to report their quarterly holdings on form 13F within 45 days of the end of the quarter. Holdings of less than 10,000 shares or $200,000 in market value are exempted.

[12] https://accounting-faculty.wharton.upenn.edu/bushee/ and https://www.koijen.net/index.html, respectively.

[13] A 'Mutual Fund Company' from the s34 file reflects a family of potentially many mutual funds. A mutual fund from the s12 file reflects a single mutual fund. Koijen and Yogo (2019) categorize institutions from the s34 file as 'Mutual Fund Companies' if 1) their type code is 3, 4, or 5, and 2) their name and assigned number match a record from the s12 file.

from I/B/E/S, which is available beginning in 1976. The final CRSP/Compustat/13f/IBES merged dataset, including the institutional investor type classifications, covers the period January 1980 – December 2021.

I construct two measures of aggregate institutional ownership. The first, denoted $FNIO_{i,q}$, is equal to the number of institutions that hold shares of stock i in quarter q, scaled by the total number of institutions in the 13f dataset in quarter q. The second, denoted $FSIO_{i,q}$, is equal to the fraction of stock i's shares outstanding held by institutions in q.[14] Measures of mutual fund ownership, $FNMF$ and $FSMF$, are defined similarly using the s12 mutual fund holdings data. Analyst coverage is defined as the total number of analysts covering a firm, and is equal to the number of quarterly earnings forecasts made by unique analysts.

Table 1 reports summary statistics for the institutional holdings data, mutual fund holdings data, and analyst coverage data. The average $FNIO$ has increased over time, from approximately 3.6% in the 1980s to approximately 4.6% in the 2010s. Likewise, the average $FSIO$ has increased over time, from approximately 16% in the 1980s to approximately 58% in the 2010s. These summary statistics are consistent with other many studies (e.g., Hong et al., 2000; Gompers and Metrick, 2001; Edelen et al., 2022).

4 Missing Data in Compustat

Compustat provides financial statement information for tens of thousands of firms. However, Compustat's data coverage is not comprehensive. Panel A of Figure 1 reports the fraction of publicly listed firms (with data available in CRSP) with no annual (solid-red line) and no quarterly (dotted-blue line) data coverage in the Compustat North America database. Results show that when Standard & Poor's began building Compustat in the early 1960s, they did not collect annual accounting data for over 45% of public firms, and that Compustat's data coverage subsequently gradually expanded over time. Regarding this evolution in Compustat's data coverage, S&P Global client support stated the following: When Standard

[14] Following Lewellen (2011), observations where the fraction of shares outstanding held by institutions exceeds 100% are truncated at 100%. This occurs in approximately 2.5% of cases.

& Poor's began collecting data in 1962, they started with the S&P 425 industrial firms. In 1967, S&P expanded Compustat's coverage to include 900 NYSE firms. In 1973, they expanded Compustat's coverage again to include (almost) all NYSE and AMEX firms. Finally, in 1978, they expanded Compustat's coverage to include 3,000 OTC firms, approximately 1,600 of which were listed on the NASDAQ, and added around five years of annual financial statement data for the new additions.

These expansions to Compustat's data coverage explain the decreasing fraction of firms with no annual data in the 1960s and 1970s: in 1962, Compustat has financial statement information available for nearly 70% of NYSE firms and approximately 30% of AMEX firms. This corresponds to just over 1,000 companies. (Financial statement information covering fiscal-years pre-1967 for non-S&P 425 industrial firms likely reflects back-filled data; however, without Point-in-Time information covering this period, this is difficult to confirm.) There is a subsequent decrease in the fraction of firms with no annual data across the 1960s and early 1970s as coverage of NYSE firms becomes close to comprehensive, and coverage of AMEX firms reaches over 80%. There is then a discrete, precipitous increase in the fraction of firms with no annual data coverage in 1972. This corresponds to the addition of the NASDAQ stock exchange, which began operations in February 1971, and was added to the CRSP database in December 1972. At that time, NASDAQ firms had no financial statement information available in Compustat. S&P subsequently expanded their annual data coverage to include approximately 80% of NASDAQ firms in 1978, with information back-filled to 1973 for many of the new additions. By 1980, Compustat provides annual accounting information for 80-85% of all public firms.

The dotted-blue line in Panel A of Figure 1 indicates that nearly 75% of firms did not have quarterly data available in Compustat in the early 1960s. This data subsequently becomes available throughout the 1960s. I am not able to identify precisely when Standard & Poor's began collecting quarterly 10-Q information. However, in discussions with S&P Global client support, they stated that the quarterly data file was expanded to incorporate 40

data items and 20 quarters in 1973. Thus, Compustat definitively began to include quarterly financial statement information no later than 1973, with quarterly information back-filled to at least 1968 for many (primarily NYSE) firms. In December 1972, when NASDAQ firms enter the CRSP database, the fraction of firms with no quarterly Compustat information increases to around 50%, and remains in excess of 45% from 1973 until early 1981. This is because Standard & Poor's did not collect 10-Q information for NASDAQ firms until the early 1980s. In 1983, S&P expanded their data coverage to include 10-Q data for the majority of NASDAQ firms. This expansion involved adding 1-3 years of quarterly data for many previously uncovered NASDAQ firms, which accounts for the gradual change in quarterly data availability over the period 1981-1983.

The fraction of firms with no annual or quarterly Compustat data coverage remains around 15-20% from 1983 through the early 1990s. These uncovered firms correspond primarily to firms in financial services industries, most of which are also listed on the NASDAQ exchange. There is discrete reduction in the fraction of firms with no data coverage in 1994. This corresponds to a project, initiated at Standard & Poor's, to expand data coverage to include all financial services firms. I discuss this change in data coverage in more detail in Section 6.1. Following the reduction in missing data in the mid-1990s, there is a subsequent small increase in the fraction of firms with no data coverage in the late-1990s, which tapers off again by the early 2000's. These uncovered firms are newly-listed financial services firms, whose accounting data does not become available in Compustat until approximately 1-3 years after their intial public offerings.[15] By the early 2000's, Compustat's data coverage is approximately comprehensive.

Panel A of Figure 1 focuses specifically on whether or not a firm has any data available in Compustat. Panel B provides an alternative perspective, and reports the fraction of firm-month observations over time with missing values of a variety of Compustat input variables

[15]In a conversation with S&P Global Client Support in 2023, they indicated that the limited coverage of newly listed financial firms in the mid-late 1990s was due to S&P's limited processing capacity while they pursued massive growth in data coverage for both North American and International firms during this time.

that are used to construct popular accounting-based firm characteristics. Financial statement variables such as total assets ('at'), net income ('ni'), and other bottom-line items from the balance sheet and income statement are typically only missing if Compustat does not cover a firm in a given period. For this reason, the fraction of firms with no data coverage, reported in Panel A, is approximately equal to the fraction of firms with missing values of 'at', 'ni', and similar bottom-line financial statement items.

More nuanced accounting items can be missing for a variety of alternative reasons, including both Compustat's data collection and processing procedures, as well as the underlying structure of firms' financial statements. As a concrete example, annual SG&A expense ('xsga'), which is often used to construct measures of operating profitability (Fama and French, 2015), is missing for approximately 15% of public firms between 2010 and 2020, even though nearly 100% of public firms are covered in Compustat during this time. Likewise, annual deferred taxes and investment tax credits ('txditc'), which is often used to construct book equity (Fama, 1991), is consistently missing for 15-20% of public firms between 2010 and 2020. In a companion paper, Easterwood (2024) provides a more detailed discussion of when and why individual financial statement items are missing in Compustat.

Because Compustat data coverage is largely a function of index membership, exchange listing, industry membership, and time since IPO, missing Compustat data is negatively correlated with firm size – for example, NASDAQ firms and young firms tend to be small and are less likely to have Compustat data available. However, while small firms are more likely to have missing Compustat data relative to large firms, there is not an overly strong association between firm size and missing data. Figure 2 reports the fraction of firms in each size quintile with no Compustat data coverage over time. Results indicate that, while the largest size quintile consistently has the most comprehensive data coverage, quintiles 1–4 consistently have a similar and non-trivial fraction of firms with no data coverage.

5 Missing Data and Investor Demand

If many investors rely on Compustat to obtain firms' financial statement information, then those investors are implicitly relying on the completeness of Compustat's data coverage. I establish in Section 4 that Compustat's data coverage is, historically, not comprehensive. In this Section, I evaluate the connection between missing data in Compustat and institutional investors' equity holdings. I begin by estimating the following regression:

$$IO_{i,q} = a + b \text{ Missing Data}_{i,q} + cX_{i,q} + FE_q + FE_{SIC2} + FE_{exch} + \epsilon_{i,q} \quad (1)$$

where $IO_{i,q}$ is firm i's level of institutional ownership in quarter q. Missing Data$_{i,q}$ is an indicator variable defined as 1 if firm i has no data available in Compustat for the fiscal year-end immediately prior to quarter q. $X_{i,q}$ is a vector of additional firm characteristics (including firm size, age, and index membership), FE_q is a time fixed effect, FE_{SIC2} is an industry fixed effect, FE_{exch} is an exchange-listing fixed effect, a is the intercept, and $\epsilon_{i,q}$ is the error term. Table 2 reports regression results.

5.1 Firms with No Compustat Data Coverage

I hypothesize that a significant fraction institutions rely on Compustat to access firms' financial statement information, and that those institutions will not invest in firms with missing Compustat data. This implies that missing data in Compustat should affect the extensive margin, or an institution's decision of *whether* to invest, and that $FNIO$ is the most relevant measure of institutional ownership. Consistent with this hypothesis, results in column 1 in Panel A of Table 2 indicate that the average fraction of institutional owners is 3.3% lower for firms with no Compustat data, relative to firms that are covered in the database. Given that the unconditional mean $FNIO$ is 4%, this is an economically very large effect; however, this regression fails to control for other firm characteristics known to predict variation in institutional demand (e.g., Falkenstein, 1996; Gompers and Metrick,

2001; Edelen et al., 2022).

Multivariate regression results in Panel A, column 2 indicate that, upon controlling for important firm characteristics such as size, age, index inclusion, and a variety of relevant fixed effects, the fraction of institutional owners is 1.5% lower on average for firms with no Compustat data, relative to covered firms. This continues to be an economically large effect: institutional investment in firms with no Compustat coverage is over 36% below its mean. This effect is not driven by micro-cap stocks. I find very similar results in Panel A, column 7, which excludes the smallest 20% of firms. Results are also similar for the Poisson pseudo-likelihood regressions reported in columns 4–6. The Poisson model is an important robustness check because it better accounts for the fractional nature and (approximately) log-normal distribution of the institutional holdings data.

Insofar as the quantity of shares held by institutions is a function of the number of institutional owners, $FSIO$ should also be correlated with missing Compustat data. However, conditional on an institution investing in a firm, it is not clear whether missing data should impact the intensive margin, or an institution's decision of how much to invest. Panel B in Table 2 reports regression results where institutional ownership is defined as $FSIO$. Results in column 2 indicate that when a firm is not covered in the database, the fraction of shares outstanding held by institutions is approximately 5.7% lower (>16% below the unconditional mean) relative to firms with data coverage.

13f institutions are far from a homogeneous group of investors. Different types of institutions are governed by different regulatory standards and face different investment objectives and mandates. In Panels C and D in Table 2, I report regression results for institutional ownership measures constructed based on the legal type of institution and institutions' size, measured via total equity under management. All missing data coefficient estimates in Panels C and D are significantly negative and economically large in magnitude. Firms with no Compustat coverage have 2.2% (>26% below the unconditional mean) lower bank ownership, 1.3% (>20% below the unconditional mean) lower mutual fund company ownership,

and 0.9% (>35% below the unconditional mean) lower investment company ownership relative to firms with data coverage. Results in Panel D show that the negative association between institutional demand and missing data is largest in magnitude for the largest institutions. This effect is likely mechanical: the larger the institution, on average, the more firms they will invest in and therefore the more binding their aversion is to missing data. Results are also robust to inflation-adjusting the institution size reporting threshold.

Missing data in Compustat is most relevant at the institution level because database subscriptions are likely purchased by the institution and made available to all fund managers within the institution. Thus, all fund managers within an institution should face similar data constraints: either the institution subscribes to Compustat or it does not, and either the individual funds are limited by Compustat's data coverage or they are not. For this reason, in the majority of empirical analyses, I focus on investor demand at the institution (fund family) level. However, as a robustness check, I also consider demand at the individual fund level. In these cases, I use the mutual fund holdings data from the Thomson Reuters s12 file. Results in Panel E in Table 2 focus on the association between individual mutual funds' portfolio holdings and missing data. I find that firms with missing Compustat data have approximately 0.33% lower mutual fund ownership (>28% below the unconditional mean).

Analysts are themselves information intermediaries, and analyst reports are used to inform and advise investors (Healy and Palepu, 2001). It is plausible that analysts utilize firms' past financial statement information when developing earnings forecasts and investment recommendations. Similar to institutional investors, if many of the brokerage firms employing analysts and delegating analyst coverage rely primarily on Compustat to obtain this information, then these firms are also implicitly relying on the completeness of Compustat's data coverage. I evaluate this possibility in Panel F of Table 2. In this case, I define the left-hand-side variable from eq. (1) as Analyst Coverage$_{i,q}$. Consistent with this hypothesis, results indicate that firms with missing Compustat information are covered by approximately 2-3 fewer analysts (>44% below the unconditional mean).

5.2 Extensions and Robustness Checks

Results in Table 2 indicate that institutional ownership is over 36% below the unconditional mean for firms with no Compustat data coverage. This suggests that whether a firm has any data available in Compustat is a critical determinant of institutional demand. Panel B in Figure 1 illustrates that there is considerable heterogeneity in missing data for many potentially important financial statement variables. For example, Selling, General, and Administrative Expenses ('xsga'), Interest Expenses ('xint'), and Deferred Taxes ('txdb') are each missing for around 10-20% of firm-fiscal-years *with* Compustat coverage.

In the Online Appendix, I evaluate the relative importance of missing values for different types of financial statement items. I find that coefficient estimates for Missing Data indicator variables that reflect whether a firm has any Compustat coverage (e.g., missing values of total assets or total sales revenue) are significantly negative and similar in magnitude to estimates reported in Table 2. In contrast, institutional demand is not consistently related to missing values of other, more nuanced accounting items such as capital expenditures, total inventory, and SG&A expenses. This suggests that the investors relying on Compustat to obtain firms' financial statement information may either: 1) focus primarily on a subset of financial statement items which are both non-missing for nearly all covered firms and broadly representative of firms' overall operations and performance, or 2) identify and adjust missing values of granular accounting data by, e.g., assuming a missing value is equal to zero.

I consider many additional robustness checks with respect to the results reported in Table 2. These include Fama MacBeth regression specifications, alternative definitions of 'Missing Data' including changes in data coverage, and alternative combinations of control variables. Selected results appear in the Online Appendix.

6 Identification

The previous Section establishes that institutional ownership is lower on average for firms that are not covered in Compustat. While these results are suggestive, significant endogeneity concerns limit their interpretation. First and foremost, data vendors have an incentive to cater their data coverage to client demand, and S&P readily states that client requests influence data coverage decisions. Thus, Compustat is more likely to contain a firm's financial statement data when investor demand for that information is high. This reverse causality concern is similar to themes highlighted in D'Souza et al. (2010), who emphasize the interdependence between Compustat's information dissemination speeds and institutional ownership. In addition, firms' underlying salience likely influences both institutional demand and Compustat's data coverage. If this salience is time-varying, even the regression specifications including firm fixed effects in Table 2 cannot fully address this correlated omitted variable issue.

I introduce a novel, quasi-natural experiment to evaluate whether there is a plausibly causal connection between Compustat data coverage and institutional investor demand. Section 6.1 describes the empirical setting. Section 6.2 presents results for the associated difference-in-differences analysis. Section 6.3 describes various robustness exercises.

6.1 Background

Standard & Poor's has always maintained two internal data collection systems: one for firms in financial services industries, and one for firms in all other industries. S&P regularly refers to these two categories of firms as 'banks' and 'industrials', respectively.[16] S&P maintains a separate system for the financial firms because the disclosures for these firms are structured very differently from industrial firms in other industries. S&P then uses a 'balancing model' to convert financial firms' accounting data into the industrial firm format.

[16]To be clear, S&P's 'bank' category refers to firms with SIC codes in the 6000s, which includes financial services firms such as insurance companies and brokerage firms. Thus, it is *not* accurate to interpret this designation as referring to only commercial and investment banks, or to only bank holding companies.

The financial firm data collection system lagged behind the industrial firm data collection system for several decades. As a result, until the early 1990s, Compustat provided accounting data for only a subset of banks and financial services institutions. In the early 1990s, S&P was able to significantly enhance the financial firm internal data collection system. This positive technological shock enabled them to 1) expand coverage on a going-forward basis to include all (non-newly publicly listed) firms in the financial services industry, 2) back-fill data for a subset of previously uncovered financial firms, and 3) update the balancing model so that many variables which were previously missing for many covered financial firms were no longer missing.

This database expansion led to a discrete, precipitous reduction in the fraction of financial firms with missing data in Compustat, and is not related to any changes in those firms' 10-K or 10-Q disclosures. Figure 3 shows the fraction of financial versus non-financial firms with missing values of a variety of popular Compustat variables over time. In order to focus on the effects of the technology shock and avoid variation in data coverage related to newly listed firms in the mid-late 1990s, all firms in Figure 3 are required to be publicly listed in or before Q1 1988. Panel A focuses on data availability in the Compustat North America database. There is a clear, abrupt reduction in missing data in 1994 for financial firms: approximately 45% of financial services firms do not have any data available in the Compustat North America database prior to the 1993 fiscal year-end. In contrast, the fraction of non-financial firms with missing Compustat data is consistently very close to 0 for all of the period 1988-1999.

Panel B of Figure 3 also shows the fraction of financial versus non-financial firms with missing data over time, in this case focusing on data availability in the Compustat Point-In-Time database. The PIT database states *when* data became available in the Compustat North America database. This is an important robustness check because, during the database expansion in the early 1990s, Standard & Poor's back-filled financial data for a subset of financial services firms. Thus, from a backward-looking perspective, the Compustat North

America database overstates the amount of information that was available to investors in real-time. Results in Panel B of Figure 3 show that nearly 65% of financial services firms (>800 firms) had no data available prior to 1993. Likewise, approximately 80% of financial firms did not have data available for common financial statement items, including accounts payable ('ap'), cost of goods sold ('cogs'), and total inventory ('invt'). The gradual darkening of the 'Financial Firms' figure in Panel B over the period 1990–1994, with the largest change occurring in 1993 and 1994, reflects 1) the release of back-filled information for some of the previously uncovered firms, 2) the attribution of 1993 fiscal year-end information for all financial firms, and 3) updates to the balancing model.

Panel B of Figure 3 illustrates that current financial statement information was not comprehensively available for all financial services firms until around the end of 1994. This delay occurs for two related reasons. First, the majority of firms have December 31 fiscal year-ends, which means that accounting data from the 1993 fiscal year is not filed with the SEC until early 1994. This change in data coverage focused primarily on financial firms' financial statements from the 1993 *fiscal year* onward. As such, it is sensible that accounting data does not become available for most firms in the database until 1994. Second, because the Compustat database was undergoing a significant expansion during this time, there was a significant production lag in the attribution of financial statement information from the 1993 fiscal year into the database. This meant that, for some firms, data from the 1993 fiscal year-end did not become available in Compustat until the third or fourth quarter of 1994. From 1995 onward, nearly 100% of the financial and non-financial firms (excluding new listings) are covered in Compustat and have non-missing values of basic financial statement items such as income, assets, and stockholder's equity.

6.2 Difference-in-Differences Analysis

The increase in Compustat's coverage of financial services firms in 1993-1994 was driven by a positive shock to S&P's data collection technologies. It therefore provides a setting to

evaluate a plausibly causal connection between Compustat's data coverage and institutional investor demand. If institutional investors rely on Compustat to access firms' financial statement information, then, following the positive shock to coverage, ownership should increase for the subset of affected firms.

I use a difference-in-differences approach to test the impact of the technology shock on institutional investor demand. There are a number of important methodological details to consider regarding the definition of appropriate treatment and control samples. First, the technology shock relates specifically to Standard & Poor's financial firm data collection system. Therefore, the treatment sample is composed only of financial services firms. To facilitate comparability, the control sample is also composed only of financial services firms. I follow Standard & Poor's definition of financial services, and classify all firms with SIC codes ranging from 6000–6999, excluding codes 6411, 6792, 6794, and 6795, as financials. To avoid confounding results with issues related to changes in data coverage for new listings, I require all candidate treatment and control firms to be publicly listed in or before Q1 1988.

Second, the technology shock led to three related changes in data availability in Compustat: firms that were previously not covered began to have data coverage on a going-forward basis; accounting information was back-filled for a subset of firms that were previously not covered; and the number of available accounting items increased for a subset of firms that previously had limited data coverage. My aim is to evaluate if a firm's inclusion in Compustat drives variation in institutional investment. Thus, I define the treatment sample as financial services firms with no data available in the Compustat Point-in-Time database prior to 1993. All of these firms then begin to be covered in Compustat between 1993 and 1994. Among financial firms that have at least some data available in the PIT database prior to 1993 and that were publicly listed in or before 1988, there are firms that 1) have complete data coverage for the period 1988-1992, 2) have some accounting information available over the period 1988-1992 and have more information added following the shock in 1993-1994, and 3) have no data available for part of the period 1988-1992, and then subsequently have their

data added to Compustat sometime before 1993. I further restrict the sample of candidate control firms to the first group: financial services firms with complete data coverage for the period 1988-1992. Because firms in the second and third samples experience different but related changes in data coverage compared to the treatment group, they are not suitable controls.[17]

Finally, to ensure that the treatment and control samples are similar in all aspects except their Compustat data coverage, I match firms in the control group to firms in the treatment group based on observable firm characteristics. None of the treated firms are members of the S&P500 index. For this reason, I drop all S&P500 firms from the control sample. I then use K-nearest neighbor ('KNN,' K=1 with replacement) to match control firms to treated firms based on size (log market cap), institutional ownership (FNIO), and trading volume (log turnover), all measured in Q1 1988. Because the treatment group is composed of firms with no Compustat data coverage prior to 1993, I cannot observe accounting-based characteristics for these firms in 1988. However, I confirm that characteristics such as profitability, asset growth, dividend yield, and book-to-market ratios, all measured using Compustat data from the 1993 and 1994 fiscal years, are equal on average for the two samples.

I estimate the following regression to test the impact of the technology shock on institutional investment:

$$IO_{i,q} = a + b(\text{Treated}_i \times \text{Post}_q) + c\ \text{Treated}_i + dX_{i,q} + FE_q + FE_{SIC2} + FE_{exch} + \epsilon_{i,q} \quad (2)$$

where $IO_{i,q}$ is firm i's level of institutional ownership in quarter q. Treated is an indicator that equals one for firms in the treatment group, and zero otherwise. Because the treatment affect is dispersed across 1993 and 1994, and is not complete until the final quarter of 1994, Post is defined as an indicator that equals one in and after 1995, and zero otherwise. The

[17]The data coverage requirements for the control group implicitly require these firms to be publicly listed (and have data available in CRSP and Compustat) throughout the period 1988-1992. To ensure that the treatment and control samples are similar in all aspects except their Compustat data coverage, I also require firms in the treatment group to be publicly listed (and have data available in CRSP) throughout 1988-1992.

sample spans the period Q1 1988 – Q4 1999.

Regression results are reported in Table 3. The b coefficient estimate for $\text{Treated}_i \times \text{Post}_q$ is consistently significant and positive across all specifications. Results in Panel A, column 2 indicate that, following the change in data coverage, institutional ownership increased by around 0.76% for treated firms relative to control firms. Notably, results are robust to time, industry, exchange, and firm fixed effects, to the inclusion of time-varying controls, to alternative functional forms, and to the exclusion of micro-cap stocks.

Panel A of Table 3 reports results for aggregate institutional ownership. Panel B reports results for legal types of institutions. Panel C reports results for various institution sizes. Panels D and E report results for individual mutual fund holdings and for analyst coverage, respectively. Consistent with the discussion in Section 5.1, ownership levels increase for all types and sizes of institutions following the initiation of Compustat coverage. Results are also similar for analyst coverage and alternative measures of mutual fund ownership.

Figure 4 evaluates the parallel trends requirement. Panel A reports dynamic treatment effects over time, and illustrates that the interaction coefficient estimate from regression eq. (2) becomes large, positive, and statistically significant following the positive shock to Compustat coverage. Panel B reports the cross-sectional average institutional ownership for treated (dotted-blue line) versus control (solid orange line) firms from Q1 1988 through Q4 1999. The left-hand figure reports the average FNIO, and the right-hand figure reports the average NIO (i.e., the average unscaled number of institutional owners). The grey, diagonal slash shaded region indicates the period over which treated firms' financial statement data began to appear in Compustat.

Results in Panel B in Figure 4 show that treated and control firms have equal average levels of institutional investment from 1988–1992. Then, in and after 1993, the average level of ownership for treated firms begins to increase relative to the average level of ownership for control firms. Across the period 1993–1999, treated firms' average ownership increases from 10 to over 33 institutions while controls firms' average ownership increases from 10 to

approximately 18 institutions. Thus, the increase in average ownership for treated firms is over 2.8 times larger than the increase in average ownership for control firms. The gradual divergence of treated firms' average level of institutional investment relative to control firms' average level of investment is exactly what is expected under the hypothesis that institutions avoid investing in firms with missing data – if many institutional investors rely on Compustat to obtain firms' financial statement information, then these institutions will not invest in firms with missing Compustat data. Once a firm begins to be covered in Compustat, these institutions *might* invest in the newly covered firm; however, the decision of whether or not to invest likely depends on many other factors, such as the values of various financial ratios.

Altogether, results in Figure 4 and Table 3 are consistent with the following causal interpretation: a meaningful fraction of institutional investors rely on Compustat to access firms' accounting information. Thus, when a firm is not covered in Compustat, those institutions do not invest in the firm. Once the firm begins to be covered, institutional investment increases.

6.3 Robustness Checks and Alternative Explanations

I consider a variety of robustness checks with respect to the empirical results presented in Section 6.2, including several placebo tests where the treatment effect is randomly assigned across firms. Explicit results appear in the Online Appendix, as do additional summary statistics for the treated and control samples.

A primary concern with the difference-in-differences analysis is that the technology shock coincides with other firm-level shocks that may affect investor actions. Importantly, however, in order to provide a plausible alternative explanation, such shocks must effect the subset of treated firms differently than the subset of control firms. At least two notable events coincide with the Compustat technology shock: the implementation of the SEC's EDGAR system, which took place over the period 1993–1996, and the 1994 passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act. I explore these events, and their relation

to the regression results in Section 6.2, in the following subsections.

6.3.1 EDGAR

The SEC's EDGAR database was implemented on a staggered schedule between 1993–1996. Studies such as Gao and Huang (2020), Kim et al. (2024), and Hirshleifer and Ma (2024) highlight the role that EDGAR played in massively reducing information acquisition costs and improving access to firms' public disclosures, including their accounting data. It is possible that many institutional investors began using EDGAR as soon as it became available. However, the introduction of EDGAR alone is an unlikely explanation for the differential trends in institutional investment for the treated versus control firms reported in Figure 4 and Table 3. Specifically, to provide a plausible alternative explanation, EDGAR's implementation would need to differentially affect disclosure access for the treatment versus control samples. *All* public firms were required to begin filing their public disclosures on EDGAR between April 1993 and May 1996.[18] Thus, 10-K and 10-Q disclosures for all firms in both the treatment and control groups became available on EDGAR during this time.

Empirically, I confirm that phase-in schedules across the treatment and control groups are similar. Figure 5 reports the fraction of firms in the treatment and control samples that were EGDAR filers over the period 1992–1996.[19] Results show that the majority of treated and control firms began to file electronically in 1994 or 1995. The control firms did begin to file on EDGAR approximately one year earlier on average compared to the treatment sample. Ex ante, it is unlikely that such timing differences led to material differences in institutional demand for the treated versus control groups. However, I consider several robustness checks to evaluate whether differences in the timing of firms' launch on EDGAR influences the differences-in-differences analysis in Section 6.2.

[18]The SEC allocated each public firm to one of ten groups, which determined precisely when over the three year period the firm was required to commence electronic filing. Gao and Huang (2020) state that they inquired with the SEC to investigate how companies were assigned to different groups, and that the SEC did not locate any information related to this request (see Footnote 9 in Gao and Huang, 2020).

[19]I include only those firms that have an exact CIK or name match in SEC Release No. 33-6977–Appendix B, which identifies when each firm was required to begin filing on EDGAR. This corresponds to over 80% of the sample of firms. Results are similar if I also include firms with approximate name matches.

In Panel A of Table 4, I examine whether there is any variation in the treatment effect across firms that began filing on EDGAR in each year between 1993 and 1996. I first estimate the regression in eq. (2), where the sample is restricted to firms that began filing on EDGAR in 1993 (columns 1 and 2), 1994 (columns 3 and 4), or 1995/1996 (columns 5 and 6). (I combine the 1995 and 1996 samples because only one control firm began filing in 1996.) In columns 7 and 8, I consider the following triple-differences regression analysis:

$$\begin{aligned} IO_{i,q} = a + b\left(\text{Treated}_i \times \text{Post}_q\right) + c\left(\text{EDGAR(93)}_i \times \text{Treated}_i \times \text{Post}_q\right) + \\ d\left(\text{EDGAR(94)}_i \times \text{Treated}_i \times \text{Post}_q\right) + e\left(\text{EDGAR(93)}_i \times \text{Post}_q\right) + \\ f\left(\text{EDGAR(94)}_i \times \text{Post}_q\right) + gX_{i,q} + FE_q + FE_i + \epsilon_{i,q} \quad (3) \end{aligned}$$

where EDGAR(93)_i is an indicator variable equal to one for firms that began filing on EDGAR in 1993 and zero otherwise. EDGAR(94)_i is defined similarly.

I find no evidence that variation the timing of firms' launch on EDGAR is related to institutional demand. Results in Panel A of Table 4 show that the increase in institutional ownership for treated firms relative to control firms is significantly positive and consistent in magnitude for firms the began filing on EDGAR in each of 1993, 1994, and 1995/1996. Likewise, all of the triple-interaction terms (EDGAR×Treated×Post) and EDGAR×Post interaction terms in columns 7 and 8 are statistically indistinguishable from zero. This is inconsistent with the notion that EDGAR's launch provides a plausible alternative explanation for the differential increase in institutional investment for the treated versus control firms in the primary difference-in-differences analysis.

6.3.2 Riegle-Neal Interstate Banking and Branching Efficiency Act

The Riegle-Neal Interstate Banking and Branching Efficiency Act was passed in 1994. This act removed the restrictions which previously prevented banks from engaging in interstate banking and from branching across state lines. Literature examining the impact of this regulation has largely concluded that it increased the competitiveness of U.S. banking markets

(Zarutskie, 2006; Rice and Strahan, 2010). While it is possible that this affected institutional demand for banks in the mid-late 1990s, in the difference-in-differences analysis, both the treatment and control groups are composed exclusively of financial services firms. This reduces concerns that Riegle-Neal provides a plausible alternative explanation.

The Riegle-Neal Act did not affect all banks equally because states maintained the authority to create barriers to branch expansion. Specifically, states could limit interstate branching in any of the following four ways: First, states could limit interstate bank mergers by setting a minimum age requirement for all target institutions. Second, states could cap the percentage of deposits controlled by any single bank or bank holding company, thus limiting banks' ability to engage in large interstate mergers. Third, de novo interstate branching was only permitted if states decided to "opt-in" to this feature of the regulation. Finally, interstate mergers of individual branches was also only permitted if states decided to "opt-in" to this feature of the regulation. Collectively, this means that interstate branching was only possible via whole-bank mergers which met minimum age requirements and did not exceed the relevant deposit cap for states that elected not to opt-in to these provisions. Rice and Strahan (2010) exploit variation in states' adoption of these different barriers to entry to create a state-level index of branching restrictiveness.

Under the hypothesis that the Riegle-Neal Act explains the differential increase in institutional investment for treated firms relative to control firms in the 1990s, the increase in institutional ownership should be largest for firms located in states with the most open branching laws post-Riegle-Neal. This is because banks in more open states were more affected by Riegle-Neal than banks in less open states. In Panel B of Table 4, I examine whether there is any variation in the treatment effect across firms located in different states. I first estimate the regression in eq. (2), where the sample is restricted to firms with very open branching laws post-Riegle-Neal (columns 1 and 2, corresponding to firms located in states with a Rice and Strahan (2010) Branching Restrictiveness Index $\leqslant 2$) or very restrictive branching laws post-Riegle-Neal (columns 3 and 4, corresponding to firms located in

states with a Rice and Strahan (2010) Branching Restrictiveness Index $\geqslant 3$). In columns 5 and 6, I consider the following triple-differences regression analysis:

$$IO_{i,q} = a + b(\text{Treated}_i \times \text{Post}_q) + c(\text{High}_i \times \text{Treated}_i \times \text{Post}_q) + d(\text{High}_i \times \text{Post}_q) + eX_{i,q} + FE_q + FE_i + \epsilon_{i,q} \quad (4)$$

where High_i is an indicator variable equal to one for firms located in states with restrictive branching laws post-Riegle-Neal (i.e., firms located in states with a Rice and Strahan (2010) Branching Restrictiveness Index $\geqslant 3$) and zero otherwise.

I find no evidence that variation in branching restrictions is related to institutional demand. Results in Panel B of Table 4 show that the increase in institutional ownership for treated firms relative to control firms is significantly positive and consistent in magnitude for firms located in both very open and very restrictive states. Likewise, the triple-interaction term in columns 5 and 6 (High×Treated×Post) is statistically indistinguishable from zero. This is inconsistent with the notion that Riegle-Neal provides a plausible alternative explanation for the differential increase in institutional investment for the treated versus control firms in the primary difference-in-difference analysis.

Riegle-Neal facilitated many interstate mergers in the mid-late 1990s. Ex ante, it is possible that mechanical changes in institutional investment resulting from these mergers drives the differential trends in institutional investment for the treated versus control firms reported in Figure 4 and Table 3. I evaluate this possibility in Panel C of Table 4. In this case, I examine whether there is any variation in the treatment effect across firms which engaged in a merger or acquisition at any point over the period 1988-1999 versus those that did not. I first estimate the regression in eq. (2), where the sample is restricted to non-M&A firms (columns 1 and 2, corresponding to firms that did not engage in a merger or acquisition at any point over the period 1988-1999) or M&A firms (columns 3 and 4, corresponding to firms that did engage in at least one merger or acquisition over the period 1988-1999).[20] In

[20]I obtain data on mergers and acquisitions from SDC's U.S. Mergers and Acquisitions Database. Follow-

columns 5 and 6, I consider the following triple-differences regression analysis:

$$IO_{i,q} = a + b(\text{Treated}_i \times \text{Post}_q) + c(\text{M\&A}_i \times \text{Treated}_i \times \text{Post}_q) + d(\text{M\&A}_i \times \text{Post}_q) + eX_{i,q} + FE_q + FE_i + \epsilon_{i,q} \quad (5)$$

where $\text{M\&A}_i$ is an indicator variable equal to one for firms that engaged in at least one merger or acquisition (as target or acquirer) over the period 1988-1999 and zero otherwise.

I find no evidence that mechanical changes in institutional investment resulting from mergers or acquisitions drives the differential increase in institutional investment for the treated versus control firms. Results in Panel C of Table 4 show that the increase in institutional ownership for treated firms relative to control firms is significantly positive and consistent in magnitude for both the M&A and non-M&A samples. Likewise, the triple-interaction term in columns 5 and 6 (M&A×Treated×Post) is statistically indistinguishable from zero. This is inconsistent with the notion that the mergers facilitated by Riegle-Neal provide a plausible alternative explanation for the differential increase in institutional investment for the treated versus control firms in the primary difference-in-difference analysis.

7 Who Uses Compustat?

The empirical analyses in Sections 5 and 6 suggest that a significant fraction of institutions do not invest in firms with missing data in Compustat. This confirms Compustat's relevance as an information intermediary, and emphasizes the notion that data vendors' data coverage decisions can have a meaningful impact on investor actions. It also raises a number of questions: Which institutions use Compustat? Given that it is possible to supplement Compustat with self-collected data, why do many institutional investors appear not to do not do so? In this Section, I evaluate several potential explanations for why this is the case.

ing Netter et al. (2011), the sample of mergers and acquisitions is defined as completed deals which involve a domestic (U.S.) target or acquirer, where the acquirer obtains an ownership stake in the target of at least 50%.

7.1 Hypothesis Development

First and foremost, Compustat provides access to financial statement data. Thus, it is reasonable to conclude that those institutions that utilize Compustat are using accounting information when making investment decisions. There are several potential ways in which this data may be useful. Trading strategies may incorporate accounting information, investment mandates may include restrictions based on financial ratios, and financial statement analysis may help portfolio managers illustrate that they have done their due diligence as fiduciaries, and/or identify variation in risk exposure or mispricing.

Collectively, this suggests that several empirical patterns should emerge in the data. First, passive index funds should be more inclined to invest in firms with no Compustat data relative to actively managed funds, because a passive index fund's strategy depends only on index constituents and not on accounting information. Second, the very highest turnover funds, which are the most likely to engage in strategies focusing on high frequency information (e.g., price, returns, and trading volume), should be more inclined to invest in firms with no Compustat data relative to lower- and mid-turnover funds, which are more likely to engage in strategies that utilize accounting data. Similarly, funds whose trades are most (least) correlated with momentum and contrarian strategies should be more (less) inclined to invest in firms with no Compustat data. Third, institutions most (least) constrained by agency conflicts and prudent-man regulations should less (more) inclined to invest in firms with no Compustat data because these institutions are more (less) likely to utilize financial statement analysis to illustrate that they have done their due diligence as fiduciaries.

There are many alternative resources from which investors can obtain firms' 10-K and 10-Q disclosures, and any institution that uses Compustat could supplement Compustat data with self-collected accounting information. However, lower average investment in uncovered firms suggests that many institutions do not do so. From an equilibrium perspective, it is only optimal for portfolio managers to self-collect data if the marginal benefits of obtaining

the information are greater than the marginal costs associated with collecting that information (Grossman and Stiglitz, 1980). It is possible that, for many institutions, the costs associated with acquiring firms' disclosures and maintaining a database of information for these uncovered firms exceed the benefits they might accrue from obtaining and trading on the additional accounting data. Studies such as Gao and Huang (2020), Bowles et al. (2024), and Kim et al. (2024) highlight many issues associated with obtaining 10-K and 10-Q disclosures from the SEC, and emphasize that this is *not* a cost-less endeavor. These 'costs' can include the time and resources spent to acquire the information, develop a standardization method, build and maintain a database, and ensure that the use of the data satisfies any relevant due diligence requirements.

Collectively, this suggests that several similar empirical patterns should emerge in the data. First, institutions more (less) constrained by agency conflicts and prudent-man regulations should less (more) inclined to self-collect data because they face a higher (lower) burden of due diligence. Second, more (less) skilled portfolio managers, who are better (less) able to identify variation in risk exposure or missing pricing, should be more (less) inclined to self-collect data because the potential marginal benefit of obtaining that data is higher (lower) for these investors. Finally, high fixed costs associated with data collection and economies of scale suggest that larger (smaller) institutions should be more (less) inclined to self-collect data.

7.2 Empirical Analyses

In order to evaluate the hypotheses described in Section 7.1, I examine institutional investors' propensity to invest in firms with missing Compustat data and estimate the following regression at the institution-quarter level:

$$\begin{aligned}\%\ \text{Portfolio No Data}_{j,q} = a + b\ \text{Log(EUM)}_{j,q} + c\ \text{Age}_{j,q} + d\ \text{Turnover}_{j,q} + \\ e\ \text{Past Returns Trader}_{j,q} + \text{Legal Type FE}_j + \text{Time FE}_q + \epsilon_{j,q} \quad (6)\end{aligned}$$

where % Portfolio No Data$_{j,q}$ is the fraction of institution j's portfolio invested in firms with no Compustat data in quarter q. I consider two alternative definitions of this variable: the fraction of the institution's equity under management invested in firms with no Compustat data ('Fraction of EUM') and the fraction of firms that the institution holds with no Compustat data ('Fraction of Firms'). Log(EUM)$_{j,q}$ is the log of institution j's total equity under management, Age is the number of quarters that the institution has appeared in the 13f database, and Turnover is the institution's portfolio turnover defined as in Yan and Zhang (2009). Past Returns Trader reflects the correlation between institution j's trades in the most recent quarter and returns for the firms that the institution traded. High (low) values of Past Returns Trader indicate the institution's most recent trades were highly correlated (uncorrelated) with firm returns. Legal Type FE$_j$ is a set of fixed effects reflecting institutions' legal types.

Throughout the majority of this study, I focus on institution-level data because database subscriptions are likely maintained at the institution (or fund family) level, and therefore the association between Compustat's data coverage and investment is most relevant at the institution level. However, investment mandates, which often govern investment strategies, benchmark indices, and activeness, are typically specified at the fund level. For this reason, I estimate similar regressions using the s12 data regarding individual mutual funds:

$$\begin{aligned}\%\text{ Portfolio No Data}_{f,q} = a + b\ \text{Log(EUM)}_{f,q} + c\ \text{Age}_{f,q} + d\ \text{Turnover}_{f,q} + \\ e\ \text{Past Returns Trader}_{f,q} + f\ \text{Active Share}_{f,q} + g\ \text{Index Fund}_f + \\ h\ \text{Enhanced Index Fund}_f + \text{Time FE}_q + \epsilon_{f,q} \quad (7)\end{aligned}$$

where, in this case, % Portfolio No Data$_{f,q}$ is the fraction of mutual fund f's portfolio invested in firms with no Compustat data in quarter q. Active Share is defined as in Cremers and Petajisto (2009) and Petajisto (2013), and is equal to the percentage of a fund's portfolio holdings that differ from the fund's benchmark index. Data regarding funds' active share

are obtained from Annti Petajisto's website.[21] Table 5 reports regression results. Panel A focuses on institutional investors and regressions as in eq. (6), while Panel B focuses on individual mutual funds and regressions as in eq. (7).

Results indicate that both index funds and enhanced index funds invest a significantly larger fraction of their portfolios in firms with missing Compustat data compared to non-index funds. Likewise, results show that both Portfolio Turnover and Past Returns Trader are significantly positively associated with the fraction of an institution's or mutual fund's portfolio invested in firms with missing data. For example, institutions in the highest turnover quintile have approximately 0.27% more of their portfolio invested firms with no Compustat data relative to institutions in the lowest turnover quintile. This is equivalent to approximately 8% of a standard deviation difference in 'Fraction of Firms.' Collectively, these results are consistent with the conclusion that those institutions that are less likely to use accounting information when implementing their trading strategies are less dependent on Compustat's data coverage.

Studies such as Badrinath et al. (1989), Badrinath et al. (1996), and Del Guercio (1996) emphasize that insurance companies, banks, and pension funds are subject to much stricter prudent-man laws and due diligence constraints compared to investment companies and advisors. Consistent with the notion that these due diligence constraints affect both the likelihood that an institution uses accounting data and the likelihood that an institution will self-collect data, results in Panel A of Table 5 show that insurance companies, banks, and pension funds invest a significantly smaller fraction of their portfolios in firms with missing Compustat data compared to mutual fund and investment companies.

Studies such as Edelen et al. (2022) emphasize that smaller institutions are more constrained by agency conflicts. Likewise, economies of scale and high fixed costs associated with data collection imply that smaller institutions are less likely to find it optimal to self-collect data. Consistent with the notion that 1) agency costs affect the likelihood that a portfolio

[21] https://www.petajisto.net/data.html. This data is available only for domestic, all equity mutual funds, which are not sector funds and which have a minimum of $10 million in assets under management.

manager uses accounting data, and 2) both agency costs and explicit data collection costs affect the likelihood that a portfolio manager will self-collect data, results in Table 5 show that total equity under management is significantly negatively associated with the fraction of an institution's portfolio invested in firms with missing data.

Finally, studies such as Cremers and Petajisto (2009) suggest that measures of fund activeness proxy for measures of portfolio managers' skill, and show that more active managers outperform less active managers on average. This suggests that more (less) active managers should be more (less) inclined to self-collect information and less (more) reliant on Compustat data because the marginal benefit of obtaining and trading on the additional information is higher (lower). Consistent with this hypothesis, results in Panel B of Table 5 show that Active Share is significantly positively related to the fraction of a mutual fund's portfolio invested in firms with missing data. For example, mutual funds in the highest active share quintile (the 'active stock pickers') have 0.97% more of their equity under management invested firms with no Compustat data coverage relative to mutual funds in the lowest active share quintile (the 'closest indexers'). This is equivalent to nearly 50% of a standard deviation difference in 'Fraction of EUM.' Fund activeness may also proxy for a portfolio manager's investment constraints – that is, managers more (less) constrained by agency conflicts may have less (more) freedom to deviate from their benchmark indices. Thus, the positive association between Active Share and a mutual fund's propensity to invest in firms with missing Compustat data is also potentially consistent with the aforementioned hypotheses regarding the role of agency costs.[22]

[22]The hypotheses related to agency costs also suggest that younger institutions with weaker reputations should be more constrained by Compustat's data coverage relative to older, more established institutions. However, the institution and mutual fund age coefficient estimates vary considerably in sign and significance. This inconsistency may be arise because institution and mutual fund age are imperfect indicators of individual portfolio managers' reputational capital.

8 Missing Data and Information Assimilation

I conclude this study by evaluating the economic consequences of lower institutional investment in firms with missing Compustat data. Merton (1987) links investor attention to market efficiency, and suggests that 'neglected' firms that face significantly less scrutiny by many market participants will have less informationally efficient equity prices. Several recent studies provide empirical support for this hypothesis (Boone and White, 2015; Ben-Rephael et al., 2017; Kacperczyk et al., 2021; Chen et al., 2022). This suggests that Compustat's data coverage should affect market efficiency because of its impact on investor attention.

I consider several alternative empirical settings to evaluate the connection between Compustat's data coverage and information assimilation. In each setting, I construct firm-level empirical proxies for stock price informational inefficiency ('II'), and estimate the following regression:

$$\begin{aligned} II_{i,t} = a + b\text{Missing Data}_{i,t-1} + c\left(\text{Missing Data}_{i,t-1} \times \text{Investor Attention}_{i,t-1}\right) \\ + d\ \text{Investor Attention}_{i,t-1} + eX_{i,t-1} + FE_t + FE_{SIC2} + FE_{exch} + \epsilon_{i,t} \quad (8) \end{aligned}$$

where $II_{i,t}$ is the relevant informational inefficiency measure for firm i at time t, $\text{Missing Data}_{i,t-1}$ is an indicator variable defined as 1 if firm i is not covered in Compustat in the $t-1$ fiscal year, and $\text{Investor Attention}_{i,t-1}$ is a proxy for market participation, measured as either institutional ownership or analyst coverage. I hypothesize that missing Compustat data mitigates information assimilation, and that this effect will be partially offset by increased investor attention – that is, that $b > 0$ and $c < 0$ in regression (8).

8.1 Quarterly Earnings Announcements

I begin by evaluating the connection between Compustat data coverage and returns during and after quarterly earnings announcements. I use quarterly earnings announcements as a laboratory from which to study information assimilation because 1) these announce-

ments provide firm-specific, valuation-relevant, fundamental information at definitive points in time, and 2) it is well-documented in the empirical literature that there is a significant price drift following these announcements (e.g., Ball and Brown, 1968; Fink, 2020).

I follow prior literature (Blankespoor et al., 2020; Fink, 2020) and estimate announcement period cumulative abnormal returns ('CARs') over the window $\tau = [-1, 1]$, where $\tau = 0$ is the earnings announcement date. I define post-announcement CARs over the window $\tau = [2, 60]$. In main results, I focus on two alternative models for the normal return: the single-factor market model and the Fama and French (1993) three-factor model.[23] I consider alternative windows and alternative normal return models as robustness checks. Finally, I estimate regressions as in eq. (8), where the dependent variable is defined as the announcement period or post-announcement absolute cumulative abnormal return, $ACAR_{i,\tau}$. Earnings announcements are measured in each quarter of year t, the 'Missing Data' indicator reflects whether a firm has any Compustat data coverage for the $t-1$ fiscal year-end, and all other variables are measured as of the end of the quarter prior to the earnings announcement.

Table 6 reports regression results. Panel A focuses on earnings surprises. Results in columns 1–4 indicate that announcement period returns are 0.3–0.5% larger in magnitude for firms with no Compustat coverage for the most recent fiscal year-end. Results also indicate that this effect is offset if there are sufficiently many analysts covering the firm and/or sufficiently many institutions investing in the firm. Specifically, the interaction coefficient estimates suggest that an increase of approximately 6-8 ($\approx\geqslant$1 standard deviation) analysts or an increase in institutional ownership of 5-15% ($\approx\geqslant$1 standard deviation) negates the impact of missing Compustat data.

Panel B of Table 6 focuses on post-announcement drift. Results in columns 1–4 indicate that post-announcement returns are 0.7–1% larger in magnitude for firms with no Compustat coverage for the most recent fiscal year-end. Results also indicate that this effect is offset if

[23]Under the single-factor market model, I set all market betas equal to 1. This avoids methodological issues associated with estimating betas. Under the Fama and French (1993) three-factor model, I use a 250-trading-day window ending on day $\tau - 2$ to estimate factor loadings.

there are sufficiently many analysts covering the firm and/or sufficiently many institutional investors, however the effect is both statically weaker and smaller in magnitude than that for earnings surprises. Specifically, the interaction coefficient estimates suggest that an increase of at least 7 analysts or an increase in institutional ownership of at least 9% negates the impact of missing Compustat data.

The results in Table 6 support the conclusion that limited access to financial statement data limits the informational efficiency of equity prices: when Compustat does not cover a firm, earnings surprises are larger, post-earnings announcement drift is larger, and information assimilation is slower. These effects are mitigated if investor attention is sufficiently high. This suggests that Compustat data coverage affects information assimilation in financial markets via its impact on market participation and investor attention.

8.2 Return Autocorrelations and Price Delay

French and Roll (1986) argue that the absolute levels of firms' daily return autocorrelations should be positively related to investors' mis-reactions to new, firm-specific information. Thus, a firm's autocorrelation coefficient (ρ) serves as a measure for the informational inefficiency of the firm's stock price: in an informationally efficient market, prices reflect all public information and returns should follow a random walk (i.e., $\rho \approx 0$). For this reason, I evaluate the connection between Compustat's data coverage and firms' return autocorrelations. I first estimate the following AR(1) regression:

$$R_{i,d} = \alpha_i + \rho_i R_{i,d-1} + \epsilon_{i,d} \tag{9}$$

where $R_{i,d}$ is the daily return for firm i on trading day d. I estimate these regressions at the firm-level using one year of daily returns, requiring a minimum of 100 trading days of data.

In addition to measures of daily return autocorrelations, I consider the three alternative measures of 'price delay' proposed by Hou and Moskowitz (2005), which are designed to estimate the delay with which firms' stock prices incorporate market-wide information. These

measures are obtained from the following regressions:

$$R_{i,w} = \alpha_i + \beta_i^0 R_{MKT,w} + \sum_{n=1}^{4} \left(\beta_i^{-n} R_{MKT,w-n}\right) + \epsilon_{i,w} \tag{10}$$

where $R_{i,w}$ is the weekly return for firm i in week w, and $R_{MKT,w}$ is the value-weighted market return in week w. Weekly returns are measured from Wednesday–Tuesday. I estimate these regressions at the firm-level using one year of weekly returns, requiring a minimum of 24 weeks of data. Each measure of price delay (D1, D2, and D3) is constructed from the regression R^2 values, the estimated β coefficients, or the β coefficient standard errors. Appendix Table 1 includes formal definitions of each variable.

In order to evaluate the connection between Compustat data coverage and these alternative measures of information assimilation, I estimate regressions as in eq. (8) at the annual frequency. In daily return autocorrelation regressions, II is defined as $|\widehat{\rho}|$, $|\frac{\widehat{\rho}}{se(\widehat{\rho})}|$, or the r-squared from regression (9). In price delay regressions, II is defined as D1, D2, or D3. In all cases, the informational inefficiency measures are constructed using stock return data from July in year t through June in year $t+1$. The 'Missing Data' indicator reflects whether a firm has any Compustat data coverage for the $t-1$ fiscal year-end, and all other variables are measured as of the end of June in year t.

Results are reported in Table 7 and uniformly indicate that missing Compustat data is associated with stronger return autocorrelations and stronger price delays. Results in Panel A indicate that the autocorrelation coefficients are over 0.02 larger in magnitude (≈20% of a standard deviation) for firms with no Compustat data relative to firms with Compustat coverage. Similarly, the autocorrelation t-statistics more than 0.3 higher (≈20% of a standard deviation), and the r-squared values from the AR(1) regressions are 1.5% higher (≈50% of a standard deviation), for firms with missing data. Likewise, results in Panel B indicate that the fraction of stock-specific return variance captured by lagged market returns is approximately 4.6% larger (≈15% of a standard deviation) for firms with no Compustat

data relative to firms with Compustat coverage. Results are again similar across alternative price delay measures.

Collectively, results in Table 7 are consistent with the notion that equity prices are less informationally efficient for firms that are not covered in Compustat: when Compustat does not cover a firm, return autocorrelations are larger, price delay measures are larger, and information assimilation is slower. Results in Panel B are also consistent with the conclusion that this effect is offset if there are sufficiently many analysts covering the firm and/or sufficiently many institutions investing in the firm. Although the interaction coefficient estimates in Panel A are uniformly insignificant, the interaction coefficients in Panel B suggest that an increase of approximately 6-10 analysts or an increase in institutional ownership of 6-10% offsets the impact of missing Compustat data on price delay measures. These results are collectively consistent with the earnings announcement analysis in Section 8.1, and suggest that limited access to financial statement data reduces the informational efficiency of equity prices via its impact on market participation and investor attention.

9 Conclusion

Over the last several decades, data has become a pivotal component of the global economy (Farboodi and Veldkamp, 2023), leading to the rise of a robust industry of data aggregators. These data vendors act as information intermediaries in a variety of contexts by collecting and aggregating data on clients' behalf. Despite their popularity in financial research and practice, however, it is unclear how a data vendor's data coverage ultimately affects information access and investor actions. Specifically, when building and maintaining any database, the data vendor must decide what information is collected and when, and when and how that information is updated over time. This raises an important question: if a significant fraction of market participants rely on a common data vendor, can these data coverage decisions impact real outcomes, such as investment? This paper sheds light on this issue.

Standard & Poor's Compustat database provides subscribers with decades of 10-K and

10-Q information. However, Compustat does not cover every public firm in every period. I examine how the completeness of Compustat's data coverage affects institutional investor demand. I first show that institutional investment in firms with no Compustat coverage is over 36% below its unconditional mean. Importantly, I introduce a novel, quasi-natural experiment to confirm a plausibly causal connection between Compustat data coverage and institutional ownership: a technology shock at S&P in the 1990s causes a discrete reduction in missing data. This change in data coverage is followed by a significant increase in institutional investment for treated firms relative to control firms. I confirm that institutional investors more likely to use accounting data are significantly more reliant on Compustat's data coverage. Finally, I evaluate the connection between Compustat coverage and information assimilation, and show that missing Compustat data is associated with lower informational efficiency of equity prices.

This study highlights the role that data vendors play in capital markets, and emphasizes the impact their data coverage decisions can have on investor actions. Although many of the frictions related to the intermediation of *financial statement* information have attenuated in recent years, financial statement data is only one subset of potentially relevant information. Data vendors such as Glassdoor, the Carbon Disclosure Project, the Privacy Rights Clearinghouse, online retailers such as Amazon, and social media platforms such as Facebook and Twitter now provide information related to employee satisfaction, pollution, cybersecurity risk, consumer attention, retail investor sentiment, and other firm characteristics. All of this information is plausibly relevant to a significant fraction of investors. As such, the role that data vendors play as information intermediaries, and their impact on investor actions, will continue to be relevant in studies of financial markets.

References

Akbas, F., Markov, S., Subasi, M., Weisbrod, E., 2018. Determinants and consequences of information processing delay: Evidence from the thomson reuters institutional brokers' estimate system. Journal of Financial Economics 127, 366–388.

Badrinath, S., Gay, G. D., Kale, J. R., 1989. Patterns of institutional investment, prudence, and the managerial "safety-net" hypothesis. Journal of Risk and Insurance 56, 605–629.

Badrinath, S., Kale, J. R., Ryan, H. E., 1996. Characteristics of common stock holdings of insurance companies. Journal of Risk and Insurance 63, 49–76.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159–178.

Ben-Rephael, A., Da, Z., Israelsen, R. D., 2017. It depends on where you search: Institutional investor attention and underreaction to news. Review of Financial Studies 30, 3009–3047.

Bird, A., Karolyi, S. A., 2016. Do institutional investors demand public disclosure? Review of Financial Studies 29, 3245–3277.

Blankespoor, E., deHaan, E., Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. Journal of Accounting and Economics 70, 101344.

Bochkay, K., Markov, S., Subasi, M., Weisbrod, E., 2022. The roles of data providers and analysts in the production, dissemination, and pricing of street earnings. Journal of Accounting Research 60, 1695–1740.

Boone, A. L., White, J. T., 2015. The effect of institutional ownership on firm transparency and information production. Journal of Financial Economics 117, 508–533.

Boritz, J. E., No, W. G., 2020. How significant are the differences in financial data provided by key data sources? a comparison of xbrl, compustat, yahoo! finance, and google finance. Journal of Information Systems 34, 47–75.

Bowles, B., Reed, A. V., 2024. Mutual fund shorts and the benefits of acquiring information.

Working Paper .

Bowles, B., Reed, A. V., Ringgenberg, M. C., Thornock, J. R., 2024. Anomaly time. Journal of Finance, forthcoming .

Bryzgalova, S., Lerner, S., Lettau, M., Pelger, M., 2024. Missing financial data. Review of Financial Studies 00, 1–80.

Bushee, B. J., Matsumoto, D. A., Miller, G. S., 2003. Open versus closed conference calls: The determinants and effects of broadening access to disclosure. Journal of Accounting and Economics 34, 149–180.

Bushee, B. J., Noe, C. F., 2000. Corporate disclosure practices, institutional investors, and stock return volatility. Journal of Accounting Research 38, 171–202.

Chen, A. Y., McCoy, J., 2024. Missing values handling for machine learning portfolios. Journal of Financial Economics 155, 1–15.

Chen, S., Miao, B., Shevlin, T., 2015. A new measure of disclosure quality: The level of disaggregation of accounting data in annual reports. Journal of Accounting Research 53, 1017–1054.

Chen, Y., Kelly, B., Wu, W., 2022. Sophisticated investors and market efficiency: Evidence from a natural experiment. Journal of Financial Economcics 138, 316–341.

Chuk, E., Matsumoto, D., Miller, G. S., 2013. Assessing methods of identifying management forecasts: Cig vs. researcher collected. Journal of Accounting and Economics 55, 23–42.

Chychyla, R., Kogan, A., 2015. Using xbrl to conduct a large-scale study of discrepancies between the accounting numbers in compustat and sec 10-k filings. Journal of Information Systems 29, 37–72.

Correia, S., Guimarães, P., Zylkin, T., 2019. Verifying the existence of maximum likelihood estimates for generalized linear models.

Correia, S., Guimarães, P., Zylkin, T., 2020. Fast Poisson estimation with high-dimensional fixed effects. The Stata Journal 20, 95–115.

Crane, A. D., Crotty, K., Umar, T., 2023. Hedge funds and public information acquisition.

Management Science 69, 3241–3262.

Cremers, K. M., Petajisto, A., 2009. How active is your fund manager? a new measure that predicts performance. Review of Financial Studies 22, 3329–3365.

Del Guercio, D., 1996. The distorting effect of the prudent-man laws on institutional equity investment. Journal of Financial Economics 40, 31–62.

Drake, M. S., Roulstone, D. T., Thornock, J. R., 2015. The determinants and consequences of information acquisition via edgar. Contemporary Accounting Research 32, 1128–1161.

D'Souza, J., Ramesh, K., Shen, M., 2010. The interdependence between institutional ownership and information dissemination by data aggregators. The Accounting Review 85, 159–193.

Du, K., Huddart, S., Jiang, X., 2023. Lost in standardization: Effects of financial statement database discrepancies on inference. Journal of Accounting and Economics 75, 1–28.

Easterwood, S., 2024. Why is data missing in crsp and compustat? Working Paper .

Edelen, R. M., Hosseinian, A., Kadlec, G. B., 2022. The investable universe of 13f institutions. Working Paper .

Edelen, R. M., Ince, O., Kadlec, G. B., 2016. Institutional investors and stock return anomalies. Journal of Financial Economics 119, 472–488.

Falkenstein, E., 1996. Preferences for stock characteristics as revealed by mutual fund portfolio holdings. Journal of Finance 51, 111–135.

Fama, E. F., 1991. Efficient capital markets: Ii. Journal of Finance, 46, 1575–1617.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33, 3–56.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics, 116, 1–22.

Farboodi, M., Matray, A., Veldkamp, L., Venkateswaran, V., 2022. Where has all the data gone? Review of Financial Studies 35, 3101–3138.

Farboodi, M., Veldkamp, L., 2020. Long-run growth of financial data technology. American

Economic Review 110, 2485–2523.

Farboodi, M., Veldkamp, L., 2023. Data and markets. Annual Review of Economics 15, 23–40.

Farboodi, M., Veldkamp, L., 2024. A model of the data economy. NBER Working Paper pp. 1–56.

Fink, J., 2020. A review of the post-earnings-announcement drift. Journal of Behavioral and Experimental Finance 29, 1–13.

French, K. R., Roll, R., 1986. Stock return variances: The arrival of information and the reaction of traders. Journal of Financial Economics 17, 5–26.

Freyberger, J., Hoppner, B., Neuhierl, A., Weber, M., 2024. Missing data in asset pricing panels. Review of Financial Studies 00, 1–43.

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

Goldfarb, A., Tucker, C., 2019. Digital economics. Journal of Economic Literature 57, 3–43.

Gompers, P., Metrick, A., 2001. Institutional investors and equity prices. Quarterly Journal of Economics 116, 229–259.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393–408.

Harris, T., Morsfiled, S., 2012. An evaluation of the current state and future of xbrl and interactive data for investors and analysts. Columbia Business School Center for Excellence in Accounting and Security Analysis, White Paper Number Three.

Healy, P., Palepu, K., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of Accounting and Economics, 31, 405–440.

Hirshleifer, D., Ma, L., 2024. The effect of new information technologies on asset pricing anomalies. Working Paper pp. 1–41.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and

the profitability of momentum strategies. Journal of Finance 55, 265–295.

Hou, K., Moskowitz, T. J., 2005. Market frictions, price delay, and the cross-section of expected returns. Review of Financial Studies, 18, 981–1020.

Jones, C. I., Tonetti, C., 2020. Nonrivalry and the economics of data. American Economic Review 110, 2819–2858.

Kacperczyk, M., Sundaresan, S., Wang, T., Jiang, W., 2021. Do foreign institutional investors improve price efficiency? Review of Financial Studies 34, 1317–1367.

Kaplan, Z., Martin, X., Xie, Y., 2021. Truncating optimism. Journal of Accounting Research 59, 1827–1884.

Karpoff, J. M., Koester, A., Lee, D. S., Martin, G., 2017. Proxies and databases in financial misconduct research. The Accounting Review 92, 129–163.

Kim, S., Kim, S., 2023. Fragmented securities regulation, information-processing costs, and insider trading. Management Science 70, 4407–4428.

Kim, Y. H., Ivkovich, Z., Muravyev, D., 2024. Causal effect of information costs on asset pricing anomalies. Working Paper .

Koijen, R., Yogo, M., 2019. A demand system approach to asset pricing. Journal of Political Economy 127, 1475–1515.

Kolanovic, M., Krishnamachari, R. T., 2017. Big data and AI strategies: machine learning and alternative data approach to investing. Global Quantitative & Derivatives Strategy, J.P. Morgan.

Kothari, S., Zhang, L., Zuo, L., 2023. Disclosure regulation: Past, present, and future. Handbook of Financial Decision Making pp. 215–234.

Lewellen, J., 2011. Institutional investors and the limits of arbitrage. Journal of Finance 102, 62–80.

Ljungqvist, A., Malloy, C., Marston, F., 2009. Rewriting history. Journal of Finance 64, 1935—1960.

Merton, R. C., 1987. A simple model of capital market equilibrium with incomplete infor-

mation. Journal of Finance 42, 483–510.

Netter, J., Stegemoller, M., Wintoki, M. B., 2011. Implications of data screens on merger and acquisition analysis: A large sample study of mergers and acquisitions from 1992 to 2009. Review of Financial Studies, 24, 2316–2357.

Noble, K. B., 1982. Sec data: Difficult hunt. The New York Times .

Petajisto, A., 2013. Active share and mutual fund performance. Financial Analysts Journal 69, 73–93.

Pricewaterhouse Coopers, LLP, 2006, June 8. Sec letter. Available at: https://www.sec.gov/news/press/4-515/4515-8.pdf .

Rice, T., Strahan, P. E., 2010. Does credit competition affect small-firm finance? Journal of Finance 65, 861–889.

Schaub, N., 2018. The role of data providers as information intermediaries. Journal of Financial and Quantitative Analysis 53, 1805–1838.

Standard & Poor's, 2003. Standard & Poor's Compustat User's Guide. The McGraw-Hill Companies, Inc.

Thomson Reuters, 2010. Worldscope Database – Data Definitions Guide. Thomson Reuters.

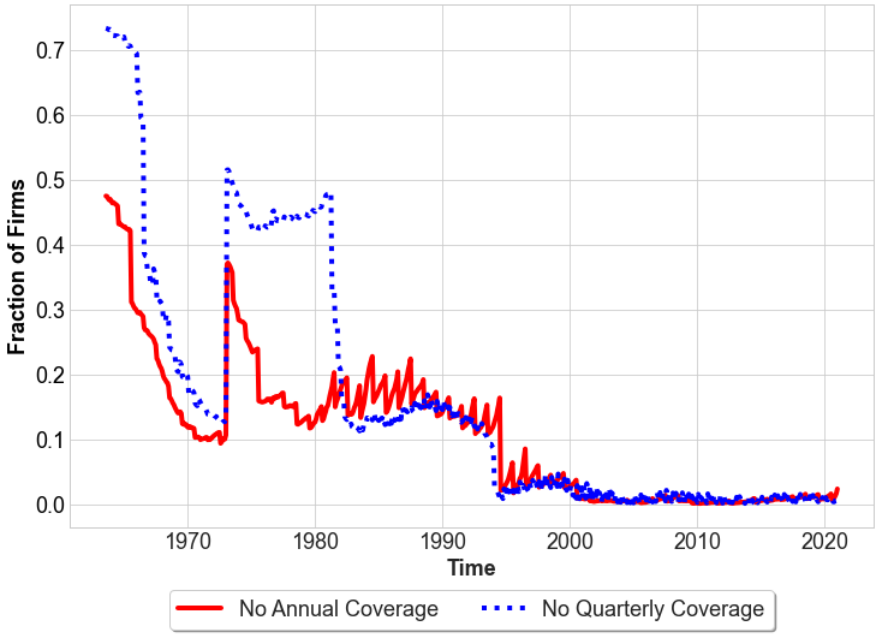
Yan, X., Zhang, Z., 2009. Institutional investors and equity returns: Are short-term institutions better informed. Review of Financial Studies 22, 893–924.

Zarutskie, R., 2006. Evidence on the effects of bank competition on firm borrowing and investment. Journal of Financial Economics 81, 503–537.

Figure 1: **Missing Compustat Data Over Time**

This figure reports the fraction of firms with missing Compustat data over time. Panel A shows the fraction of firm-month observations with no Compustat coverage. Panel B shows the fraction of firm-month observations with missing values for a variety of Compustat variables used to construct popular accounting-based firm characteristics.

(a) Fraction of Firms with No Compustat Data



(b) Fraction of Firms with Missing Characteristic Data

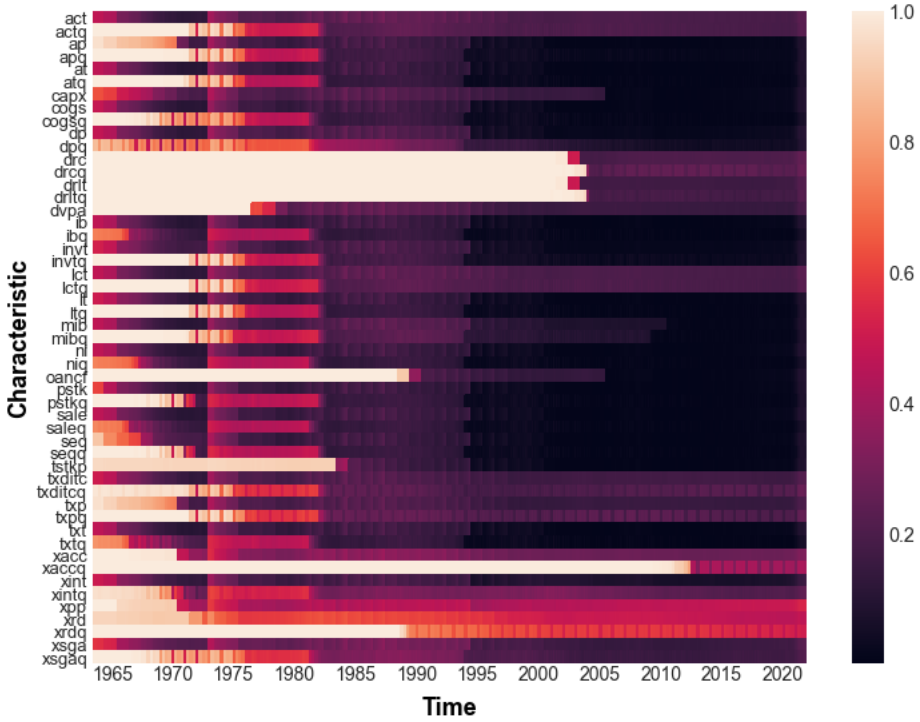
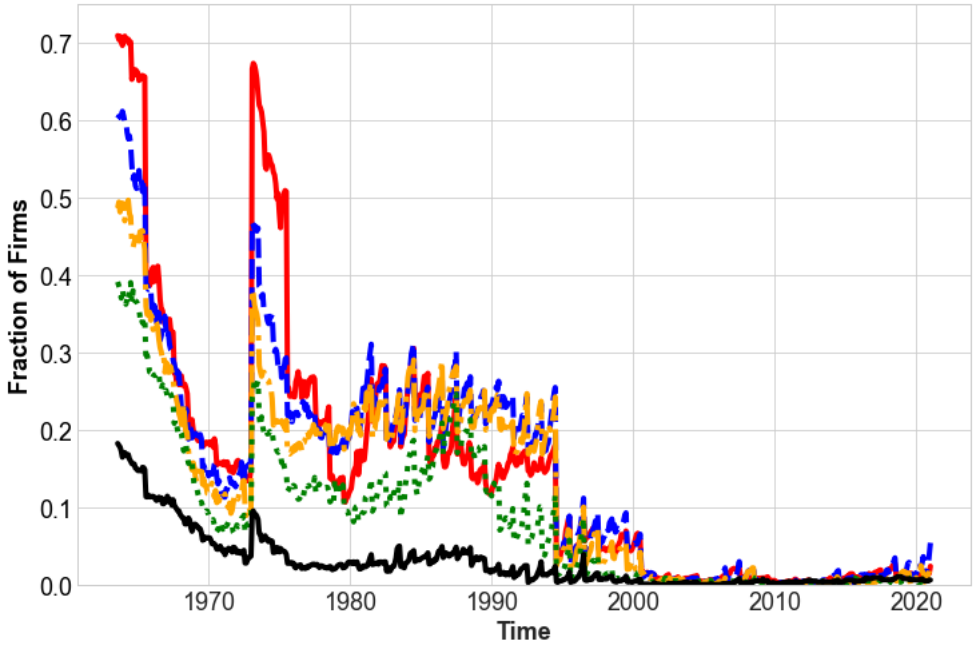


Figure 2: **Missing Compustat Data versus Firm Size**

This figure displays the fraction of firms with no Compustat data coverage in each of five size quintiles over the time. Size quintiles are defined based on market capitalization measured as of the end of the prior month. Panel A focuses on annual Compustat data coverage. Panel B focuses on quarterly Compustat data coverage.

(a) Firms with No Annual Data Coverage



(b) Firms with No Quarterly Data Coverage

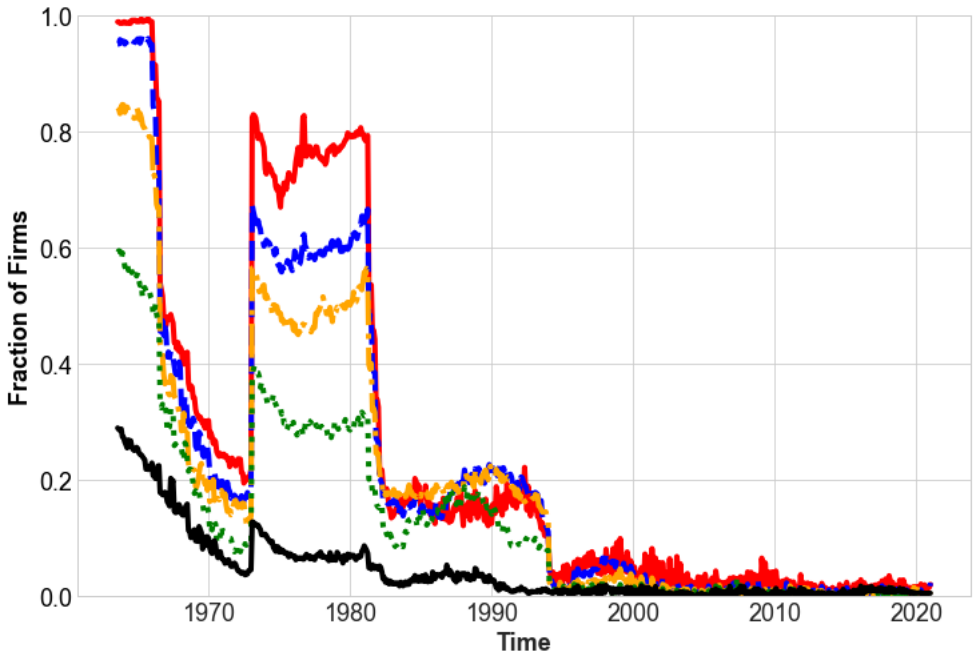
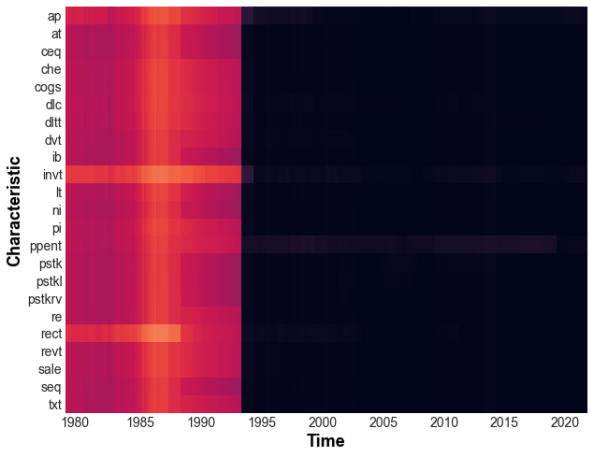


Figure 3: **Change in Compustat Data Coverage for Financial Firms**

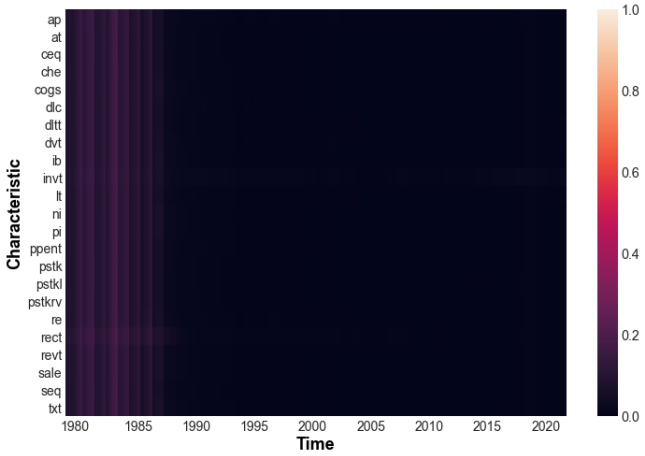
This figure reports the fraction of firm-month observations with missing values of a variety of Compustat input variables over time. Panel A defines data as missing if it is missing in the Compustat North America database. Panel B defines data as missing if it was not available in Compustat in real-time; this is measured using the Compustat Point-in-Time database. Left-hand figures show results for financial services firms. Right-hand figures show results for all other firms. Standard & Poor's classifies firms with SIC codes ranging from 6000 – 6999, excluding codes 6411, 6792, 6794, 6795, as financial services. All firms are required to be publicly listed in or before Q1 1988.

(a) Compustat North America Database: 1980–2021

Financial Firms

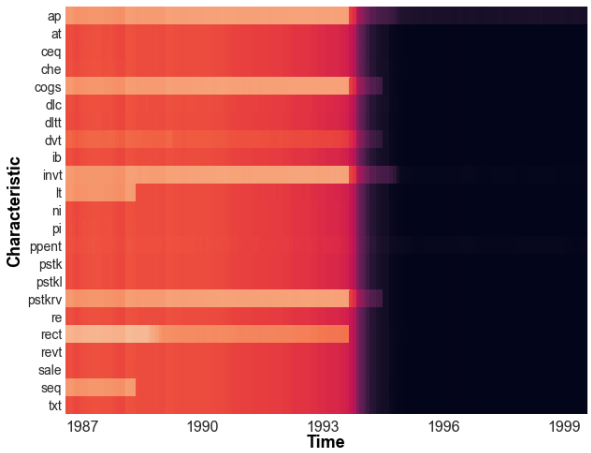


Non-Financial Firms



(b) Compustat Point-in-Time Database: 1987–1999

Financial Firms



Non-Financial Firms

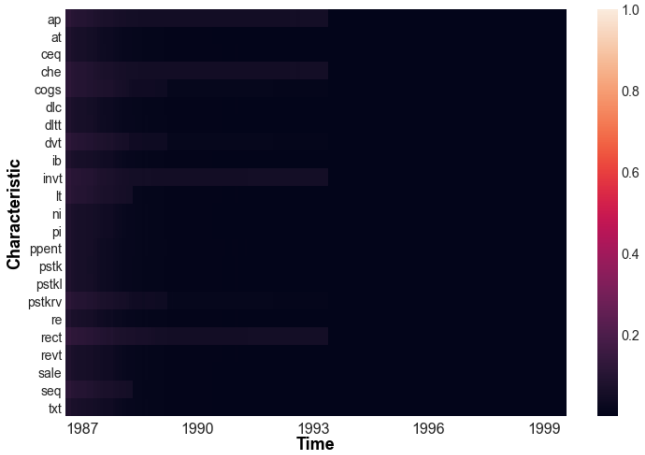
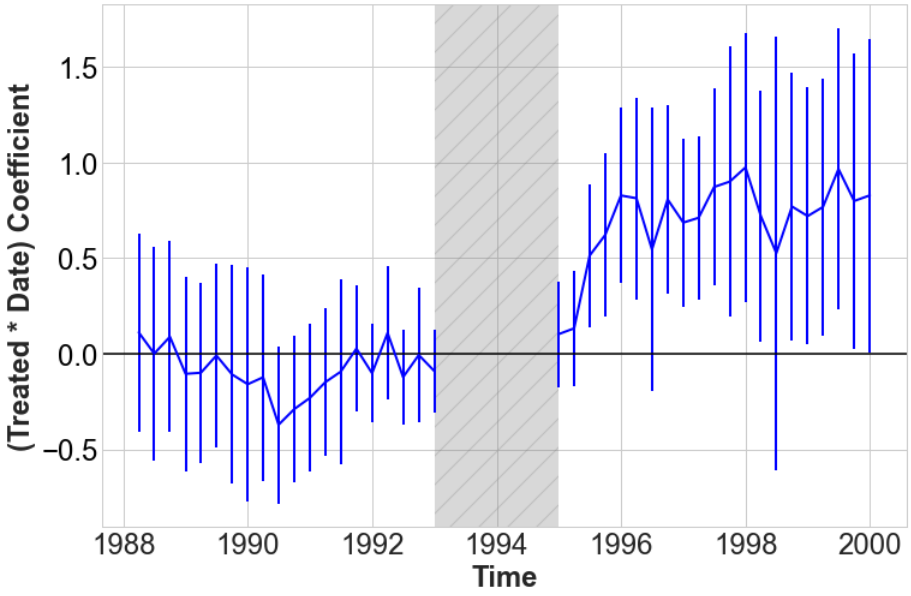


Figure 4: **Parallel Trends**

This figure displays results for various tests of parallel trends for the difference-in-differences analysis decribed in Section 6.2. Panel A reports dynamic Treated×Date coefficient estimates from regression eq. (2), including a 99% confidence interval. Panel B reports the cross-sectional average institutional ownership for treated (dotted-blue line) versus control (solid-yellow line) firms between Q1 1988 and Q4 1999. The left-hand figure reports average FNIO. The right-hand figure reports average NIO. The grey, diagonal slash shaded region indicates the period over which treated firms' financial data began to appear in the Compustat North America database.

(a) Dynamic Treatment Effects



(b) Average Institutional Ownership

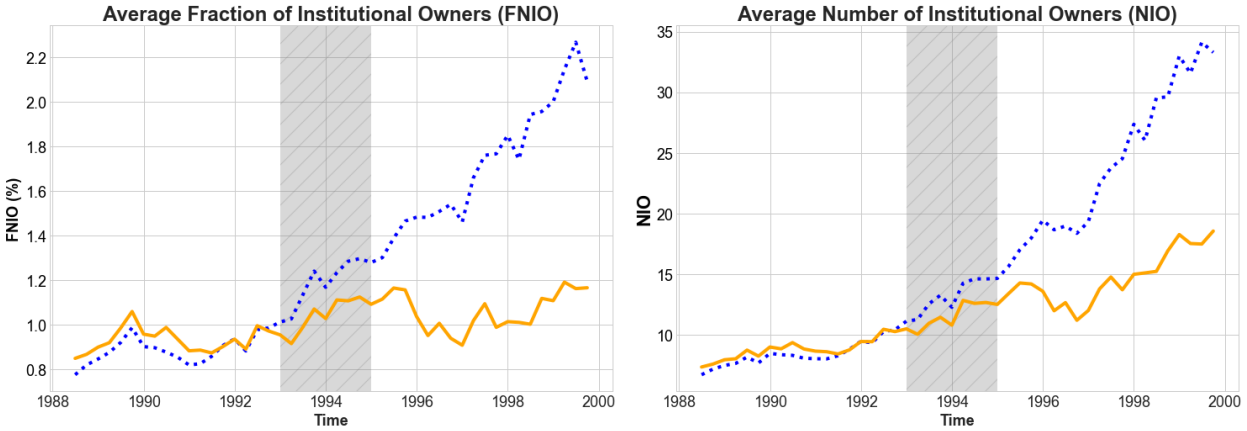


Figure 5: **Staggered Implementation of EDGAR**

This figure reports the fraction of firms subject to mandatory electronic filing on EDGAR over the period April 1993 – May 1996. I include only those firms that have an exact CIK or name match in SEC Release No. 33-6977 – Appendix B, which identifies when firms were required to begin filing on EDGAR. Treated (dotted-blue line) and control (solid-orange line) firms are defined as in the differences-in-differences analysis in Table 3. The grey, diagonal slash shaded region indicates the period over which treated firms' financial data began to appear in the Compustat North America database.

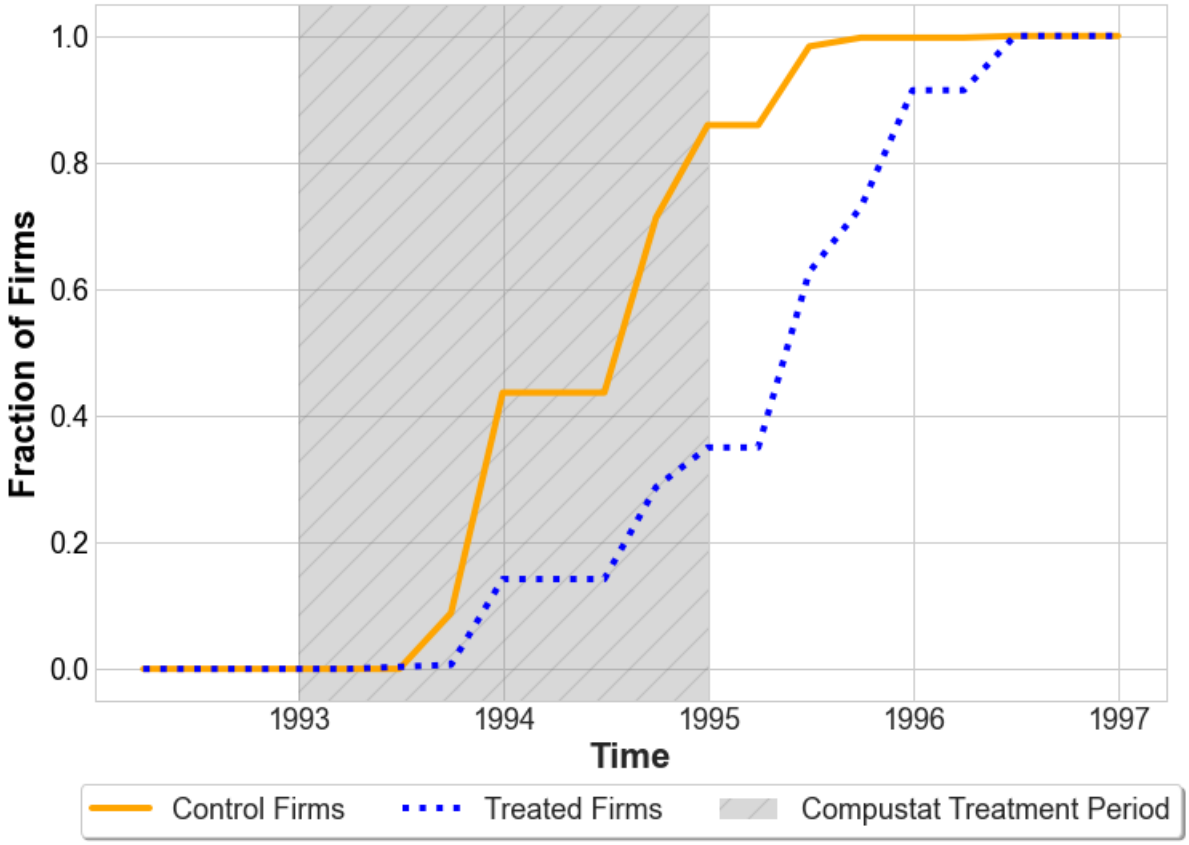


Table 1: **Summary Statistics**

This table presents summary statistics for institutional ownership, mutual fund ownership, and analyst coverage. $FNIO_{i,q}$ is equal to the number of institutions that hold shares of stock i in quarter q, scaled by the total number of institutions in the 13f dataset (s34 file) in quarter q. $FSIO_{i,q}$ is equal to the fraction of stock i's shares outstanding held by institutions in quarter q. $FNMF_{i,q}$ is equal to the number of mutual funds that hold shares of stock i in quarter q, scaled by the total number of mutual funds in the 13f dataset (s12 file) in quarter q. Analyst Coverage is equal to the number of unique analysts covering a stock.

Panel A: FNIO (%)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	4.03	6.96	0.12	1.58	10.23	6.6
1980s	3.61	7.84	0.0	0.79	9.90	17.8
1990s	3.73	6.83	0.11	1.32	9.64	5.4
2000s	4.39	6.42	0.26	2.31	10.50	0.9
2010s	4.56	6.48	0.27	2.64	10.88	1.0

Panel B: FSIO (%)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	34.77	30.96	0.39	26.53	83.78	6.6
1980s	16.11	18.42	0.0	8.82	45.17	17.8
1990s	27.14	24.17	0.58	20.81	64.26	5.4
2000s	44.50	31.38	3.64	42.28	88.71	0.9
2010s	57.95	32.85	5.68	66.50	96.34	1.0

Panel C: FNIO by Institution Type (%)

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/ Advisor	Miscellaneous
Mean	8.16	8.17	9.37	6.40	2.48	2.23
Std	12.00	13.53	14.79	9.18	4.94	4.89
10%	0	0	0	0	0	0
Median	2.56	3.18	2.17	2.43	0.87	0
90%	23.94	20.75	30.19	18.06	6.14	6.45

Panel D: Fraction of Institutions w/in Type Classifications (%)

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/ Advisor	Miscellaneous
Full Sample	1.99	7.74	2.15	5.14	72.49	10.49
1980s	7.74	27.43	6.17	15.88	37.20	5.58
1990s	4.09	15.47	3.17	17.46	56.07	3.74
2000s	1.64	5.53	2.15	8.61	72.13	9.94
2010s	1.01	3.30	1.54	4.30	80.92	8.93

Panel E: Mutual Fund Ownership (FNMF, %)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	1.18	2.36	0	0.35	3.01	18.80
1980s	0.89	2.11	0	0.19	2.42	40.60
1990s	0.90	2.05	0	0.21	2.36	20.80
2000s	1.22	2.46	0.02	0.43	3.00	4.60
2010s	1.91	2.76	0.03	1.07	5.43	5.20

Panel F: Analyst Coverage (#)

	Mean	Std	10%	Median	90%	Fraction = 0
Full Sample	4.52	6.39	0	2	14	36.0
1980s	3.35	6.11	0	0	12	53.5
1990s	3.97	6.10	0	1	12	38.2
2000s	4.61	5.81	0	2	13	29.9
2010s	6.67	7.18	0	4	17	18.0

Table 2: **Investor Demand and Missing Data**

This table reports regressions of investor demand, primarily institutional ownership, on an indicator for missing Compustat data, as in eq. (1). 'Missing Data' is an indicator variable defined as 1 if a firm has no Compustat data available for its most recent fiscal year-end, and 0 otherwise. In Panels A and B, institutional ownership (FNIO or FSIO) is measured for all institutions in aggregate. In Panels C and D, institutional ownership (FNIO) is measured for individual types and sizes of institutions. 'Inflation-Adjusted Size Cutoff' measures FNIO with respect to only those institutions whose total EUM falls above or below the inflation-adjusted 13f reporting threshold. (The threshold is equal to $100 million in Q1 1980, and is adjusted over time for inflation.) In Panel E, investor demand is measured via mutual fund ownership (FNMF, s12 file). In Panel F, demand is measure via analyst coverage, which is a count variable equal to the number of unique analysts covering the firm. All measures of demand (except analyst coverage) are expressed in percentage points. Standard errors are clustered by firm and date, t-statistics are reported in parentheses, and marginal effects for non-linear regressions are reported in brackets. Linear regressions are estimated via ordinary least squares. Poisson pseudo-likelihood regressions are estimated as in Correia et al. (2019) and Correia et al. (2020). All regressions are linear unless indicated otherwise. Continuous control variables are winsorized at the 1% and 99% levels. Controls include: the Koijen and Yogo (2019) market cap instrument, an S&P500 indicator, age, and age squared. Industry is defined as 2-digit SIC code. The sample period is 1980–2021.

Panel A: Aggregate Institutional Ownership with Dependent Variable FNIO

	All Firms						Drop Smallest 20%	
Missing Data	-3.2925 ***	-1.4729 ***	-0.3439 ***	-1.4950 ***	-0.3560 ***	-0.1696 ***	1.5054 ***	-0.3934 ***
	(-31.09)	(-24.45)	(-8.54)	(-33.85)	(-10.50)	(-7.08)	(-20.76)	(-8.24)
				[-6.0225]	[-1.4343]	[-0.6919]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	2.13%	66.11%	90.19%	3.86%	63.08%	69.36%	66.25%	89.91%
N	848,838	848,838	848,499	848,838	848,834	837,947	678,999	678,360

Panel B: Aggregate Institutional Ownership with Dependent Variable FSIO

	All Firms						Drop Smallest 20%	
Missing Data	-12.0925 ***	-5.6754 ***	-2.4382 ***	-0.6456 ***	-0.3136 ***	-0.1787 ***	-5.0561 ***	-2.3527 ***
	(-36.10)	(-18.53)	(-9.57)	(-21.58)	(-12.43)	(-14.32)	(-14.46)	(-8.18)
				[-22.4436]	[-10.9021]	[-6.2922]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	29.02%	58.15%	83.23%	23.67%	51.80%	72.26%	60.08%	83.03%
N	848,838	848,838	848,499	848,838	848,834	837,947	678,999	678,360

Table 2: **Investor Demand and Missing Data**

Panel C: Institution Legal Type

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/Advisor	Miscellaneous	
Missing Data	-1.1041 ***	-2.1734 ***	-1.0672 ***	-1.2958 ***	-0.8793 ***	-0.5063 ***	
	(-11.50)	(-20.65)	(-7.52)	(-18.29)	(-17.95)	(-11.40)	
Controls	Y	Y	Y	Y	Y	Y	
Time FE	Y	Y	Y	Y	Y	Y	
Industry FE	Y	Y	Y	Y	Y	Y	
Exchange FE	Y	Y	Y	Y	Y	Y	
Adjusted R^2	71.84%	69.37%	74.65%	66.39%	56.61%	52.77%	
N	848,838	848,838	848,838	848,838	848,838	848,838	

Panel D: Institution Size

EUM Quintile:	Q1 (Small)	Q2	Q3	Q4	Q5 (Large)	30 Largest	Inflation-Adjusted Size Cutoff (Below)	(Above)
Missing Data	-0.5216 ***	-0.7647 ***	-0.9921 ***	-1.4906 ***	-3.6030 ***	-8.1773 ***	-0.5321 ***	-1.7158 ***
	(-12.62)	(-13.65)	(-15.92)	(-20.11)	(-29.98)	(-26.54)	(-11.72)	(-25.74)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	39.91%	45.81%	52.06%	62.60%	75.80%	74.67%	43.03%	71.28%
N	848,838	848,838	848,838	848,838	848,838	848,838	848,838	848,838

Panel E: Mutual Fund Ownership

	All Firms						Drop Smallest 20%	
Missing Data	-0.8692 ***	-0.3329 ***	-0.1037 ***	-1.5460 ***	-0.2649 ***	-0.0504 *	-0.3293 ***	-0.1077 ***
	(-26.24)	(-17.52)	(-6.70)	(-30.67)	(-6.64)	(-1.94)	(-14.36)	(-5.54)
				[-1.8278]	[-0.3132]	[-0.0636]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	4.28%	59.07%	84.90%	5.75%	51.38%	56.99%	58.51%	84.44%
N	848,838	848,838	848,499	848,838	848,799	794,630	678,999	678,360

Panel F: Analyst Coverage

	All Firms						Drop Smallest 20%	
Missing Data	-3.3346 ***	-1.9895 ***	-1.2005 ***	-1.4874 ***	-0.6583 ***	-0.4881 ***	-2.1765 ***	-1.4950 ***
	(-37.38)	(-33.09)	(-22.92)	(-29.55)	(-15.32)	(-14.92)	(-31.47)	(-24.32)
				[-6.7260]	[-2.9770]	[-2.5526]		
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	N	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	N	Y
Exchange FE	N	Y	N	N	Y	N	N	Y
Firm FE	N	N	Y	N	N	Y	N	N
Adjusted R^2	5.84%	60.23%	82.98%	6.39%	52.32%	61.01%	59.88%	82.10%
N	848,838	848,838	848,499	848,838	848,777	733,861	678,999	678,360

Table 3: **Difference-in-Differences Analysis**

This table reports difference-in-differences regression results, as in eq. (2). The dependent variables are measures of demand, primarily institutional ownership. In Panel A, institutional ownership (FNIO) is measured for all institutions in aggregate. In Panels B and C, institutional ownership is measured for individual types and sizes of institutions. 'Inflation-Adjusted Size Cutoff' is defined as in Table 2. In Panel D, investor demand is measured via mutual fund ownership (FNMF, s12 file). In Panel E, demand is measured via analyst coverage. All measures of demand (except analyst coverage) are expressed in percentage points. Treated firms are defined as all financial services firms with no data in the Compustat Point-in-Time database prior to 1993. Control firms are defined as financial services firms not listed in the S&P500 with complete data coverage in the Compustat Point-in-Time database from 1988–1992, and are matched to treated firms based on size, institutional ownership (FNIO), and turnover, all measured in Q1 1988, using KNN matching with replacement (K=1). Both treated and control firms are required to be publicly listed in or before Q1 1988. Linear regressions are estimated via ordinary least squares. Poisson pseudo-likelihood regressions are estimated as in Correia et al. (2019) and Correia et al. (2020). All regressions are linear unless indicated otherwise. Standard errors are clustered by firm and date, t-statistics are reported in parentheses, and marginal effects for non-linear models are reported in brackets. Continuous control variables are winsorized at the 1% and 99% levels. Controls include: the Koijen and Yogo (2019) market cap instrument, age, and age squared. Industry is defined as 2-digit SIC code. The sample period covers Q1 1988 – Q4 1999.

Panel A: Aggregate Institutional Ownership

	All Firms						Drop Smallest 20%	
Treated×Post	0.6105 ***	0.7613 ***	0.6480 ***	0.4473 ***	0.4616 ***	0.3503 ***	0.8373 ***	0.7060 ***
	(3.90)	(5.15)	(5.01)	(3.54)	(5.82)	(4.34)	(5.03)	(5.01)
				[0.4928]	[0.5086]	[0.3884]		
Treated	0.0211	-0.2843		0.0220	-0.2940 **		-0.4250 **	
	(0.15)	(-1.42)		(0.15)	(-2.35)		(-2.09)	
				[0.0242]	[-0.3239]			
Post	0.1038			0.1036				
	(0.82)			(0.89)				
				[0.1142]				
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	Y	Y
Time FE	N	Y	Y	N	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	Y	N
Exchange FE	N	Y	N	N	Y	N	Y	N
Firm FE	N	N	Y	N	N	Y	N	Y
Adjusted R^2	4.42%	35.98%	80.46%	2.13%	21.79%	33.84%	38.99%	80.43%
N	27,923	27,923	27,922	27,923	27,923	27,752	25,089	25,088

Panel B: Institution Legal Type

Institution Type:	Insurance Company	Bank	Pension Fund	Mutual Fund Company	Investment Company/Advisor	Miscellaneous
Treated×Post	1.4851 ***	1.5037 ***	1.5471 ***	0.7096 ***	0.2704 ***	0.0721
	(3.92)	(5.40)	(3.15)	(3.21)	(3.54)	(0.43)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	36.23%	39.43%	27.14%	35.13%	24.32%	11.15%
N	27,923	27,923	27,923	27,923	27,923	27,923

Table 3: **Difference-in-Differences Analysis – Financial Firm Matched Sample**

Panel C: Institution Size

EUM Quintile:	Q1 (Small)	Q2	Q3	Q4	Q5 (Large)	30 Largest	Inflation-Adjusted Size Cutoff (Below)	(Above)
Treated×Post	0.0971 **	0.1197 ***	0.3402 ***	0.5255 ***	2.1519 ***	7.6243 ***	0.1015 **	0.8536 ***
	(2.05)	(2.72)	(4.25)	(3.45)	(4.29)	(5.19)	(2.33)	(4.41)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y	Y	Y
Adjusted R^2	12.40%	15.12%	20.49%	23.43%	39.69%	46.97%	14.50%	36.02%
N	27,923	27,923	27,923	27,923	27,923	27,923	27,923	27,923

Panel D: Mutual Fund Holdings

	All Firms						Drop Smallest 20%	
Treated×Post	0.0933 ***	0.1342 ***	0.0835 ***	0.4378 ***	0.4415 ***	0.2923 ***	0.1506 ***	0.0933 ***
	(3.04)	(3.37)	(3.12)	(2.79)	(3.38)	(2.59)	(3.34)	(3.12)
				[0.0909]	[0.0917]	[0.0635]		
Treated	-0.0210	-0.0436		-0.1041	-0.2223		-0.0704	
	(-0.51)	(-0.81)		(-0.53)	(-1.37)		(-1.19)	
				[-0.0216]	[-0.0462]			
Post	-0.0296			-0.1504				
	(-1.09)			(-1.02)				
				[-0.0312]				
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	Y	Y
Time FE	N	Y	Y	N	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	Y	N
Exchange FE	N	Y	N	N	Y	N	Y	N
Firm FE	N	N	Y	N	N	Y	N	Y
Adjusted R^2	0.50%	24.23%	74.02%	0.27%	15.65%	26.70%	26.25%	73.95%
N	27,923	27,923	27,922	27,923	27,923	27,715	25,089	25,088

Panel E: Analyst Coverage

	All Firms						Drop Smallest 20%	
Treated×Post	1.2910 ***	1.4604 ***	1.3777 ***	0.8859 ***	0.8091 ***	0.8164 ***	1.5557 ***	1.4586 ***
	(6.63)	(6.52)	(7.54)	(3.21)	(3.68)	(3.71)	(6.23)	(7.39)
				[1.1103]	[1.0140]	[1.2936]		
Treated	0.7245 ***	0.1896		0.7421 ***	-0.0297		0.1274	
	(4.33)	(0.59)		(3.43)	(-0.15)		(0.37)	
				[0.9301]	[-0.0372]			
Post	-0.1661			-0.2907				
	(-1.21)			(-1.06)				
				[-1.07]				
Regression	Linear	Linear	Linear	Poisson	Poisson	Poisson	Linear	Linear
Controls	N	Y	Y	N	Y	Y	Y	Y
Time FE	N	Y	Y	N	Y	Y	Y	Y
Industry FE	N	Y	N	N	Y	N	Y	N
Exchange FE	N	Y	N	N	Y	N	Y	N
Firm FE	N	N	Y	N	N	Y	N	Y
Adjusted R^2	9.93%	25.20%	75.99%	8.96%	30.24%	43.05%	28.26%	76.66%
N	27,923	27,923	27,922	27,923	27,923	22,087	25,089	25,088

Table 4: **Difference-in-Differences Sub-Sample Analysis**

This table reports difference-in-differences regression results for various sub-samples. The dependent variable is aggregate institutional ownership (FNIO), and is expressed in percentage points. Treated and controls firms are defined as in Table 3. In Panel A, columns 1-6 report regression results corresponding to eq. (2) and sub-samples are defined based on when firms began filing electronically on EDGAR. I include only those firms that have an exact CIK or name match in SEC Release No. 33-6977 – Appendix B, which identifies when firms were required to begin filing electronically on EDGAR. Columns 7 and 8 in Panel A report regression results corresponding to eq. (3). In Panel B, columns 1-4 report regression results corresponding to eq. (2) and sub-samples are defined based the 'Branching Restrictiveness Index' described in Rice and Strahan (2010). Firms' states are defined based on their address recorded in Compustat. Columns 5 and 6 in Panel B report regression results corresponding to eq. (4). In Panel C, columns 1-4 report regression results corresponding to eq. (2) and sub-samples are defined based on whether or not a firm participates in a merger or acquisition as target or acquirer at any point between Q1 1988 and Q4 1999. ('M&A firms' are defined as firms which participate in a merger or acquisition). Columns 5 and 6 in Panel C report regression results corresponding to eq. (5). In all regressions, standard errors are clustered by firm and date and t-statistics are reported in parentheses. Continuous control variables are winsorized at the 1% and 99% levels. Controls include: the Koijen and Yogo (2019) market cap instrument, age, and age squared. The sample period covers Q1 1988 – Q4 1999.

Panel A: Aggregate Institutional Ownership and State Subgroups

Began Filing On EDGAR:	1993		1994		1995-1996		All	
Treated×Post	0.9032 ***	0.8222 ***	0.6065 **	0.8073 ***	0.7363 ***	0.6698 **	0.7363 ***	0.6271 **
	(2.71)	(3.22)	(2.46)	(3.45)	(3.78)	(2.30)	(3.78)	(2.26)
EDGAR(93)×Treated×Post							0.1669	0.0895
							(0.47)	(0.24)
EDGAR(94)×Treated×Post							-0.1298	0.3171
							(-0.45)	(0.96)
EDGAR(93)×Post							0.2948	0.1125
							(1.12)	(0.33)
EDGAR(94)×Post							0.0703	-0.2960
							(0.56)	(-1.10)
EDGAR(95/6)×Post							0.0936	
							(0.80)	
EDGAR(93)							1.1683 ***	
							(3.37)	
EDGAR(94)							0.1652	
							(0.59)	
Treated and Post Indicators	Y	N	Y	N	Y	N	N	N
EDGAR×Treated Indicators	N	N	N	N	N	N	Y	N
Controls	N	Y	N	Y	N	Y	N	Y
Time FE	N	Y	N	Y	N	Y	N	Y
Firm FE	N	Y	N	Y	N	Y	N	Y
Adjusted R^2	6.82%	80.71%	6.72%	73.72%	10.00%	80.93%	14.74%	80.05%
N	5,564	5,564	7,301	7,301	8,808	8,807	21,673	21,672

Table 4: **Difference-in-Differences Sub-Sample Analysis**

Panel B: Aggregate Institutional Ownership and State Subgroups

Branching Restrictiveness Index:	$2 \leqslant$		$\geqslant 3$		All	
Treated×Post	0.6641 ***	0.5895 ***	0.5744 **	0.6731 ***	0.6641 ***	0.5843 ***
	(5.18)	(2.94)	(2.58)	(4.14)	(5.18)	(3.18)
High×Treated×Post					-0.0897	0.1218
					(-0.36)	(0.55)
High×Treated					-0.0549	
					(-0.15)	
High×Post					0.1340	-0.0547
					(0.69)	(-0.30)
Treated	0.0363		-0.0186		0.0363	
	(0.11)		(-0.13)		(0.11)	
Post	0.0148		0.1488		0.0148	
	(0.42)		(0.80)		(0.42)	
High					-0.0732	
					(-0.21)	
Controls	N	Y	N	Y	N	Y
Time FE	N	Y	N	Y	N	Y
Firm FE	N	Y	N	Y	N	Y
Adjusted R^2	4.59%	81.88%	3.69%	79.43%	4.55%	80.48%
N	11,784	11,784	16,139	16,138	27,923	27,922

Panel C: Aggregate Institutional Ownership and M&A Subgroups

M&A:	No		Yes		All	
Treated×Post	0.4233 **	0.5262 ***	0.6621 ***	0.6822 ***	0.4233 **	0.5889 ***
	(2.45)	(3.50)	(2.77)	(3.52)	(2.45)	(4.07)
M&A×Treated×Post					0.2388	0.0316
					(0.82)	(0.15)
M&A×Treated					-0.1884	
					(-0.67)	
M&A×Post					0.1119	0.1691
					(0.43)	(0.92)
Treated	0.0958		-0.0926		0.0958	
	(0.48)		(-0.46)		(0.48)	
Post	0.0695		0.1814		0.0695	
	(0.49)		(0.84)		(0.49)	
M&A					0.1974	
					(0.76)	
Controls	N	Y	N	Y	N	Y
Time FE	N	Y	N	Y	N	Y
Firm FE	N	Y	N	Y	N	Y
Adjusted R^2	2.45%	85.19%	6.07%	77.77%	5.06%	80.55%
N	12,041	12,040	15,882	15,882	27,923	27,922

Table 5: **Institutions' Propensity to Invest in Firms with Missing Data**

This table reports regressions of the fraction of an institution's or mutual fund's portfolio invested in firms with no Compustat data on a variety of institution/mutual fund characteristics, as in eq. (6) and (7). Dependent variables are defined as the 'Fraction of EUM,' equal to the fraction of an institution's equity under management invested in firms with no Compustat coverage, or the 'Fraction of Firms,' defined as the fraction of firms that the institution is invested in with no Compustat coverage. Panel A focuses on 13f institutions (s34 data), and observations are recorded at the institution-quarter frequency. The sample in Panel A is 1986–2021. Panel B focuses on the individual mutual funds (s12 data), and observations are recorded at the mutual fund-quarter frequency. The sample in Panel B is 1980–2009. Standard errors are clustered by institution/mutual fund and date, and t-statistics are reported in parentheses.

Panel A: Institutional Investors

Dependent Variable:	Fraction of EUM			Fraction of Firms		
Log(EUM)	-0.0443 ***	-0.0436 ***	-0.0171	-0.0222 **	-0.0225 **	0.0062
	(-3.89)	(-3.79)	(-0.73)	(-2.27)	(-2.29)	(0.29)
Age	0.0018 ***	0.0019 ***	-0.0112	0.0007 *	0.0008 **	-0.0163 ***
	(3.19)	(3.25)	(-1.35)	(1.68)	(2.02)	(-3.02)
Portfolio Turnover	0.3555 ***		0.1816 **	0.4328 ***		-0.0347
	(3.13)		(2.28)	(2.85)		(-0.54)
Turnover: Quint 2		-0.0180			-0.0050	
		(-0.45)			(-0.23)	
Turnover: Quint 3		-0.0429			0.0003	
		(-0.86)			(0.01)	
Turnover: Quint 4		0.0167			0.1347 ***	
		(0.30)			(2.96)	
Turnover: Quint 5		0.1714 **			0.2772 ***	
		(2.31)			(2.88)	
Past Returns Trader	0.3833 ***	0.3829 ***	0.1481 ***	0.5543 ***	0.5400 ***	0.1639 ***
	(5.36)	(5.42)	(2.85)	(6.17)	(6.20)	(3.21)
Insurance Company	-0.2677 ***	-0.2648 ***		-0.2067 **	-0.1959 **	
	(-2.86)	(-2.83)		(-2.19)	(-2.08)	
Bank	-0.1450	-0.1450		-0.2578 ***	-0.2334 ***	
	(-1.54)	(-1.52)		(-3.43)	(-3.17)	
Pension Fund	-0.2709 ***	-0.2711 ***		-0.3143 ***	-0.2983 ***	
	(-3.03)	(-3.01)		(-3.52)	(-3.37)	
Mutual Fund Company	-0.0520	-0.0416		0.1085	0.1133	
	(-0.58)	(-0.47)		(1.23)	(1.30)	
Investment Company	0.0226	0.0257		0.0507	0.0528	
	(0.27)	(0.31)		(0.74)	(0.77)	
Time FE	Y	Y	Y	Y	Y	Y
Institution FE	N	N	Y	N	N	Y
Adjusted R^2	5.14%	5.16%	33.04%	12.05%	12.13%	38.53%
N	310,742	310,742	310,344	310,742	310,742	310,344

Table 5: **Institutions' Propensity to Invest in Firms with Missing Data**

Panel B: Mutual Funds

Dependent Variable:	Fraction of EUM			Fraction of Firms		
Log(EUM)	-0.0984 ***	-0.0514 ***	-0.1607 ***	-0.0785 ***	-0.0289	-0.1442 ***
	(-4.90)	(-3.22)	(-4.47)	(-3.71)	(-1.61)	(-3.73)
Age	-0.0011	-0.0008	-0.0001	-0.0016 *	-0.0013	0.0020
	(-1.31)	(-0.96)	(-0.02)	(-1.68)	(-1.34)	(0.41)
Portfolio Turnover	1.4419 ***		0.8395 ***	1.6147 ***		0.7148 ***
	(4.25)		(3.83)	(4.06)		(3.16)
Turnover: Quint 2		0.1713 ***			0.2055 ***	
		(4.42)			(4.61)	
Turnover: Quint 3		0.2251 ***			0.2693 ***	
		(4.31)			(4.44)	
Turnover: Quint 4		0.3132 ***			0.3765 ***	
		(4.79)			(4.69)	
Turnover: Quint 5		0.5242 ***			0.5963 ***	
		(5.11)			(4.95)	
Past Returns Trader	0.8265 ***	0.5153 ***	0.8395 ***	0.9617 ***	0.6316 ***	0.5244 ***
	(7.08)	(5.23)	(3.83)	(6.72)	(5.16)	(4.90)
Active Share	1.1002 ***		0.0627	1.2991 ***		0.1252
	(4.98)		(0.23)	(4.74)		(0.38)
Active Share: Quint 2		0.1615 ***			0.2009 ***	
		(3.74)			(3.49)	
Active Share: Quint 3		0.3461 ***			0.4208 ***	
		(4.90)			(4.59)	
Active Share: Quint 4		0.6251 ***			0.7377 ***	
		(6.68)			(6.30)	
Active Share: Quint 5		0.9692 ***			1.0742 ***	
		(6.68)			(6.66)	
Index Fund	1.0176 ***	0.6466 ***		1.2096 ***	0.7818 ***	
	(4.71)	(4.61)		(4.60)	(4.59)	
Enhanced Index	0.7452 ***	0.5249 ***		0.8935 ***	0.6357 ***	
	(4.75)	(5.57)		(4.84)	(5.64)	
Time FE	Y	Y	Y	Y	Y	Y
Mutual Fund FE	N	N	Y	N	N	Y
Adjusted R^2	26.79%	28.43%	45.12%	33.14%	34.57%	51.19%
N	62,868	62,868	62,749	62,868	62,868	62,749

Table 6: **Earnings Announcements and Missing Data**

This table reports regressions of abnormal returns measured during and after quarterly earnings announcements on an indicator for missing Compustat data. Regressions are estimated as in eq. (8), standard errors are clustered by time, and t-statistics are reported in parentheses. 'Missing Data' is an indicator variable defined as 1 if a firm does not have any Compustat data available in the most recent fiscal year-end. The dependent variables are defined as absolute cumulative abnormal returns ('ACAR') around quarterly earnings announcements. Trading day $\tau = 0$ is defined as the earnings announcement date. In Panel A, ACARs are constructed over the window $\tau = [-1, 1]$. In Panel B, post-announcement ACARs are constructed over the window $\tau = [2, 60]$. All observations are recorded at the firm-announcement frequency, and controls are measured as of the most recent quarter-end. Controls in all regressions include: analyst coverage, institutional ownership (FNIO), stock beta, log market cap, an S&P500 indicator, prior 1-year return, prior return measured over years -5:-1, net stock issuance, share turnover, and age. Continuous control variables are winsorized at the 1% and 99% levels. Industry is defined as 2-digit SIC code. The sample period is 1980–2021.

Panel A: Earnings Announcement Returns, ACAR[-1,1]

Missing Data	0.5372 ***	0.3566 ***	0.4426 ***	0.3157 ***	0.9743 ***	1.1653 ***	0.7893 ***	1.0097 ***
	(5.38)	(3.37)	(5.03)	(3.22)	(8.06)	(8.97)	(7.42)	(8.41)
Missing Data×	-0.0745 ***	-0.0616 ***			-0.1301 ***	-0.1553 ***		
Analyst Coverage	(-4.48)	(-3.50)			(-6.50)	(-6.29)		
Missing Data×			-0.0396 ***	-0.0569 ***			-0.0564 ***	-0.1305 ***
FNIO			(-3.98)	(-4.44)			(-4.39)	(-4.85)
Normal Return	Market	Market	Market	Market	FF3	FF3	FF3	FF3
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	N	Y	N	Y	N	Y	N
Exchange FE	Y	N	Y	N	Y	N	Y	N
Firm FE	N	Y	N	Y	N	Y	N	Y
Adjusted R^2	10.31%	14.92%	10.31%	14.92%	10.72%	15.71%	10.72%	15.71%
N	586,970	586,319	586,970	586,619	550,779	550,321	550,779	550,321

Panel B: Post-Announcement Drift, ACAR[2,60]

Missing Data	1.0242 ***	0.8523 ***	0.8502 ***	0.6862 ***	0.6677 **	1.0716 **	0.3788	0.8167 **
	(3.89)	(2.81)	(3.70)	(2.57)	(2.01)	(2.37)	(1.32)	(2.06)
Missing Data×	-0.0635	-0.0887 *			-0.1007 *	-0.1656 **		
Analyst Coverage	(-1.44)	(-1.79)			(-1.87)	(-2.26)		
Missing Data×			0.0212	-0.0255			0.0530	-0.0836
FNIO			(0.73)	(-0.84)			(1.39)	(-1.39)
Normal Return	Market	Market	Market	Market	FF3	FF3	FF3	FF3
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	N	Y	N	Y	N	Y	N
Exchange FE	Y	N	Y	N	Y	N	Y	N
Firm FE	N	Y	N	Y	N	Y	N	Y
Adjusted R^2	13.07%	18.27%	13.07%	18.27%	13.46%	18.66%	13.46%	18.66%
N	586,970	589,619	586,970	589,619	550,779	550,321	550,779	550,321

Table 7: **Information Assimilation and Missing Data**

This table reports regressions of various measures of informational efficiency on an indicator for missing Compustat data, as in eq. (8). 'Missing Data' is an indicator variable defined as 1 if a firm does not have any Compustat data available in the most recent fiscal year-end. In Panel A, the dependent variables are defined as measures of return autocorrelations, estimated as the absolute coefficient value ($|\widehat{\rho}|$), t-statistic ($|\frac{\widehat{\rho}}{se(\widehat{\rho})}|$, or r-squared from the regression in eq. (9). In Panel B, the dependent variables are defined as measures of price delay obtained from the regression in eq. (10) and defined in Appendix Table 1. All observations are recorded at the annual frequency. Autocorrelation and price delay measures are constructed using return data from July in year t through June in year $t+1$. Missing Compustat data is measured as of the $t-1$ fiscal year-end. Controls are measured as of the end of June in year t. Standard errors in all regressions are clustered by time, and t-statistics are reported in parentheses. Controls in all regressions include: analyst coverage, institutional ownership (FNIO), stock beta, log market cap, an S&P500 indicator, prior 1-year return, prior return measured over years -5:-1, net stock issuance, log share turnover, and age. Continuous control variables are winsorized at the 1% and 99% levels. Industry is defined as 2-digit SIC code. The sample period is 1980–2021.

Panel A: Daily Return Autocorrelation

Dependent Variable:	$\lvert\widehat{\rho}\rvert$	$\lvert\widehat{\rho}\rvert$	$\lvert\frac{\widehat{\rho}}{se(\widehat{\rho})}\rvert$	$\lvert\frac{\widehat{\rho}}{se(\widehat{\rho})}\rvert$	R^2(%)	R^2(%)
Missing Data	0.0208 **	0.0218 **	0.3232 **	0.3420 **	1.4677 ***	1.5339 ***
	(2.25)	(2.27)	(2.02)	(2.06)	(2.79)	(2.80)
Missing Data×	0.0002		0.0098		-0.0466	
Analyst Coverage	(0.22)		(0.59)		(-0.82)	
Missing Data×		-0.0009		-0.0120		-0.1177
FNIO		(-0.67)		(-0.57)		(-1.41)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	29.42%	29.42%	26.16%	26.16%	28.62%	28.64%
N	190,559	190,559	190,559	190,559	190,559	190,559

Panel B: Price Delay

Dependent Variable:	D1 (%)	D1 (%)	D2	D2	D3	D3
Missing Data	4.6485 ***	4.6829 ***	0.2066 ***	0.2026 ***	0.1559 ***	0.1591 ***
	(4.97)	(5.29)	(3.84)	(3.87)	(2.98)	(3.19)
Missing Data×	-0.4511 **		-0.0249 **		-0.0278 **	
Analyst Coverage	(-2.60)		(-2.35)		(-2.43)	
Missing Data×		-0.4521 *		-0.0184 **		-0.0291 **
FNIO		(-1.91)		(-2.28)		(-2.61)
Controls	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y
Exchange FE	Y	Y	Y	Y	Y	Y
Adjusted R^2	36.51%	36.51%	4.32%	4.32%	3.94%	3.94%
N	194,365	194,365	194,365	194,365	194,365	194,365

Appendix Table 1: **Characteristic Definitions**

Panel A: Firm Characteristics	
Analyst Coverage	The number of analysts covering a firm in a given quarter, equal to the number of quarterly earnings forecasts made by unique analysts ('NUMEST' from the I/B/E/S Summary file).
Beta	Monthly CAPM beta estimated using the prior 60 months of returns; require a minimum of 36 months of data.
Fractional Number of Institutional Owners (FNIO)	The total number of institutions that own shares of a firm in a given quarter, scaled by the total number of institutions in the 13f data in that quarter.
Fractional Number of Mutual Fund Owners (FNMF)	The total number of mutual funds that own shares of a firm in a given quarter, scaled by the total number of mutual funds in the 13f data (s12 file) in that quarter.
Fraction of Shares Held by Institutions (FSIO)	The fraction of a firm's total shares outstanding held by institutions in a given quarter.
Fraction of Shares Held by Mutual Funds (FSMF)	The fraction of a firm's total shares outstanding held by mutual funds in a given quarter.
Market Cap Instrument	Constructed as in Koijen and Yogo (2019), and equal to the log of the counter-factual market equity if all institutions held an equally-weighted portfolio of their investable universe. Each institution's investable universe is defined as all stocks that they currently hold or have held over the prior three years. Institutions' counter-factual investments are defined as total equity under management multiplied by 1/N, where N is the total number of firms in the institutions' investable universe. Each firm's counter-factual market equity is defined as the sum of all counter-factual investments for each institution.
Net Issuance	Annual log change in split-adjusted shares outstanding. Shares outstanding are measured as of the prior December-end.
Number of Institutional Owners (NIO)	The total number of institutions that own shares of a firm in a given quarter.
Price Delay (D1)	Constructed as in Hou and Moskowitz (2005), $D1 = 1 - \frac{R^2_{\beta_i^{-n}=0,\forall n\in[1,4]}}{R^2}$, where R^2 is the r-squared from regression (10), and $R^2_{\beta_i^{-n}=0,\forall n\in[1,4]}$ is the r-squared from regression (10) when restricting $\beta_i^{-n} = 0$ for $\forall n \in [1,4]$.
Price Delay (D2)	Constructed as in Hou and Moskowitz (2005), $D2 = \frac{\sum_{n=1}^{4} n\beta^{-n}}{\beta^0 + \sum_{n=1}^{4} \beta^{-n}}$, where β^{-n} is the relevant coefficient estimate from regression (10).
Price Delay (D3)	Constructed as in Hou and Moskowitz (2005), $D3 = \frac{\sum_{n=1}^{4}\left(n\beta^{-n}/se(\beta^{-n})\right)}{(\beta^0/se(\beta^0)) + \sum_{n=1}^{4}\left(\beta^{-n}/se(\beta^{-n})\right)}$, where $se(\beta^{-n})$ is the standard error of the relevant coefficient estimate from regression (10).
Turnover	Trading volume, averaged over the prior M months, divided by total shares outstanding. In annual regressions, $M = 12$. In quarterly regressions, $M = 3$.

Appendix Table 1: **Characteristic Definitions**

Panel B: Institution and Mutual Fund Characteristics	
Active Share	Constructed as in Cremers and Petajisto (2009) and Petajisto (2013), and equal to the sum of absolute differences between each mutual fund's portfolio weights and the fund's benchmark's portfolio weights. Active share data is obtained from Annti Petajisto's website, and is available only for the mutual fund (s12) data.
Age	The number of quarters that the institution (s34 data) or mutual fund (s12 data) has appeared in the 13f database.
Past Returns Trader	The sum of absolute betas from a regression of an institution's (s34 data) or mutual fund's (s12 data) current quarter trades on the firms' returns measured over the previous four quarters: $\frac{Shares_{i,j,q}-Shares_{i,j,q-1}}{TSO_{i,q-1}} = \alpha_{j,q} + \sum_{n=1}^{4}\left(\beta_{j,q}^{n} * Return_{i,q-n}\right) + \epsilon_{i,j,q}$, where $Shares_{i,j,q}$ denotes the number of shares of stock i that institution j holds at the end of quarter q, $TSO_{i,q}$ is firm i's total shares outstanding at the end of quarter q, and $Return_{i,q}$ is firm i's return, cross-sectionally ranked and scaled to fall in the interval [-0.5,0.5], measured over quarter q. The estimated $\widehat{\beta_{j,q}^{n}}$ are cross-sectionally winsorized at the 1% and 99% levels, and the characteristic 'past returns trader' is defined as: $\sum_{n=1}^{4}\left\lvert\widehat{\beta_{j,q}^{n}}\right\rvert$. Each institution j must hold more than five stocks in a given quarter q to construct this characteristic. The trade variable, $\frac{Shares_{i,j,q}-Shares_{i,j,q-1}}{TSO_{i,q-1}}$, is defined only if $Shares_{i,j,q} > 0$ and/or $Shares_{i,j,q-1} > 0$.
Portfolio Turnover	Constructed following Yan and Zhang (2009), and equal to the four-quarter average of an institution's or mutual fund's churn rate. Churn rate is equal to the minimum of aggregate purchases and aggregate sales, both measured over the most recent quarter, scaled by the average of equity under management at the beginning and end of the quarter.