# Information Acquisition and Expected Returns: Evidence from EDGAR Search Traffic*

Frank Weikai Li† Chengzhu Sun‡

This Draft: September 2019

**Abstract**

Using a novel dataset containing investors' access of company filings through the SEC's EDGAR system, we show that the abnormal number of IPs searching for firms' financial statements strongly predicts future stock returns and firm fundamentals. Consistent with theories of costly information acquisition, the return predictability is more pronounced for firms with larger and lengthier financial filings that are more costly to process, and for IPs searching current and historical filings simultaneously. Our findings suggest investors' costly information acquisition activities reveal their expectation of future firm performance.

*JEL classification*: G12, G14

*Keywords*: Information Acquisition, EDGAR Search, SEC Filings

*We are grateful to Hengjie Ai, Utpal Bhattacharya, Justin Birru, Ekkehart Boehmer, Hao Chen, Si Cheng, Darwin Choi, Lauren Cohen, Zhi Da (Discussant), Sudipto Dasgupta, Fangjian Fu, Amit Goyal, Johan Hombert, Charles Hsu, Claire Yurong Hong, Dashan Huang, Wenxi Jiang, Ron Kaniel, Xinlei Li, Cen Ling, Hongqi Liu, Roger Loh, Roni Michaely, Chase Potter (Discussant), Ronnie Sadka, Rik Sen, Chenyu Shan (Discussant), Tao Shen, Elvira Sojli (Discussant), Derrald Stice, Melvyn Teo, Baolian Wang, Chishen Wei, K.C. John Wei, Jialin Yu, Huai Zhang (Discussant), Bohui Zhang, Zilong Zhang and seminar participants at the EFA 2018, CICF 2018, AAA 2018, Asian FA 2018, the 2017 Singapore Scholars Symposium, the Inaugural Conference on Intelligent Information Retrieval in Accounting and Finance, and LKCSB Summer Research Camp, Singapore Management University, HKUST, Chinese University of Hong Kong, City University of Hong Kong, Peking University, Tsinghua SEM, University of Macau, CUHK Shenzhen, and Southern University of Science and Technology for helpful comments. We also thank Robert Stambaugh and Yu Yuan for sharing the mispricing factor returns and Kewei Hou, Chen Xue, and Lu Zhang for sharing the q-factor returns. All errors are our own.

†Corresponding author: Singapore Management University, Lee Kong Chian School of Business (Email: wkli@smu.edu.sg)

‡Hong Kong University of Science and Technology (Email: csunab@connect.ust.hk)

# 1 Introduction

Information acquisition and dissemination is key to understanding asset price movements and market efficiency. When information is costly to acquire and price is only partially revealing, economic agents will expend resources and effort to become informed (Grossman and Stiglitz (1980); Verrecchia (1982)), and in doing so, will move prices closer to the fundamental value. A central prediction from theories of costly information acquisition is that more investors will choose to become informed when they perceive greater benefits from doing so, holding the cost of information acquisition constant.[1] Although theories offer clear and rich predictions, empirical evidence of the relation between information acquisition behavior and the value of information is sparse in financial markets, potentially due to the difficulty of directly measuring the information acquisition activities of investors.

In this paper, we take advantage of a novel dataset containing investors' access of regulatory filings through the Securities and Exchange Commission (SEC)'s EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system to study the implications of information acquisition activities on firm value. Because the EDGAR system is the main source of firms' regulatory filings, and the SEC maintains a log file of all activities performed by users on EDGAR, we are able to directly observe investors' information acquisition activity for a broad cross-section of firms over a sample period of more than 10 years.

Our research objectives in this paper are twofold. First, we examine the determinants of investors' information acquisition through the EDGAR website. Motivated by theories of information acquisition,[2] we posit that information acquisition activities should be negatively related to the cost of gathering and analyzing information, and positively related to the (perceived) benefits of information. To test this, we use the number of unique IP addresses searching for SEC filings through EDGAR as a proxy for investors' information acquisition. We then run cross-sectional regressions of our information acquisition proxy on several firm characteristics associated with the costs and benefits of information acquisition. Specifically,

[1]The definition of "information acquisition", as is commonly used in the literature, not only includes cost of acquiring information, but also the cost of analyzing and interpreting information.

[2]There is a large body of theoretical literature on information acquisition, e.g., Grossman and Stiglitz (1980), Diamond and Verrecchia (1981), Verrecchia (1982), Hellwig (1980), Admati (1985), Veldkamp (2011) and Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016).

we hypothesize that firms with higher investor visibility and attention will attract more information acquisition, as these stocks are more accessible in investors' minds and less costly to analyze. We conjecture that the strength of firms' information environments would affect information acquisition, although the direction of the effect is not clear ex-ante.[3] We also expect investors to have stronger incentives to acquire information about firms with higher valuation uncertainty (Mele and Sangiorgi (2015)). Using firm size as a proxy for investor visibility, trading volume as a proxy for investor attention (Gervais, Kaniel, and Mingelgrin (2001); Barber and Odean (2007)), analyst coverage as a proxy for information environment (Hong, Lim, and Stein (2000)), and idiosyncratic volatility as a proxy for valuation uncertainty (Zhang (2006)), we find evidence consistent with the theories. These four firm characteristics explain 55% of the cross-sectional variation of information acquisition activities across firms. Further tests show that information acquisition through EDGAR also increases following negative stock returns, for firms belonging to the S&P 500 index, held by more institutional investors and during earnings announcement months, but these additional characteristics do not dramatically improve the explanatory power of our baseline model.

After implementing a simple characteristic-based model of expected information acquisition, we proceed to examine our second research question, that an abnormal level of information acquisition reflects investors' expected benefits of trading on information. This prediction is based on the simple premise that when attention-constrained investors decide how to allocate their limited time and effort, they will have a strong preference for firms with the largest price appreciation or depreciation potential. In reality, investors will more likely engage in costly information acquisition when the expected return of a stock is positive, due to the asymmetry in buying and selling decision (Barber and Odean (2007)). When deciding which stocks to buy, investors have to choose from thousands of available stocks, hence information acquisition becomes an important part of decision-making. On the other hand, unless using short selling, investor can only sell the stocks they currently own, and the selling decision is more likely motivated by liquidity and tax considerations, and less likely

[3]On one hand, firms with abundant public information will be less costly to analyze, so we expect information acquisition to increase with the quality of a firm's information environment. On the other hand, a better information environment also means that the stock is less likely to be mispriced ex-ante, so investors' incentives to acquire private information will be reduced.

to require information acquisition.

To that end, we extract the number of IPs unexplained by firm characteristics to infer investors' private expectations of future payoffs. Consistent with the idea that information acquisition embeds the value of information, we show that an abnormal number of IPs (denoted as AIP) requesting EDGAR filings strongly predicts subsequent stock returns. An equal-weighted, monthly rebalanced, long-short strategy that buys stocks in the highest decile of AIP and sells stocks in the lowest decile of AIP generates 59 to 80 basis points per month after adjustment for the Carhart (1997) four factors and is highly significant. Adjusting for the recently proposed factor models – the Fama and French (2015) five-factor model, the Hou, Xue, and Zhang (2015) q-factor model, and the Stambaugh and Yuan (2016) mispricing-factor model – does not affect the return spread of the long-short portfolio much. The abnormal return of AIP strategy is much weaker for value-weighted portfolios. The high-minus-low AIP strategy generates approximately 30 basis points per month, which is mostly insignificant. One potential explanation is that short-sale constraints are less binding for big stocks, so the direction of the information contained in AIP is more ambiguous for big stocks. Using several proxies of short-sale constraints including lendable supply and lending fees, we confirm that the positive expected return information embedded in EDGAR searching activities is more pronounced for stocks that are more costly to short.

With a Fama-MacBeth regression setting, we confirm that AIP has additional explanatory power for future stock returns when we control for the standard cross-sectional return predictors, such as firm size, book-to-market ratio, momentum, short-term reversal, idiosyncratic volatility, turnover, and institutional ownership. The return predictability of AIP persists for two quarters, and is not reversed in the subsequent 24 months. This persistence in return predictability alleviates concerns that our finding is the result of temporary price pressure caused by noise traders, which should reverse over the long-run (Da, Engelberg, and Gao (2011)). Furthermore, we show that within-firm change of AIP (relative to its 12-months moving average) also significantly predict future returns, suggesting that our result is unlikely driven by unobserved risk exposure which should be persistent at the firm level. The return predictability of AIP is also *not* explained by alternative channels such as investor recognition, media coverage, firm events, extreme returns, and investor disagreement. Examining within-industry return predictability, we find that AIP is able to significantly predict future returns for 10 out of 12 industries based on the Fama-French industry classification.

Looking into different types of EDGAR filings, we find that the return predictability of AIP comes mainly from those searching for firms' annual reports 10-Ks (AIP_10K). As analyzing 10-Ks is more costly than other types of SEC filings and those searching activities are more indicative of deliberate information acquisition, the stronger predictability of AIP for 10-Ks is consistent with theories of costly information acquisition.[4] To further substantiate our argument, we conduct two tests that explore the heterogeneity of return predictability by varing information acquisition costs. First, we use the filing size and word count of 10-Ks as proxies for the complexity of financial filings (Loughran and McDonald (2014)), and find that the return predictability of AIP is significantly stronger among firms with larger and lengthier 10-Ks that are more costly to process. Second, we show that the return predictability of AIP is more pronounced when we focus on IPs searching for the current and historical 10-Ks simultaneously. The evidence supports the hypothesis that in equilibrium, the expected benefits from information acquisition are proportional to the cost of acquiring and analyzing information, as predicted by theories of endogenous information acquisition.

Having established the robustness of the return predictability of the abnormal number of IPs, we examine the sources of return predictability. The underlying assumption in this paper is that under short-sale constraints, investors rationally allocate more effort and attention to underpriced stocks. As mispricing implies the separation of stock price from the fundamental value of a firm, we conjecture two non-mutually exclusive channels through which investors can identify mispricing. The first channel is investors' information acquisition activity revealing their private expectation of firms' fundamental performance that are yet to be priced in the market.[5] Consistent with the first channel, we find that AIP strongly predicts the future *changes* in firms' fundamentals such as quarterly Return-on-

[4]Cohen, Malloy, and Nguyen (2018) show that the length of the average 10-K has grown 6 times longer over the last 20 years.

[5]Investors may get informed about firm fundamentals, for example, by being exposed to advertisement on firms' product or major events in economically-linked firms (Liaukonyte and Zaldokas (2019); Madsen (2017)).

Assets (ROA), standardized unexpected earnings (SUE), and *revisions* in analyst consensus EPS forecast. Moreover, AIP significantly predict future earnings announcement returns, suggesting that the information contained in AIP is not immediately incorporated into stock prices and is (partially) revealed during earnings announcements.

The second channel of investors identifying mispricing is that investors may observe changes in stock prices due to exogeneous reasons. Supporting the second channel, we show that the abnormal number of IPs searching for EDGAR filings increases significantly after firms experiencing mutual fund outflow-induced fire sale (Coval and Stafford 2008; Edmans, Goldstein and Jiang 2012). Taken together, our evidence suggests that investors expend greater effort on undervalued stocks and these findings are much more difficult to reconcile with alternative explanations such as omitted risk factors or changes in investor recognition (Merton (1987))[6].

Lastly, we examine the incremental value of information acquisition through EDGAR given that some investors may already know potential misvaluation opportunities even before accessing EDGAR filings. We hypothesize that acquiring fundamental information through EDGAR could help investors distinguish truely mispriced stocks from those sharing similar mispricing characteristics. Our empirical tests support such a conjecture. Specifically, among the most undervalued quintile of stocks based on the composite mispricing measure of Stambaugh, Yu, and Yuan (2015), those with highest abnormal number of IPs generate a monthly four-factor alpha of 1.05%. In sharp contrast, these similarly undervalued stocks with lowest AIP do not have any abnormal returns. This result supports our premise that investors' costly information acquisition activity via EDGAR is being compensated as it allows them to identify truely mispriced stocks.

The remainder of this paper is organized as follows. Section 2 briefly surveys related literature and discusses the contribution of this study. Section 3 describes the data, presents summary statistics, and examines the determinants of information acquisition through EDGAR. Section 4 shows that the abnormal level of information acquisition reveals investors' expec-

[6] Alternative explanations based on omitted risk factor or changes in investor base all work through the discount-rate channel, while the return predictability of AIP operates (partially) through the cashflow channel.

tations of future payoffs. Section 5 tests the channels underlying the return predictability results. Section 6 conducts some additional analyses and robustness checks. Section 7 concludes the paper.

## 2 Related Literature and Contribution

This paper contributes to several strands of the existing literature. First, our results offer strong empirical evidence supporting information acquisition theories that information acquisition is endogenous to the value of information. Using the novel EDGAR log file dataset, we construct a direct measure of investors' information acquisition activity, and show its strong predictability for firms' future returns and fundamentals. Du (2015) shows that the number of web visits to SEC filings of insider trades predicts post-filing stock return in the short-run. Although similar in spirit, our paper differs from his paper as we study a much broader sample of SEC regulatory filings and longer horizon returns. We also test the channels underlying the return predictability results. Using EDGAR search data, Chen, Cohen, Gurun, Lou, and Malloy (2018) find that mutual funds tend to track a particular set of firms and insiders, and that their tracked trades generate abnormal performance.[7] Lee and So (2017) study the information content of analysts' selective coverage decisions and show that an abnormal amount of analyst coverage reflects analysts' favorable expectation of firms' fundamental performances. By extracting the information acquisition activities of all internet users through the EDGAR site, our measure captures the expected return information embedded in the collective behavior of a much larger set of market participants, i.e., millions of unique end-users of financial information. In addition, analysts' incentives have been found to be distorted by generating underwrting revenues (Lin and McNichols (1998)) or trading commissions for their brokerage houses (Cowen, Groysberg, and Healy (2006)); such distortions are less likely among EDGAR users. Empirically, we construct the AIP measure by controlling for analyst coverage proxies.

This paper also contributes to the growing literature on the effect of investor attention

[7]Several recent studies examine other market participants' access of SEC filings through EDGAR website, including financial analysts (Gibbons, Iliev, and Kalodimos (2019)), the Federal Reserve (Li, Lind, Ramesh, and Shen (2018)), and hedge funds (Crane, Crotty, and Umar (2018)).

and information acquisition on asset prices and capital market efficiency. Da, Engelberg, and Gao (2011) show that the abnormal attention of retail investors, as captured by Google search volume, causes transitory price pressures on attention-grabbing stocks. Using news-searching activity via the Bloomberg terminal as a proxy for institutional investors' attention, Ben-Rephael, Da, and Israelsen (2017) find that institutional attention facilitates the timely incorporation of fundamental information into asset prices. More pertinent to this study, Drake, Roulstone, and Thornock (2015) show that EDGAR-based information acquisition affects the efficient pricing of earnings-related news. However, the aforementioned papers mainly examine the effect of information acquisition on the pricing of *publicly* announced news, while this paper directly infers investors' *private* expectations of firm value through their collective actions.

Third, our work contributes to the emerging literature on extracting intelligence latent in the collective "wisdom of crowds". Chen, De, Hu, and Hwang (2014) document that investors' social media posts help predict stock return. Lee, Ma, and Wang (2015) show that investors' co-search patterns via the EDGAR website could help identify peer firms better than traditional industry benchmarks. Huang (2016) finds that consumer opinions of firms' products on Amazon.com contain value-relevant information about firm fundamentals and stock prices. Similarly, Green, Huang, Wen, and Zhou (2019) and Sheng (2019) document that employer reviews on Glassdoor reveal valuable information about employers' fundamentals. This paper complements the above studies as we infer agents' expectations not from what they "say", but from what they actually "do".

Finally, the sheer number of EDGAR IPs suggests that the majority of them should come from individual investors.[8] Thus our study also contributes to a recent literature documenting that individual investors as a group exhibit stock picking ability and their aggregate trading activities predict future stock returns and fundamental news (Kaniel, Saar, and Titman (2008); Kaniel, Liu, Saar, and Titman (2012); Kelley and Tetlock (2013); Boehmer, Jones, and Zhang (2017)). These papers speculate that retail investors may have valuable information obtained from geographic proximity to firms, relations with employees,

[8]Our sample contains more than 30 million unique IP addresses that ever searched any type of company filing through EDGAR server.

and or insights into consumer preferences. The finding of our paper suggests that retail investors actively use financial filings in EDGAR system to confirm their privately observed (noisy) information, which is likely a channel through which individual investors become informed.[9]

Our finding that information acquisition activity predicts future returns does not necessarily imply that the market is inefficient. As pointed out by Grossman and Stiglitz (1980), a fully efficient market where prices instantaneously reflect all available information cannot sustain an equilibrium when information is costly to acquire and analyze. Rather, our evidence is mostly consistent with the idea of "*efficiently inefficient markets*" (Pedersen (2015)), where competition among investors makes the market almost efficient, but the market also remains inefficient enough that these investors are compensated for their costs of acquiring and analyzing information.

# 3 Data and Methodology

## 3.1 Data

Our IP search volume data comes from the Securities and Exchange Commission's (SEC) EDGAR log file database, which has recorded all website search traffic for SEC filings since 2003.[10] Each search record contains information about the user's unique Internet Protocol (IP) address (partially anonymized)[11], timestamp, searched company (identified by the Central Index Key (CIK)) and searched specific filing (identified by the unique SEC accession number).[12] Following Lee, Ma, and Wang (2015) and Ryans (2017), we first filter the raw log data to eliminate the requests made by robots or automated webcrawlers, since such nu-

[9]Consistent with this idea, Gao and Huang (2019) show that trades by retail investors become more informative about future stock returns following the staggered implementation of the EDGAR system in 1993-1996.

[10]The data is available for download at https://www.sec.gov/data/edgar-log-file-data-set.html.

[11]The EDGAR log file dataset provides the first three octets of the IP address with the fourth octet obfuscated with a three character string that preserves the uniqueness of the last octet without revealing the full identity of the IP.

[12]The detailed log file record elements are described at https://www.sec.gov/files /EDGAR_variables_FINAL.pdf.

merous and indiscriminate requests are uninformativeness for our research question.[13] Next, we match the CIK in the EDGAR log filings to that in COMPUSTAT to identify public companies, and retrieve the filing type and filing date for each requested file by linking the accession number to the Master Index files maintained by the SEC.[14] We classify these filings into six groups: 10-K, 10-Q, 8-K, insider, registration, and proxy.[15] Finally, we calculate the monthly IP search volume for each filing category at firm level by counting the total number of unique IP addresses that searched one category of SEC filings of a specific company within a one-month window. We define IP_total as the total number of unique IP addresses searching all six types of SEC filings. Drake, Roulstone, and Thornock (2015) report that periodic accounting reports are the type of SEC filings most frequently requested by investors through the EDGAR website. We therefore also construct two additional measures of information acquisition specifically targeting firms' periodic accounting reports. IP_funtl (IP_10K) is the total number of unique IP addresses searching 10-K, 10-Q, and 8-K (10-K) filings. Our sample runs from January 2003 to December 2014.[16]

It is important to note that there are other ways for investors to access financial filings, such as a firm's investor relations website and Yahoo! Finance. Data vendors such as Bloomberg and FactSet also provide investors with access to these financial statements. As a

[13]First, following Lee, Ma, and Wang (2015), we exclude the searching records of those users who download more than 50 unique firms' filings in one day. The user is identified by their unique IP address. Second, following Ryans (2017) and Drake, Roulstone, and Thornock (2015), we remove log records that reference an "index" (idx=1), as index pages only provide the links to filings rather than the filings themselves. Third, following Ryans (2017), we keep the request records with successful document delivery (code=200). We then further exclude the search records of users who make more than 25 filing requests per minute or more than 500 requests per day, or with more than three unique CIKs searching per minute. Finally, we only keep one search record for a specific filing (unique accession number) to each user in a given day. This step is to avoid duplicated records due to users viewing the same document multiple times, a particular concern after the adoption of XBRL filing in 2009. For users who view the financial reports of XBRL-adopted firms in interactive data format, every click on a different footnote will generate a new search record, although it references the same document.

[14]Further details of the EDGAR index files can be found at https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm

[15]We define the 10-K category as the filing type in "10-K", "10-K/A", "10-K405", "10-K405/A", "10-KSB", "10KSB", "10-KSB-A", "10KSB/A", "10-KT", "NT 10-K", and "10-KSB40"; the 10-Q category as the filing type in "10-Q", "10-Q/A", "10QSB", "10-QSB", "10QSB-A", and "NT 10-Q"; the 8-K category as the filing type in "8-K" and "8-K/A"; the insider category as the filing type in "SC 13G", "SC-13D", "SC 13G/A", "SC 13D/A", "3", "4", and "5"; the registration category as the filing type in "S-1", "S-1/A", "S-3", "S-3/A", "S-3ASR", "424B5", "424B4", "424B3", "424B2", and "FWP"; and the proxy category as the filing type in "DEF 14A", "DEF 14C", "DEFA14A", "DEFM14A", "DEFR14A", and "DEFM14C".

[16]There are significant gaps in the data between September 2005 and May 2006, due to lost or corrupt log file. As a result, we exclude these months from our sample in our analysis.

result, our analysis of the EDGAR server log cannot capture all the views/downloads that the entire universe of investors are conducting on company filings. However, the EDGAR server still possesses several advantages over other information sources. First, it is questionable that investors primarily use the company website to retrieve SEC filings. As an example, Monga and Chasan (2015) quote General Electric (GE) CFO Jeffrey Bornstein, who noted that GE's 2013 annual report was downloaded from their investor relations website just 800 times.[17] However, for the same annual report, the EDGAR logs record 21,987 (4,325) downloads in the year (two months) following its filing. Second, other sources of company information often condense income-statement and balance-sheet information into pre-specified bins. As a result, some critical components of firms' financial information may be misrepresented. Third, many important accounting information such as information regarding operating lease is only available from annual reports' footnotes, not from a Bloomberg terminal or the Yahoo Finance web page. Finally, investors could better assess a firm's future prospects by reading the qualitative information contained in 10-K filings, which is not freely available in these data consolidators (Loughran and McDonald (2011)).

We obtain monthly stock returns from the Center for Research in Security Prices (CRSP), and annual accounting data from Compustat. Our sample of stocks starts with all common stocks traded on the NYSE, Amex, and NASDAQ. We adjust the stock returns by delisting. If a delisting return is missing and the delisting is performance-related, we set the delisting return at -30% (Shumway (1997)). We remove stocks with month end price less than $3.

We use standard control variables in our empirical analysis. *Size* (LnME) is defined as the natural logarithm of market capitalization at the end of June in each year. *Book-to-market ratio* (LnBM) is the most recent fiscal year-end report of book value divided by the market capitalization at the end of calendar year t-1. Book value equals the value of common stockholders' equity, plus deferred taxes and investment tax credits, and minus the book value of preferred stock. *Momentum* (Mom) is defined as the cumulative holding-period return from month t-12 and t-2. We follow the literature by skipping the most recent month's return when constructing the *Momentum* variable. The *short term reversal measure* (REV) is the prior month's return. *Turnover12* is the monthly trading volume over shares

[17] https://www.wsj.com/articles/the-109-894-word-annual-report-1433203762.

outstanding, averaged from the past 12 months. Since the dealer nature of the NASDAQ market makes its turnover difficult to compare with the turnover observed on NYSE and AMEX, we follow Gao and Ritter (2010) by adjusting the trading volume for NASDAQ stocks.[18] *Institutional ownership* (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by total shares outstanding. *Idiosyncratic volatility* (IVOL) is the standard deviation of the residuals from the regression of daily stock excess returns on the Fama and French (1993) three-factor returns within a month (Ang, Hodrick, Xing, and Zhang (2006)). Institutional ownership data of stocks are available from the Thomson Reuters (formerly CDA/Spectrum) Institutional Holdings database (13F). *Coverage* is the log one plus the number of analysts following a firm. Both the analyst coverage and recommendation data are from I/B/E/S. We get the filing size and number of words of the 10-Ks for all publicly-traded firms from WRDS SEC Analytics.

Finally, we obtain stock lendable supply (lendable shares divided by shares outstanding) and stock lending costs from the Markit Securities Finance (formerly Data Explorer) database.[19] We use the Markit provided *DCBS* score (Daily Cost of Borrowing Score) to measure short selling costs. DCBS is a score from 1 to 10 created by Markit using their proprietary information. This score is intended to capture the cost of borrowing the stock: A score of 1 represents the cheapest to short and 10 represents the most difficult.

## 3.2 Summary Statistics

Panel A of Table 1 displays the time-series average of the cross-sectional means and standard deviations of the variables for the full sample. The average number of unique IPs searching for all six types of SEC filings of a firm is 155 in a month. The cross-sectional standard deviation is 317, indicating a large cross-sectional variation among firms. Consistent with Drake, Roulstone, and Thornock (2015), the annual report 10-K is the most frequently searched type of SEC filings, with an average of 60 IPs requesting it in a month. IPs searching for 10-Q and 8-K are relatively less frequent. The average institutional ownership

[18]Specifically, we divide NASDAQ volume by 2.0, 1.8, 1.6, and 1.0 for the periods before February 2001, between February 2001 and December 2001, between January 2002 and December 2003, and after January 2004, respectively.

[19]See Saffi and Sigurdsson (2010) for a detailed account of Markit equity lending database.

in our sample is 55%, reflecting the rapid growth of assets managed by institutional investors during our sample period. The remaining summary statistics are well known and do not require additional discussion.

Panel B reports the pairwise rank correlation among our variables. The three IP variables are highly correlated. This is expected as periodic accounting reports consist of the largest fraction of EDGAR search requests. The number of IPs is also positively correlated with firm size, analyst coverage, and turnover, suggesting that firms with high investor visibility and attention have more EDGAR users. The number of IPs is negatively correlated with stock idiosyncratic volatility. However, this is mainly due to the size effect: small firms with high return volatility attract less EDGAR searching. As will be explained later, once we control for firm size, the number of IPs becomes positively correlated with idiosyncratic volatility, potentially because the incentives of acquiring information are greater when stock price is noisier (Grossman and Stiglitz (1980)).

Figure 1 plots the average number of IPs searching for EDGAR filings in each calendar month. The average is first calculated across stocks within a particular year-month and then averaged across all years. As we can see, there is no large seasonal variation for IP_total. The number of IPs searching for 10-Ks do spike during March and April. This could be explained by more investors searching for 10-Ks during earnings season as most public firms file annual report in these two months. In our subsequent analysis, we design tests to rule out the alternative explanation that our result is simply driven by earnings announcement.

### 3.3 Cross-sectional Determinants of Number of IPs

Theories of endogenous information acquisition suggest that information acquisition activity is a function of both the cost of acquiring information and the benefits of trading on acquired information. In order to isolate investors' expected benefits from information acquisition activity, we need a model of expected information acquisition activities. To this end, we develop and implement a simple characteristics-based model of expected information acquisition, and identify the discrepancies between the realized and expected level of information acquisition. Calculating these discrepancies requires proxies for information acquisition and firm characteristics useful in estimating the expected level of information acquisition activities.

Our proxy for information acquisition activity is the number of unique IP addresses searching for EDGAR filings for each firm in a given month. To mitigate data mining concerns, we use three measures capturing information acquisition activities for different types of SEC filings. IP_total is the total number of unique IPs searching for all types of SEC filings, and IP_funtl (IP_10K) is the total number of unique IPs searching for 10-K, 10-Q and 8-K (10-K) filings. Our choice of firm characteristics is guided by information acquisition theories. Specifically, we hypothesize that firms with higher visibility and investor attention would attract more information acquisition, as these firms are more accessible in investors' minds. We also conjecture that the strength of firms' information environments would affect information acquisition, although the direction of the effect is not clear. On one hand, firms with abundant public information will be less costly to analyze, so we expect information acquisition to increase with the quality of a firm's information environment. On the other hand, a better information environment also means that the stock is less likely to be mispriced, so investors' incentives to acquire additional information will be reduced. Finally, we expect investors to have stronger incentives to acquire information about firms with higher valuation uncertainty. Following prior literature, we use firm size as a proxy for investor visibility, trading volume as a proxy for investor attention (Gervais, Kaniel, and Mingelgrin (2001); Barber and Odean (2007)), analyst coverage as a proxy for information environment[20] (Hong, Lim, and Stein (2000)), and idiosyncratic volatility as a proxy for valuation uncertainty (Zhang (2006)).

We calculate the abnormal number of IPs by fitting monthly cross-sectional regressions of the raw number of IPs to isolate the components of the number of IPs not attributable to firms' size, turnover, analyst coverage, and idiosyncratic volatility. To mitigate the effect of outliers, we use the log of one plus the number of IPs when estimating the abnormal number of IPs for firms. Specifically, we calculate the abnormal number of IPs for firm $i$ in month $t$

[20] Another motivation for including analyst coverage is that according to Lee and So (2017), analyst coverage contains information about future stock return. By including analyst coverage as a regressor, any expected return information embedded in the number of IPs will be incremental to that contained in analyst coverage proxies.

by estimating the following cross-sectional regression[21]:

$$Log(1 + IP_{i,t}) = \beta_0 + \beta_1 LnME_{i,t} + \beta_2 Coverage_{i,t} + \beta_3 Turnover12_{i,t} + \beta_4 IVOL_{i,t} + \epsilon_{i,t} \quad (1)$$

where LnME is the log of market capitalization, Coverage is the log of one plus analyst coverage, Turnover12 is the monthly turnover averaged over the past 12 months, and IVOL is the daily idiosyncratic volatility calculated following Ang, Hodrick, Xing, and Zhang (2006). We define the abnormal number of IPs for each firm-month as the regression residuals from equation (1). We use the notation AIP to refer to the abnormal number of IPs, where higher values correspond to firms that have greater number of IPs searching for their SEC filings given their size, trading volume, analyst coverage, and volatility.

Table 2 reports the time-series average coefficients and Fama-MacBeth *t*-statistics from estimating equation (1). The three panels correspond to three different measures of IPs as dependent variables. To see the improvement of $R^2$, we add the explanatory variables one by one from Column (1) to Column (9). Consistent with our hypothesis, information acquisition activities increase with firm size (*t*-stat=69.44), as larger firms are more visible to investors. Size alone explains 40% of the cross-sectional variation of the number of IPs. Columns (2) and (3) show that information acquisition increases with the strength of firms' information environments and investor attention, proxied by analyst coverage and turnover, respectively. Column (4) further shows that the number of IPs increases with return volatility after controlling for firm size. This finding suggests that investors' demand for information is larger for firms with more uncertain value. Column (4) also shows that these four firm characteristics explain 55% of the cross-sectional variation of the number of IPs on average. The results are similar in Panels B and C, where the dependent variables are IP_fundl and IP_10K, respectively.

The four firm characteristics used in equation (1) were selected based on theories and parsimony, and may therefore omit other firm characteristics that drive variation in the expected level of information acquisition activity. For example, investors may be attracted

[21] We run pure cross-sectional regression in the first stage so that the abnormal number of IPs (regression residuals) we use later on does not have look-ahead bias.

to firms with extreme past returns and glamour characteristics (Barber and Odean (2007)). In addition, firms included in S&P 500 index may attract more attention from investors. To examine the explanatory power of other firm characteristics, we add the stock's past 12-month return, book-to-market ratio, institutional ownership, a dummy indicating whether it belongs to S&P 500 index, and a dummy indicating quarterly earnings announcement month iteratively from Column (5) to Column (9). The results suggest that more investors search for EDGAR filings when the firm has performed poorly over the past year, has high B/M ratio, is held by more institutional investors, belongs to S&P 500 index, and is announcing earnings. However, adding these additional characteristics improves the average $R^2$ of equation (1) by only 3 percentage points, suggesting the limited incremental explanatory power of these additional characteristics. In the robustness test below, we show that the inclusion of other firm characteristics in equation (1) does not significantly affect the return predictability of AIP.

As there might be a nonlinear relationship between the abnormal number of IPs and firm characteristics, we further look at average stock characteristics across decile portfolios sorted by abnormal number of IPs searching for 10-Ks (AIP_10K). Higher (lower) deciles correspond to firms with abnormally high (low) number of IPs. Panel C of Table 1 reports the time-series average of the cross-sectional mean values of each variable for each decile. First, the observation counts show that each month there are about 330 firms in each decile, suggesting that our measure of information acquisition is available for a broad cross-sectional sample of 3,300 firms per month. Second, the table shows that AIP is positively correlated with the raw number of IPs searching for EDGAR filings. Third, AIP is, by construction, uncorrelated with firm size, analyst coverage, turnover, and volatility, although middle portfolios are slightly larger in terms of size and turnover. Finally, the panel shows that firms in the extreme deciles have lower institutional ownership and are more likely to be value stocks.

# 4 Information Acquisition and Future Stock Returns

Theories of endogenous information acquisition predict that when investors expend effort and time to acquire information, they must perceive some benefits of utilizing such information. Hence a key hypothesis in this paper is that costly information acquisition activities reveal investors' perceptions of expected payoffs. Although in theory, the information content could be either positive or negative, in reality we expect firms with intensive information acquisition activities to have positive performance due to short-sale constraints. In addition, the positive return predictability of AIP should be stronger for smaller firms with more binding short-selling constraints. In this section, we examine the relation between information acquisition and future returns using both portfolio sorts and the Fama-MacBeth regression.

## 4.1 Portfolio Sorts

In this section, we show that stocks sorted based on their abnormal numbers of IPs generate significant return spreads. We conduct the decile portfolio sorts as follows. At the end of each month, we sort stocks into deciles by their AIP. We then compute the average return of each decile portfolio over the next month, which provide a time series of monthly returns for each decile. We use these time series to compute the average excess return of each decile over the entire sample. As we are most interested in the return spread between the two extreme portfolios, we also report the return to a long–short portfolio (i.e., a zero-investment portfolio that longs the stocks in the highest AIP decile and shorts the stocks in the lowest decile).[22]

Table 3 reports the average monthly excess return of each decile portfolio. Panel A reports the equal-weighted portfolio return, and Panel B reports the value-weighted return. The three columns in each panel correspond to sorting based on the abnormal number of IPs searching for three different types of SEC filings. Panel A shows a strong positive relation between AIP and future returns, regardless of which IP variables are used. For sorts based on AIP_total, firms in the highest decile of AIP outperform the firms in the lowest decile by 71 basis points per month on an equal-weighted basis ($t$-stat=3.18). The results are stronger when we do the portfolio sorts based on AIP_funtl and AIP_10K.[23] Specifically, the

[22]The advantage of conducting analysis at monthly frequency is that it is easier to correct for known determinants of expected returns (size, book-to- market and momentum) using factor regressions, and the estimates of alpha thus obtained have a clear interpretation in terms of asset pricing theory.

[23]The larger return spread based on IPs searching for 10-K compared with IPs searching for other types

high-minus-low monthly return spread is 100 basis points ($t$-stat=4.70) based on AIP_10K, which corresponds to an annualized return of 12%.[24] The result suggests that information acquisition activities aggregated across EDGAR users reveal an economically large source of predictable return across firms.

The return spread of the high-minus-low-AIP portfolio is considerably smaller and less significant when returns are value weighted. The high-minus-low return is only about 30 basis points per month, and mostly insignificant. This is consistent with our prior that for big firms with less binding short-sale constraints, the information content embedded in EDGAR searching could be either positive or negative. Investors could take (less costly) short positions on big stocks to benefit from the negative information they obtained through EDGAR filings. This implies that, ex-ante, we do not have a clear *directional* prediction of a relationship between the abnormal number of IPs and future returns.

Table 4 examines the relation between the abnormal number of IPs and firms' future return after controlling for the portfolios' exposure to standard asset-pricing factors. The table reports the monthly Carhart (1997) four-factor alpha for decile portfolios sorted on AIP, as well as the long/short hedge portfolio. The four-factor alpha is the intercept from a regression of the portfolio's excess return on the contemporaneous excess market return (MKTRF), the size factor (SMB), the value factor (HML), and the momentum factor (UMD). Panel A shows that AIP predicts a strong positive return spread cross-sectionally for equal-weighted portfolios. The four-factor alphas of the long/short portfolio range from 59 to 80 basis points per month and are highly significant. Moreover, in the case of AIP_10K, the alphas largely come from the long leg. The lowest AIP_10K decile portfolio generates a four-factor alpha of about -28 basis points ($t$-stat=-2.33), and the highest AIP_10K decile generates a positive alpha of 52 basis points ($t$-stat=2.92). Panel B of Table 4 shows the

of SEC filings is consistent with information acquisition theories. A firm's annual report is among the lengthiest and most difficult-to-read SEC filings. Annual reports contain detailed annual operating and financial performance and metrics, suggesting that digesting these reports requires a large amount of effort and time from investors. Compared with 10-Ks, 10-Q and 8-K files are usually much shorter and easier to digest, and investors driven to these types of filings are more likely to respond to current news events, and less likely to reflect a deliberate information acquisition choice. Given the substantially higher cost of acquiring and analyzing 10-Ks, the expected benefits perceived by investors should also be larger, which is consistent with our results.

[24] A caveat is that the large abnormal returns based on EDGAR searching data is only hypothetical. Investors without access to the real-time EDGAR searching data cannot trade on the information.

portfolio alphas for value-weighted returns. Again, we find the results are generally weaker, both economically and statistically. The four-factor alpha of the long/short portfolio ranges from 12 to 41 basis points, which are either insignificant or only marginally significant.

To emphasize the importance of measuring the abnormal level of information acquisition activity when uncovering expected return information, we conduct a parallel portfolio test when ranking firms into deciles based on the raw number of IPs searching for EDGAR filings, as shown in Table A1 in the Online Appendix. Panel A reports the equal-weighted excess returns and Panel B reports the equal-weighted four-factor alphas. The results show that the raw number of IPs is not significantly correlated with firms' future returns, regardless of which IP variable we use. The monthly four-factor alpha of the long-short portfolio based on the raw number of IPs ranges from -20 to 9 basis points, which are never significant. The lack of significant predictive power of the raw number of IP suggests that it is important to control for the expected level of information acquisition activities when uncovering investors' expected payoffs.[25]

## 4.2 Robustness Checks and Alternative Implementations

In Table A2 in the Online Appendix, we examine the robustness of our portfolio sorts. For brevity, we focus on the sorts based on AIP_10K. The first row shows the return spread when returns are weighted by past month gross return, as suggested by Asparouhova, Bessembinder, and Kalcheva (2013). The gross-return-weighted return spread is 1.1% ($t$-stat=5.16). Rows (2) and (3) show that our results barely change when we subtract the characteristic-matched portfolio (Daniel, Grinblatt, Titman, and Wermers (1997)) or the corresponding industry return from stock return. This suggests that the nature of the information contained in EDGAR searching activities is mostly firm-specific. In the fourth row, we augment the Carhart (1997) four-factors with the Pástor and Stambaugh (2003) liquidity factor. The Pástor and Stambaugh (2003) five-factor adjusted alpha is 0.80% ($t$-stat=4.23) for the equal-weighted portfolio and 0.35% ($t$-stat=1.78) for the value-weighted portfolio. The fifth row shows that our results hold when we use the Fama and French (2015) five factors to calculate

[25] The rational is that large raw number of IPs could be driven by low cost of information acquisition, rather than high expected benefits.

alphas, with a monthly return spread of 0.69% ($t$-stat=3.36) for the equal-weighted portfolio. The sixth row shows that our results still hold when we use the Stambaugh and Yuan (2016) mispricing factor model to compute alpha. The portfolio generates an equal-weighted alpha of 0.89%($t$-stat=4.42) and value-weighted alpha of 0.27%($t$-stat=1.35). Using Hou, Xue, and Zhang (2015)'s Q-factor model also does not affect our results, as shown in the seventh row. The eighth row of Table A2 shows that our results survive when we exclude stocks whose market capitalizations are in the bottom quintile of the NYSE size distribution. Again, the long-short strategy based on AIP generates a monthly four-factor alpha of 0.52% ($t$-stat=2.58) and 0.28% ($t$-stat=1.35) when returns are equal-weighted and value-weighted, respectively. The ninth row reports the long-short alphas if we implement a six-months interval between when we sort stocks and when we measure strategy returns. The equal-weighted alpha is quite substantially reduced in this case, but nonetheless still significant, with an equal-weighted four-factor alpha of 0.53% ($t$-stat=2.23). The tenth and eleventh rows show that the long-short portfolio generates significant alpha in two subperiods: one from 2003 to 2008 and another from 2009 to 2014. In fact, the return predictability of AIP appears to be stronger in the recent period (monthly alpha of 1.07% vs. 0.62%), consistent with the fact that the average 10-Ks have become lengthier and more costly to analyze over time (Cohen, Malloy, and Nguyen (2018)). The last row shows that the portfolio alpha is not affected by removing the financial crisis period (year 2008 and 2009) from our sample.

Our results are not sensitive to the specific model of calculating the abnormal number of IPs, as shown in Table A3 in the Online Appendix. The first row shows that the long-short portfolio based on AIP_10K calculated using model (9) of equation (1) generates a four-factor alpha of 0.67% ($t$-stat=3.92) for the equal-weighted portfolio. In the second row, we include the square terms of the four firm characteristics when calculating AIP to account for the nonlinear relation between number of IPs and firm characteristics. The four-factor alpha is 0.69% and 0.55% for the equal- and value-weighted portfolio, respectively. In the third row, we add the lagged log number of IPs in equation (1) when calculating AIP, and the alpha is still significant. This specification is equivalent to using the innovation in number of IPs to predict returns, so the return predictability of AIP is unlikely explained by any (omitted) persistent firm characteristics.

In Table A4 in the Online Appendix, we show that a positive relation between AIP and returns holds for change-based specifications, which further mitigates concerns that the return predictability of AIP is driven by an omitted firm-fixed effect not controlled for in our model of AIP. The long-short portfolio sorted on the change of AIP relative to its 12-month moving average generates an equal-weighted four-factor alpha of between 0.63% and 0.88% per month and are still highly significant.

In Table A5 in the Online Appendix, we examine the within-industry return predictability of AIP_10K, as defined by the Fama-French 12 industry classification. In the end of each month, we sort all stocks within each industry into quintile portfolios and calculate the Carhart (1997) four-factor alpha of the long-short portfolio. AIP_10K generates significant and positive abnormal returns for 10 out of 12 industries, with a monthly alpha ranging from 0.48% for financial industry to 1.06% for energy industry. In sum, we conclude that the return predictability of AIP is robust and pervasive across the entire universe of US equity market.

## 4.3 Cross-sectional Heterogeneity

### 4.3.1 The Role of Firm Size and Limits to Arbitrage

The results in section 4.1 show that the long/short portfolio alpha is only significant for equal-weighted returns, but not for value-weighted returns. To take a closer look at the role of firm size, we report the portfolio sorting results based on AIP by size quintiles in Table A6 in the Online Appendix. For each month, we group all stocks into size quintiles based on the NYSE size breakpoints. We then *independently* sort stocks into quintiles based on AIP_10K. The table reports the Carhart (1997) four-factor alpha for the 25 portfolios: equal-weighted returns in Panel A and value-weighted returns in Panel B. We also report the alpha for each size quintile of the high-minus-low-AIP portfolios. The result shows that the return predictability of AIP is strongest among the smallest size quintile, but is not limited to only the microcap stocks. The high-minus-low AIP portfolio generates a significant four-factor alpha of approximately 0.4% among the three middle-sized quintiles, both equal-weighted and value-weighted. The alpha is insignificant in the largest size quintile.

The findings in Table A6 show that the return predictability of AIP is more pronounced for small firms than for large firms, which could be explained by two non-mutually exclusive channels. The first is that the latent information embedded in information acquisition activities could be either positive or negative when shorting is less costly. Given that large firms have fewer short-sale impediments, the direction of return predictability for large firms is more ambiguous. An independent channel that could reinforce the weak return predictability among these stocks is that whatever information is contained in the EDGAR searches, they are impounded into stock prices more quickly due to less trading frictions (e.g., liquidity and non-fundamental volatility) among large firms. We next explore how the return predictability of AIP varies across firms with different level of arbitrage frictions and short-sale constraints.

Following the literature, we investigate the role of three general limits-to-arbitrage measures: idiosyncratic volatility (Stambaugh, Yu, and Yuan (2015); Pontiff (2006)), residual institutional ownership (Nagel (2005)), and residual analyst coverage (Hong, Lim, and Stein (2000)). In addition, to substantialize the short-sale constraints argument in particular, we use the lendable supply and lending fee measure provided by Markit to measure short-selling costs. At the end of each month, we sort all stocks into terciles based on each limits-to-arbitrage and short-sale constraints variable X except lending fee, for which we sort into two groups based on whether a stock's DCBS score is above or below 2[26]. We then *independently* sort stocks into quintiles based on the abnormal number of IPs searching for 10-Ks. Table 5 displays the equal-weighted four-factor alphas of the lowest and highest AIP portfolios in the lowest and highest X groups. Consistent with the limits-to-arbitrage predictions, the alpha of the high-minus-low portfolio is more pronounced among stocks with higher idiosyncratic volatility, lower institutional ownership, and less analyst coverage. For example, the high-minus-low portfolio generates 1.24% ($t$-stat=4.44) monthly alpha for high-volatility stocks, and only 0.23% ($t$-stat=1.76) for low-volatility stocks. The difference

[26]This treatment follows the short selling literature. Stocks with a DCBS score less than or equal to 2 are usually cheap to borrow and are called "general collateral". Stocks with DCBS larger than 2 are more costly to short and are called "special" stocks.

of alphas between stocks with high and low idiosyncratic volatilty is 1.01% ($t$-stat=3.74). The results based on measures of short-sale constraints also support our hypothesis: the alpha of the high-minus-low portfolio is more pronounced among stocks with lower lendable supply and higher lending fees. For example, the high-minus-low portfolio generates 1.14% ($t$-stat=2.85) monthly alpha for high-lending fee stocks, and only 0.26% ($t$-stat=1.39) for low-lending fee stocks. The difference of alphas between stocks with high and low lending fee is 0.88% ($t$-stat=2.07).

### 4.3.2 Variation in the Complexity of Financial Filings

The underlying hypothesis in this paper is that investors' costly information acquisition activity should be positively related to the expected payoff from using the information. If this is true, we would expect the payoff to be larger when the information acquisition/processing cost is higher. To test this prediction, we use the complexity of a firm's annual report as a proxy for the cost of information acquisition/processing. The idea is intuitive, as more complex filings require more effort and time for investors to process and digest. Following the recent literature (Loughran and McDonald (2014); You and Zhang (2009)), we use the natural log of the gross 10-K file size (complete submission text file) and the number of words contained in 10-K as a proxy for filing complexity.[27]

To this end, we first obtain the file size and number of words contained in firms' most recent 10-Ks. However, as big firms have more business lines and more diverse sets of operations, they would naturally have lengthier and larger 10-K filings.[28] To remove the confounding effect of firm size, we regress the logarithm of filing size and number of words on the logarithm of firms' market capitalizations, and use the regression residual as our proxy of filing complexity. At the end of each month, we sort all stocks into terciles based on either the residual file size or the residual word count. We then *independently* sort stocks into quintiles based on AIP_10K. Table 6 shows the equal-weighted four-factor alphas of the

[27]Loughran and McDonald (2014) report that the 10-K file size is positively associated with high return volatility in a one-month period following 10-K filings, supporting the use of file size as a proxy for the linguistic complexity of 10-K disclosure. You and Zhang (2009) find that investors' underreaction to information contained in 10-Ks is stronger for 10-Ks with larger numbers of words.

[28]The rank correlation is 0.34 between 10-K file size and firm size, and 0.40 between word count and firm size.

lowest and highest AIP_10K portfolios in the highest and lowest groups of filing complexity. Consistent with theories of costly information acquisition, the alpha of the high-minus-low portfolio is indeed economically larger and more significant for firms with more complex 10-Ks. For example, the high-minus-low AIP_10K portfolio generates 1.24% ($t$-stat=4.88) monthly alpha among firms with the largest residual file sizes, and 0.66% ($t$-stat=3.30) among firms with the smallest file sizes. The difference of alphas between stocks with large and small file size is 0.58% ($t$-stat=2.29). The result is similar when we use the word count in 10-K as a proxy for the complexity of financial filings. Overall, the evidence supports our hypothesis that the more costly information acquistion/processing is, the larger the expected payoff revealed by the equilibrium amount of information acquisition activity.

#### 4.3.3 Cross-sectional Heterogeneity at IP Level

We next examine the return predictability for different types of IP. Although we do not have the exact identity of IPs, we can nevertheless track the behavior of each IP given its uniqueness, such as the type of filings it requests and the timing of search.

The first cross-sectional heterogeneity we look at is whether the IP searched both the current and historical 10-K filings. This test could help distinguish the information acquisition story from the news-announcements explanation. On one hand, if the return predictability of AIP is entirely driven by news announcements, the result should be stronger when we focus on IPs only searching for current 10-K filings as investors rush to understand the implications of current news on firm value. On the other hand, although historical filings are unlikely to provide any news to investors, they still make up an important component of the information mosaic assembled by investors, and thus should be valuable to acquire.[29] To test this, for each stock-month, we compute the number of unique IPs that searched only the current 10-Ks and those searched both the current and historical 10-Ks. We define current (historical) 10-Ks as those filed after (before) the most recent 10-K filing date. We then sort stocks into deciles based on the abnormal number of IPs within each category and report the

[29]Drake, Roulstone, and Thornock (2016) document the value of historical accounting reports. Cohen, Malloy, and Nguyen (2018) show that change in firms' reporting practices conveys an important signal about future firm operations, which can only be obtained after comparing current reports to historical reports.

results in Table 7. Interestingly, rows (1) and (2) show that the return predictability of AIP is stronger when we isolate IPs searched both the current and historical 10-Ks. Specifically, the alpha of the high-minus-low portfolio generates 0.61% ($t$-stat=3.08) monthly alpha for IPs that searched only the current 10-Ks, while that figure is 1.00% ($t$-stat=5.28) for IPs that searched both the current and historical 10-Ks. The difference of alphas between the two groups is 0.39% ($t$-stat=2.53). As analyzing information in historical 10-Ks is more costly and more indicative of deliberate information acquisition, this evidence strongly supports the endogeneous information acquisition theories.

The second dimension we look at is the timing of search conducted by the IP, that is, whether the search is conducted at day time or night time. Under the assumption that nighttime searches should mostly come from retail investors, if we still find similar return predictability of nighttime IP, the evidence would suggest that our results are not entirely driven by institutional investors and at least some retail investors are sophisticated. To test this, for each stock-month, we compute separately the number of unique IPs that searched the firm's 10-Ks in night time (6pm of day $t$ to 6am of day $t+1$) and day time (6am of day $t$ to 6pm of day $t$).[30] Rows (3) and (4) report the monthly alphas of long-short portfolios sorted on nighttime and daytime IPs, respectively. The result shows that even if we focus on those IPs most likely from retail investors, the long-short portfolio still generates a significant four-factor alpha of 0.82% ($t$-stat=4.71) per month, which is very similar to the result using all IPs. The evidence is consistent with several recent studies showing that the aggregate trading activities of retail investors are informative about future stock returns and earnings news (Kelley and Tetlock (2013); Boehmer, Jones, and Zhang (2017)).

## 4.4 Fama-MacBeth Regression

We now test the return predictability of AIP using the Fama and MacBeth (1973) regression methodology. One advantage of this methodology is that it allows us to examine the predictive power of AIP while controlling for other known predictors of cross-sectional stock returns. This is important because, as shown in Table 1, AIP is correlated with some

[30]If a IP searched 10-Ks both in day time and night time within a month, we classify it as a daytime IP so that we can cleanly identify those IPs becoming active only in nighttime.

of these predictors. We conduct the Fama-MacBeth regressions in the usual way. For each month, starting in February 2003 and ending with December 2014, we run the following cross-sectional regression:

$$Ret_{i,t+1} = \beta_0 + \beta_1 AIP_{i,t} + \gamma X_{i,t} + \epsilon_{i,t} \tag{2}$$

where $Ret_{i,t+1}$ is the return of stock $i$ in month $t+1$, $AIP_{i,t}$ is the abnormal number of IPs searching for firm $i$'s EDGAR filings in month $t$, and $X$ is a set of control variables known to predict returns, including the natural logarithm of the book-to-market ratio (LnBM), the natural logarithm of the market value of equity (LnME), returns from the prior month (Rev), returns from the prior 12-month period excluding month $t-1$ (Mom), institutional ownership (IO), and idiosyncratic volatility (IVOL) and past 12-month turnover (Turnover12).

Table 8 reports the time-series averages of the coefficients of the independent variables, and the $t$-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. Columns (1) to (3) show the coefficient of AIP without any other return predictors. The coefficients of all three AIP variables are positive and significant at 1% level. This is consistent with our portfolio sorting results in which stocks with abnormally large numbers of IPs searching for their EDGAR filings have higher expected returns. In Columns (4) to (6), we add the usual controls including firm size, book-to-market ratio, past 1-month returns, and past 12-month returns. The coefficients of AIP barely change, and retain their strong predictive power. In Columns (7) to (9), we further add institutional ownership, turnover, and idiosyncratic volatility to the regression model, and AIP still positively predicts future returns. The economic magnitude is substantial. The average difference of AIP_10K between the lowest and highest decile portfolio is 2.39, which implies a monthly return spread of 105 basis points between these two extreme deciles. The magnitude estimated from the Fama-MacBeth regression is in line with our portfolio sorting results. For the control variables, the signs of the coefficients are consistent with those reported in the previous literature, except for momentum, which attracts an insignificant coefficient.[31] Due to the

[31] This is due to the 2009 momentum crash (see Daniel and Moskowitz (2016)). The coefficient of momentum becomes positive once we exclude the year 2009 from our sample.

short and recent sample period, however, the coefficients of some control variables are not significant.

### 4.5 Predicting Earnings Annoucement Returns

The strong return predictability of AIP suggests that the information contained in AIP is not immediately incorporated into stock prices, which is consistent with models of costly information acquisition where stock prices are only partially revealing. An important implication is that AIP should positively predict returns around earnings announcement when the fundamental information embedded in AIP is disclosed to the market.

To test, we extract quarterly earnings announcement dates from I/B/E/S and calculate three-day announcement period abnormal returns (CAR(-1,+1)) adjusted by returns on CRSP value-weighted index or characteristics-matched (size, book-to-market and past 1-year return) portfolio. We then run Fama-MacBeth regression of the earnings announcement CAR(-1,+1) on AIP and other control variables that are observed one month before the earnings announcement date. Table A7 in the Online Appendix shows that AIP also positively predict earnings announcement returns, and the strongest predictability is obtained for AIP_10K. The economic effect is substantial. For example, the coefficient on AIP_10K is 0.0036 ($t$-stat=2.74) when the dependent variable is market-adjusted CAR(-1,+1). This suggests that return difference between two extreme AIP decile portfolios during the three-day earnings announcement window is 0.86%, compared to a monthly return difference of 1.05% including all trading days. This means that about 27% of abnormal returns following AIP is concentrated on the three-day window around quarterly earnings announcement, which represents only 5% of all trading days. The fact that abnormal return is concentrated on a few information announcement days makes our findings difficult to square with risk-based explanations (La Porta, Lakonishok, Shleifer, and Vishny (1997); Engelberg, McLean, and Pontiff (2018)). We find similar results using DGTW-adjusted CAR as dependent variable, as shown in Columns (4) to (6) of Table A7.

# 5 Channels

The underlying hypothesis of this paper is that information acquisition activity embeds expected return information because with short-selling constraints, investors would rationally allocate greater effort to analyzing firms that are potentially undervalued. As mispricing implies the separation of stock prices from firms' fundamental value, there are two non-mutually exclusive channels through which investors can identify mispricing. The first channel is investors' costly information acquisition revealing their favorable expectations of firms' fundamental performances that are not fully priced in by the market. The second channel of investors identifying mispricing is by observing changes in stock prices that are not attributable to firms' fundamental changes. In this section, we test both channels.

## 5.1 Predicting Changes in Firm Fundamentals

We first test whether information acquisition via EDGAR reveals novel information about firms' fundamental performance. We use three measures of a firm's fundamental performance. The first is the change in quarterly Return-on-Assets (dROA) from four quarters ago, which takes into account of the seasonality of firms' operating performances. The second measure is the standardized unexpected earnings (SUE), defined as the change of quarterly earnings-per-share (EPS) from four quarters ago scaled by stock prices 12 months ago. The third measure is the monthly forecast revision of analysts' consensus Earnings-per-Share (EPS) forecast (FREV) scaled by stock prices 12 months ago, which is a higher-frequency measure of firms' fundamental performances. We run panel regressions of dROA, SUE and FREV on lagged AIP, controlling for other firm characteristics that are associated with firms' fundamental performances, including size, book-to-market, past 12-month returns, analyst coverage, turnover, institutional ownership, idiosyncratic volatility, and lagged quarterly ROA. Since dROA and SUE are measured at quarterly frequency, we construct the AIP at quarterly frequency by averaging the monthly AIP within a quarter. We also control for time-fixed effects, and standard errors are double clustered by firm and time following Petersen (2009). If the return predictability of AIP is partially driven by its predictive power for firm fundamentals, the coefficient of AIP should be significantly positive.

Table 9 reports the results of predicting fundamental performance based on AIP. The dependent variable is the change in quarterly ROA from Columns (1) to (3), SUE from Columns (4) to (6), and analyst forecast revision from Columns (7) to (9). We show the predictability result for all three AIP measures. The coefficients of AIP are significantly positive for all three measures of fundamental performance, regardless of which AIP measures we use. The economic magnitude is non-trivial. For example, Column (3) shows that an interquartile increase in AIP_10K is associated with an increase of 0.22 percentage points in dROA, which is about 17% of the interquartile range of quarterly change in ROA. This finding suggests that information acquisition via EDGAR contains investors' private expectations of firms' future operating performances. It is worth noting that the predictability of AIP is obtained after controlling for other determinants of firms' fundamental performances. For example, the past 12-month returns strongly and positively predict changes in ROA and analyst forecast revision, while turnover and idiosyncratic volatility negatively predict fundamental performance. Overall, the test supports the first channel that the source of return predictability of AIP derives (partially) from investors allocating greater effort to firms with improving fundamentals.

## 5.2 Identifying Mispricing using Mutual Fund Outflows

A second channel through which mispricing could occur is exogeneous shock to stock prices that is not attributable to firm fundamentals. In this section, we use mutual fund outflow-induced selling pressure as an exogeneous shock to stock prices. Coval and Stafford (2007), Khan, Kogan, and Serafeim (2012), and Edmans, Goldstein, and Jiang (2012) find that mutual funds sell a firm's shares roughly in proportion to its portfolio weights when the funds are facing severe outflows. The forced selling behavior results in significant downward price pressure that persists for more than a year. This is a relatively exogenous and clean measure of underpricing as it is associated with who is selling – funds facing large investor redemptions – rather than what is being sold, and so is unlikely to be driven by (unobserved) changes in firms' fundamentals.

To that end, we construct a mutual fund outflow-induced fire sale measure for each stock following Edmans, Goldstein, and Jiang (2012), which reflects fund outflow expressed as a percentage of a stock's total dollar trading volume within a quarter. Figure A1 illustrates the magnitude and persistence of the effect of mechanically driven mutual fund fire-sale on stock prices. We define an "event" as a firm-quarter in which outflow falls below the 10th percentile value of the full sample. We then trace out the cumulative abnormal returns (CAR) over the CRSP equal-weighted or value-weighted index from 15 months before the event to 24 months after. Figure A1 shows that the price pressure effects from fire sale are both significant in magnitude and long-lasting, persisting for over a year. Equally important, consistent with the literature, they are temporary rather than permanent, with the price recovering by the end of the 24th month.

To test whether more investors start to acquire information on firms experiencing fire-sale induced underpricing, we examine the change in AIP following outflow-induced fire sale. Specifically, we run the following Fama-MacBeth regression:

$$dAIP_{i,q+1} = \beta_0 + \beta_1 Outflows_{i,q} + \beta_2 X_{i,q} + \epsilon_{i,q+1} \quad (3)$$

where $Outflow_{i,q}$ is the flow-induced fire sale measure calculated in accordance with Edmans, Goldstein, and Jiang (2012). Our dependent variable $dAIP_{i,q+1}$ is the within-firm change of AIP in quarter $q+1$ following mutual fund outflows. $X$ is a set of firm characteristics that may affect the change of AIP.

Table 10 reports the result using all three AIP measures. Columns (1), (3), and (5) show that the coefficients of "Outflows" are significantly negative without other controls, for all three AIP measures. The negative coefficient means that more IPs begin to search the SEC filings of firms that are underpriced due to exogenous shocks. Columns (2), (4), and (6) show that the relation between outflow-induced selling pressure and change in AIP is robust after controlling for a large set of firm characteristics.

In sum, by using mutual funds outflow-induced selling pressure to identify stock-level underpricing, our test also supports the second channel that part of the return predictability we document is attributable to investors allocating greater efforts to firms experiencing exogenous underpricing that is not attributable to fundamentals.

## 5.3 Anomaly-based Mispricing and Abnormal Number of IPs

The results from previous sections suggest that investors are able to identify undervalued stocks and rationally allocate more effort to these firms in the form of searching/processing their SEC filings. The question remains is if investors already have a sense of which firms are undervalued even before analyzing SEC filings, what is the incremental value of acquiring information through EDGAR? Our conjecture is that acquiring fundamental information through EDGAR could help investors identify truely mispriced stocks. For example, a value investor may have a sense of which stocks are potentially undervalued based on some valuation ratios such as book-to-market (B/M) or earnings-to-price (E/P) ratios, but firms with high B/M or E/P ratios are not all undervalued. To avoid "value trap", the investor may need to analyse in great detail the fundamental information contained in a firm's SEC filings, which could be useful to identify whether a stock is truly mispriced.[32] In this section, we provide empirical evidence supporting such a conjecture.

Specifically, we use the composite mispricing measure constructed by Stambaugh, Yu, and Yuan (2015) to identify mispricing. The composite mispricing measure is the average of the percentiles produced by 11 anomaly variables.[33] We first look at how investors' information acquisition activity via EDGAR vary across stocks with differential degree of mispricing. The result is reported in Panel A of Table 11, which shows the average abnormal number of IPs (AIP) across quintile portfolios sorted on the composite mispricing measure. Consistent with our hypothesis, there is significantly greater abnormal number of IPs searching for SEC filings of the most undervalued 20% of stocks than other stocks. In fact, for all three AIP measures, the mean value of AIP almost monotonically increases from the most overvalued to the most undervalued stocks. This result suggests that investors may get a sense of which stocks are worth investigating based on firm characteristics commonly associated with mispricing.

More importantly, in Panel B of Table 11, we show that accessing SEC filings through

[32]A "value trap" is a stock that appears to be cheap because the stock has been trading at low valuation metrics such as multiples of earnings, cash flow or book value for an extended time period. The trap springs when investors buy into such companies at low prices and the stock continues to languish or drop further. Identifying such firms require reading SEC filings so that investors could better understand the company's competitive environment, its ability to innovate, its ability to contain costs, and management by the executives.

[33]These 11 anomalies include net stock issues, composite equity issues, accurals, net operating assets, asset growth, Investment-to-Assets, distress, O-score, momentum, gross profitability and return on assets.

EDGAR could help investors identify truely mispriced stocks among those with similar mispricing characteristics. Specifically, we conduct an *independent* double sort based on a stock's composite mispricing measure and its AIP_10K. Panel B reports the equally-weighted four-factor alphas of the 5*5 double sorted portfolios. The result show that among the most undervalued quintile of stocks based on the composite mispricing measure, those with the lowest AIP have an insigificant monthly alpha of -0.05%. In sharp contrast, these undervalued stocks with the highest AIP have a monthly alpha of 1.05%. The monthly return difference between the high and low AIP stocks that appear similarly undervalued is 1.10% ($t$-stat=4.42). Panel C shows that the difference in the composite mispricing measure between low and high AIP stocks within the same mispricing quintile is close to zero in magnitude. Overall, the results support our conjecture that investors' costly information acquisition activity are getting compensated as it allows them to distinguish truely mispriced stocks from those sharing similar mispricing characteristics.

# 6 Alternative Explanations and Additional Analyses

In this section, we consider several alternative explanations for the return predictability of EDGAR searching activity, including firm events, breadth of ownership, media coverage, investor recognition, price pressure, and omitted risk factors. We also conduct additional analyses to shed further light on the underlying channels.

## 6.1 Alternative Explanations

### 6.1.1 Firm Events

EDGAR searching activity is positively related to information-rich firm events such as earnings/dividends announcements or analyst recommendation changes (Drake, Roulstone, and Thornock (2015)). Since an earnings surprise (recommendation changes) leads to post-earnings (recommendations) announcement drift (Bernard and Thomas (1989); Womack (1996)) and earnings/dividends announcement months are generally associated with positive stock returns (Lamont and Frazzini (2007); Hartzmark and Solomon (2013)), the return predictability of AIP could be driven by these announcements-related return predictability effects. As a robustness check, we add standardized unexpected earnings (SUE), an earnings-announcement month dummy (EAM), an analyst upgrade and downgrade event dummy, and a dividend month dummy (DM) in the Fama-MacBeth regression.[34] Columns (1) to (3) of Table A8 in the Online Appendix show that the coefficients on AIP are still highly significant.

To the extent that earnings/dividends/recommendations may not fully capture all firm events, we consider 8-K filings as a more comprehensive measure of firm-specific material events and add the log number of 8-K filings from previous month in the regression.[35] Columns (4) to (6) of Table A8 show that the coefficients on AIP barely change. Overall, we conclude that the information contained in AIP is not driven by firm events.

Another piece of evidence suggesting our result is not fully driven by firm events is provided in Table 3 of Loughran and McDonald (2017). They show that only 10.1% (21.6%) of 10-K requests over a 401-day window occur in the first week (month) after the filing date. Thus, the majority of EDGAR requests for 10-Ks is not clustered around earnings announcement days.

#### 6.1.2 Breadth of Ownership and Extreme Returns

Chen, Hong, and Stein (2002) show that reduction of the breadth of institutional ownership is a proxy for investor disagreement when short-sale constraints are binding for some investors. To the extent that breadth of ownership is positively correlated with the number of IPs searching for EDGAR filings, our result may be explained by breadth of ownership.

To the extent that investors are being attracted to stocks with extreme daily returns (Barber and Odean (2007)), our results could also be driven by the asset pricing effect

[34]SUE is a firm's standardized unexplained earnings, defined as the realized earnings per share (EPS) minus EPS from four quarters prior, divided by the standard deviation of this difference over the prior eight quarters. EAM is a dummy variable that equals one when a given firm announces quarterly earnings in the month. Upgrade (Downgrade) is a dummy equals one when there is an analyst recommendation upgrade (downgrade) in the previous month. DM is a dummy variable that equals one when there is an ex-dividend event in this month.

[35]Section 409 of the Sarbanes-Oxley Act of 2002 requires public companies to disclose "on a rapid and current basis" material information regarding changes in financial condition or operations as the SEC, by rule, determine to be necessary or useful for the protection of investors and in the public interest. The disclosure is filed with the SEC on Form 8-K, which companies must file "to announce major events that shareholders should know about."

of extreme returns or return skewness (Bali, Cakici, and Whitelaw (2011)). To rule out these alternatives, we add change of breadth of ownership (dBreadth) and max daily return (Maxret) in the Fama-MacBeth regression. Maxret is defined as a stock's maximum daily return in the prior month. Columns (7) to (9) of Table A8 show that the coefficients of AIP becomes even stronger after controlling for change of breadth of ownership and extreme daily returns.

### 6.1.3 Media Coverage

A related concern is that higher investor attention and information acquisition activities correlate with more intensive media coverage of a firm. As a result, the return predictability of EDGAR searching behavior could be driven by news coverage and the information content of news. To directly control for the confounding effect of news coverage and news sentiment, we use data from RavenPack News Analytics, which is a leading global news database used by practitioners in quantitative and algorithmic trading and by scholars in accounting and finance research (Dang, Moshirian, and Zhang (2015)).[36] We count the number of news for each firm over a month and use the natural logarithm of this variable as the news coverage measure. We also include the event sentiment score (ESS) from RavenPack, which indicates how firm-specific news events are categorized and rated as having a positive or negative effect on stock prices by experts with extensive experience and backgrounds in linguistics, finance, and economics.

Table A9 reports the Fama-MacBeth regression results. The sample used in this test is reduced significantly due to the requirement of news coverage data. Columns (1) to (3) show that the coefficients of AIP are still highly significant after controlling for news coverage measure. Columns (4) to (6) report the results when we control for news sentiment. Unsurprisingly, the coefficients on news sentiment itself is significant and positive. Importantly, the return predictability of AIP barely changes. Overall, we conclude that the return predictability of AIP cannot be explained by media coverage.

[36]RavenPack collects and analyzes real-time, firm-level business news from leading news providers (e.g., Dow Jones Newswire, The Wall Street Journal, and Barron's) and other major publishers and web aggregators, including industry and business publications, regional and local newspapers, government and regulatory updates, and trustworthy financial websites.

### 6.1.4 Attention-Driven Price Pressure

We next examine the persistence of the return predictability of AIP. This test could help rule out another alternative explanation, namely that the short-run predictability is due to temporary price pressure driven by investors' excess demand for attention-grabbing stocks. Da, Engelberg, and Gao (2011) show that an increase in Google Search Volume for a stock predicts higher stock prices in the short-run that are eventually reversed within a year. As we hypothesize that AIP contains information about firms' fundamental changes, the return predictability of AIP should not be reversed in the long-run. To test this, we run Fama-MacBeth regression of cumulative returns from month $t + j$ to $t + k$ on AIP_10K in month $t$. The result is reported in Table A10 in the Online Appendix. We separately show the return predictability of AIP_10K for the next-quarter return skipping the immediate month in Column (1), the second-quarter return in Column (2), the second half-year return in Column (3), and the second-year return in Column (4). The table shows that AIP significantly predicts returns for up to two quarters, and eventually levels off for longer horizons. The coefficient of AIP is always positive, mitigating concerns that the predictive power of AIP comes from transitory price pressure that is subsequently reversed. The result suggests that investors searching firm fundamentals through the EDGAR system appear to be more sophisticated than those searching through Google Search Engine, and their aggregate information acquisition activities contain value-relevant information about firms.

### 6.1.5 Investor Recognition

The positive return predictability of AIP could potentially be explained by Merton (1987)'s investor recognition hypothesis. In his model, equilibrium stock return is affected by investors' recognition of a stock because investors are not aware of all securities. Stocks with lower investor recognition have higher expected returns to compensate investors who hold the stock for insufficient diversification. An increase in investor recognition (proxied by abnormal number of IPs) of a stock will reduce its expected return going forward and lead to a contemparenous increases in stock price. This could explain why AIP predicts short-run increase in stock returns. However, several pieces of evidence are not consistent with this alternative explanation. First, a stock experiencing an increase in investor recognition should have **lower** expected returns going forward, which is inconsistent with the fact that AIP also positively predicts long-horizon returns, as presented in Table A10. Second, the investor recognition hypothesis implies that the return predictability of AIP comes solely from the reduction in discount rate, which should have no predictability for firms' future cash flows. However, we show that part of the return predictability of AIP comes from its predictability for a firm's fundamental performance. Lastly, in untabulated analysis, we control for change of trading volume as a proxy for shocks to investor recognition in Fama-MacBeth regression (Gervais, Kaniel, and Mingelgrin (2001)), and the return predictability of AIP barely changes.

#### 6.1.6 Omitted Risk Factors

Last but not least, there is always the possibility that AIP captures some omitted risk factors, despite our best efforts to control for it using various asset-pricing models. First, to the extent that omitted risk factors are persistent at firm level, a within-firm change of AIP should be less able to predict returns if the return predictability of AIP is purely driven by exposure to risk factors. However, Table A4 shows a similarly strong return predictability using the within-firm change of AIP. Second, the fact that the return predictability of AIP concentrates on earnings announcement days is more difficult to square with risk-based explanations (La Porta, Lakonishok, Shleifer, and Vishny (1997)). In addition, we show explicitly that the return predictability of AIP partially comes from its predictability for firms' future fundamental changes. Overall, the omitted risk factor explanation is difficult to square with these additional evidences.

### 6.2 Additional Analyses

#### 6.2.1 Information Acquisition and Investor Trading

Given the large number of unique IPs (more than 3 millions) in the EDGAR log file database and the nature of the EDGAR system, we conjecture that a majority of EDGAR users should be individual investors.[37] Thus, the significant return predictability from information acquisition of EDGAR users is consistent with the recent literature that individual investors as a group exhibit stock picking ability and their aggregated trades predict future stock returns and earnings news. To substantiate this argument, we further examine whether information acquisition through EDGAR leads to subsequent investor trading. We examine trading by two types of investors: mutual funds and retail investors.

To test, we run Fama-MacBeth regression of net purchase by mutual funds and retail order imbalance on lagged AIP, controlling for a set of firm characteristics. Specifically, in each quarter or month, we run the following cross-sectional regression:

$$NetBuy_{i,t} = \beta_0 + \beta_1 AIP_{i,t-1} + \gamma X_{i,t-1} + \epsilon_{i,t} \quad (4)$$

where $NetBuy_{i,t}$ is the net purchases by mutual funds in quarter $t$ or retail order imbalance in month $t$, $AIP_{i,t-1}$ is the abnormal number of IPs searching for firm $i$'s SEC filings in time $t-1$, and $X_{i,t-1}$ is a vector of firm characteristics observed at time $t-1$, including firm size, book-to-market, analyst coverage, volatility, turnover, institutional ownership, and momentum. Net purchase is measured as the quarterly change in mutual fund holdings on a stock, with holdings expressed as a fraction of a firm's shares outstanding. Since mutual fund trade is inferred from quarterly holdings data, we aggregate the AIP at quarterly frequency by averaging the monthly AIP within a quarter. Retail order imbalance is calculated as the difference between daily retail buy volume and retail sell volume, scaled by total daily retail trading volume, and then aggregated to monthly level. Retail buys and sells are classified as in Boehmer, Jones, and Zhang (2017), who show that retail investors are informed about future stock returns in the cross section.[38]

Table 12 reports the time series averages of the cross-sectional regression coefficients. The

[37] Institutional investors, given their resources and capacity, more likely use Bloomberg terminal or other data providers for information acquisition (Ben-Rephael, Da, and Israelsen (2017)).

[38] The Boehmer, Jones, and Zhang (2017) approach exploits two key institutional features of retail trading. First, most equity trades by retail investors take place off-exchange, either filled from the broker's own inventory or sold by the broker to wholesalers. TAQ classifies these types of trades with exchange code "D." Second, retail traders typical receive a small fraction of a cent price improvement over the National Best Bid or Offer (NBBO) for market orders (ranging from 0.01 to 0.2 cents), while institutional orders tend to be executed at whole or half-cent increments. The BJZ approach "picks up a majority of overall retail trading activity". We thank Xiaoyan Zhang for providing us the data on retail order imbalance.

dependent variable is net purchases by mutual funds in Columns (1) to (3). The insignificant coefficients on AIP indicate that EDGAR-based information acquisition activities are not related to subsequent mutual fund trading. In sharp contrast, when the dependent variable is retail order imbalance in Columns (4) to (6), the coefficients on AIP are highly significant and positive. The result suggests that more information acquisition activities through the EDGAR system leads to significant net buying from retail investors subsequently.

### 6.2.2 IPs or Searches?

Our measure of information acquisition activity essentially equal weights each IP regardless of the number of searches the IP conducted through the EDGAR system during a one-month window. An alternative measure of information acquisition activity is the total number of searches for a firm requested by investors through the EDGAR system. This measure is problematic because, as documented by Drake, Roulstone, and Thornock (2015), the number of requests through EDGAR is dominated by a small fraction of investors who access EDGAR very frequently, and their activities are over-represented in this alternative measure.[39] Under the assumption that information is dispersed among a large group of economic agents (Hayek (1945)), we believe that our measure of the abnormal number of IPs should be more powerful in terms of inferring the latent information embedded in "the wisdom of crowd". Nevertheless, to test which measure of information acquisition activity has the stronger return predictability, we conduct a horse race between the abnormal number of searches (Asearch) and abnormal number of IPs (AIP) using the Fama-MacBeth regression approach. Using the same decomposition method, we extract the abnormal number of searches for each firm as the residual from a monthly cross-sectional regression of log one plus the raw number of EDGAR requests for SEC filings on the same set of firm characteristics used in equation (2).

The result is reported in Table A12 in the Online Appendix. Searches/IPs for all types of SEC filings are shown in Columns (1) and (2), searches/IPs for 10-Ks, 10-Qs and 8-Ks in Columns (3) and (4), and searches/IPs only for 10-Ks in Columns (5) and (6). Columns (1),

[39]Drake, Roulstone, and Thornock (2015) report that 86% of the users accessing EDGAR do so infrequently and only about 2% of the users access EDGAR actively during a given quarter.

(3), and (5) show that the return predictability of Asearch is generally positive but weaker than that of AIP. Columns (2), (4), and (6) show that once we control for AIP, the coefficient of Asearch is no longer significant. Importantly, the coefficients of AIP are still positive and highly significant. The result supports our use of the number of IPs as a cleaner measure of information acquisition activity, and indirectly supports the underlying assumption that private information is dispersed among market participants.

# 7 Conclusion

In this paper, we examine the expected return information embedded in investors' costly information acquisition activities. Specifically, we use a novel dataset of investors' requests for company filings through the SEC's EDGAR system to infer their expectations of future payoffs. We develop and implement a simple characteristic-based model to decompose the total number of IPs searching for EDGAR filings into abnormal and expected components, and show that the abnormal number of IPs searching for firms' SEC filings positively predicts subsequent stock returns. A long-short portfolio that buys stocks with an abnormal number of IPs in the top decile and sells stocks in the bottom decile generates an equal-weighted monthly four-factor alpha of up to 80 basis points that is not reversed in the long run. We also find that the abnormal number of IPs predicts firms' ascending fundamental performances, and that it also increases following exogenous underpricing, suggesting that investors rationally allocate greater attention and effort to undervalued firms with large price appreciation potential.

Taken together, our findings provide empirical support to theoretical models of endogenous information acquisition that costly information acquisition activity is positively associated with the value of information (Grossman and Stiglitz (1980)). Our research also highlights the promise of using the collective wisdom of investors – extracted from their EDGAR search behavior – to study expected returns and other important economic outcomes.

# References


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

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang, 2006, "The cross-section of volatility and expected returns," *The Journal of Finance*, 61(1), 259–299.

Asparouhova, E., H. Bessembinder, and I. Kalcheva, 2013, "Noisy prices and inference regarding returns," *The Journal of Finance*, 68(2), 665–714.

Bali, T. G., N. Cakici, and R. F. Whitelaw, 2011, "Maxing out: Stocks as lotteries and the cross-section of expected returns," *Journal of Financial Economics*, 99(2), 427–446.

Barber, B. M., and T. Odean, 2007, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *The Review of Financial Studies*, 21(2), 785–818.

Ben-Rephael, A., Z. Da, and R. D. Israelsen, 2017, "It Depends on Where You Search: Institutional Investor Attention and Underreaction to News," *The Review of Financial Studies*, p. hhx031.

Bernard, V. L., and J. K. Thomas, 1989, "Post-earnings-announcement drift: delayed price response or risk premium?," *Journal of Accounting research*, pp. 1–36.

Boehmer, E., C. Jones, and X. Zhang, 2017, "Tracking retail investor activity," .

Carhart, M. M., 1997, "On persistence in mutual fund performance," *The Journal of finance*, 52(1), 57–82.

Chen, H., L. Cohen, U. Gurun, D. Lou, and C. Malloy, 2018, "IQ from IP: Simplifying search in portfolio choice," working paper, National Bureau of Economic Research.

Chen, H., P. De, Y. Hu, and B.-H. Hwang, 2014, "Wisdom of crowds: The value of stock opinions transmitted through social media," *The Review of Financial Studies*, 27(5), 1367–1403.

Chen, J., H. Hong, and J. C. Stein, 2002, "Breadth of ownership and stock returns," *Journal of financial Economics*, 66(2), 171–205.

Cohen, L., C. Malloy, and Q. Nguyen, 2018, "Lazy prices," working paper, National Bureau of Economic Research.

Coval, J., and E. Stafford, 2007, "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 86(2), 479–512.

Cowen, A., B. Groysberg, and P. Healy, 2006, "Which types of analyst firms are more optimistic?," *Journal of Accounting and Economics*, 41(1-2), 119–146.

Crane, A. D., K. Crotty, and T. Umar, 2018, "Do hedge funds profit from public information?," *Available at SSRN 3127825*.

Da, Z., J. Engelberg, and P. Gao, 2011, "In search of attention," *The Journal of Finance*, 66(5), 1461–1499.

Dang, T. L., F. Moshirian, and B. Zhang, 2015, "Commonality in news around the world," *Journal of Financial Economics*, 116(1), 82–110.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, "Measuring mutual fund performance with characteristic-based benchmarks," *The Journal of Finance*, 52(3), 1035–1058.

Daniel, K., and T. J. Moskowitz, 2016, "Momentum crashes," *Journal of Financial Economics*, 122(2), 221–247.

Diamond, D. W., and R. E. Verrecchia, 1981, "Information aggregation in a noisy rational expectations economy," *Journal of Financial Economics*, 9(3), 221–235.

Drake, M. S., D. T. Roulstone, and J. R. Thornock, 2015, "The determinants and consequences of information acquisition via EDGAR," *Contemporary Accounting Research*, 32(3), 1128–1161.

——— , 2016, "The usefulness of historical accounting reports," *Journal of Accounting and Economics*, 61(2), 448–464.

Du, Z., 2015, "Endogenous Information Acquisition: Evidence from Web Visits to SEC Filings of Insider Trdes," working paper, Working paper, Kellogg School of Management.

Edmans, A., I. Goldstein, and W. Jiang, 2012, "The real effects of financial markets: The impact of prices on takeovers," *The Journal of Finance*, 67(3), 933–971.

Engelberg, J., R. D. McLean, and J. Pontiff, 2018, "Anomalies and news," *The Journal of Finance*, 73(5), 1971–2001.

Fama, E. F., and K. R. French, 1993, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33(1), 3–56.

——— , 2015, "A five-factor asset pricing model," *Journal of financial economics*, 116(1), 1–22.

Fama, E. F., and J. D. MacBeth, 1973, "Risk, return, and equilibrium: Empirical tests," *The Journal of Political Economy*, pp. 607–636.

Gao, M., and J. Huang, 2019, "Informing the market: The effect of modern information technologies on information production," *Review of Financial Studies, Forthcoming.*

Gao, X., and J. R. Ritter, 2010, "The marketing of seasoned equity offerings," *Journal of Financial Economics*, 97(1), 33–52.

Gervais, S., R. Kaniel, and D. H. Mingelgrin, 2001, "The high-volume return premium," *The Journal of Finance*, 56(3), 877–919.

Gibbons, B., P. Iliev, and J. Kalodimos, 2019, "Analyst information acquisition via EDGAR," *Available at SSRN 3112761.*

Green, T. C., R. Huang, Q. Wen, and D. Zhou, 2019, "Crowdsourced employer reviews and stock returns," *Journal of Financial Economics.*

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

Hartzmark, S. M., and D. H. Solomon, 2013, "The dividend month premium," *Journal of Financial Economics*, 109(3), 640–660.

Hayek, F. A., 1945, "The use of knowledge in society," *The American economic review*, pp. 519–530.

Hellwig, M. F., 1980, "On the aggregation of information in competitive markets," *Journal of economic theory*, 22(3), 477–498.

Hong, H., T. Lim, and J. C. Stein, 2000, "Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies," *The Journal of Finance*, 55(1), 265–295.

Hou, K., C. Xue, and L. Zhang, 2015, "Digesting Anomalies: An Investment Approach," *Review of Financial Studies*, 28(3), 650–705.

Huang, J., 2016, "The customer knows best: The investment value of consumer opinions," *Browser Download This Paper*.

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

Kaniel, R., S. Liu, G. Saar, and S. Titman, 2012, "Individual investor trading and return patterns around earnings announcements," *The Journal of Finance*, 67(2), 639–680.

Kaniel, R., G. Saar, and S. Titman, 2008, "Individual investor trading and stock returns," *The Journal of Finance*, 63(1), 273–310.

Kelley, E. K., and P. C. Tetlock, 2013, "How wise are crowds? Insights from retail orders and stock returns," *The Journal of Finance*, 68(3), 1229–1265.

Khan, M., L. Kogan, and G. Serafeim, 2012, "Mutual fund trading pressure: Firm-level stock price impact and timing of SEOs," *The Journal of Finance*, 67(4), 1371–1395.

La Porta, R., J. Lakonishok, A. Shleifer, and R. Vishny, 1997, "Good News for Value Stocks: Further Evidence on Market Efficiency," *Journal of Finance*, pp. 859–874.

Lamont, O., and A. Frazzini, 2007, "The earnings announcement premium and trading volume," working paper, National Bureau of Economic Research.

Lee, C. M., P. Ma, and C. C. Wang, 2015, "Search-based peer firms: Aggregating investor perceptions through internet co-searches," *Journal of Financial Economics*, 116(2), 410–431.

Lee, C. M., and E. C. So, 2017, "Uncovering expected returns: Information in analyst coverage proxies," *Journal of Financial Economics*, 124(2), 331–348.

Li, E. X., G. Lind, K. Ramesh, and M. Shen, 2018, "Externalities of accounting disclosures: evidence from the Federal Reserve," *Available at SSRN 3179251*.

Liaukonyte, J., and A. Zaldokas, 2019, "Background Noise? TV Advertising Affects Real Time Investor Behavior," *TV Advertising Affects Real Time Investor Behavior (February 17, 2019)*.

Lin, H.-w., and M. F. McNichols, 1998, "Underwriting relationships, analysts' earnings forecasts and investment recommendations," *Journal of Accounting and Economics*, 25(1), 101–127.

Loughran, T., and B. McDonald, 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 66(1), 35–65.

——— , 2014, "Measuring readability in financial disclosures," *The Journal of Finance*, 69(4), 1643–1671.

——— , 2017, "The use of EDGAR filings by investors," *Journal of Behavioral Finance*, 18(2), 231–248.

Madsen, J., 2017, "Anticipated earnings announcements and the customer–supplier anomaly," *Journal of Accounting Research*, 55(3), 709–741.

Mele, A., and F. Sangiorgi, 2015, "Uncertainty, information acquisition, and price swings in asset markets," *The Review of Economic Studies*, 82(4), 1533–1567.

Merton, R. C., 1987, "A simple model of capital market equilibrium with incomplete information," *The journal of finance*, 42(3), 483–510.

Monga, V., and E. Chasan, 2015, "The 109,894-Word Annual Report: As Regulators Require More Disclosures, 10-Ks Reach Epic Lengths; How Much Is Too Much?," *Wall Street Journal*.

Nagel, S., 2005, "Short sales, institutional investors and the cross-section of stock returns," *Journal of Financial Economics*, 78(2), 277–309.

Pástor, L., and R. F. Stambaugh, 2003, "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy*, 111(3), 642–685.

Pedersen, L. H., 2015, *Efficiently inefficient: how smart money invests and market prices are determined*. Princeton University Press.

Petersen, M. A., 2009, "Estimating standard errors in finance panel data sets: Comparing approaches," *Review of Financial Studies*, 22(1), 435–480.

Pontiff, J., 2006, "Costly arbitrage and the myth of idiosyncratic risk," *Journal of Accounting and Economics*, 42(1), 35–52.

Ryans, J. P., 2017, "Using the EDGAR Log File Data Set," .

Saffi, P. A., and K. Sigurdsson, 2010, "Price efficiency and short selling," *The Review of*

*Financial Studies*, 24(3), 821–852.

Sheng, J., 2019, "Asset pricing in the information age: Employee expectations and stock returns," *Available at SSRN 3321275*.

Shumway, T., 1997, "The delisting bias in CRSP data," *The Journal of Finance*, 52(1), 327–340.

Stambaugh, R. F., J. Yu, and Y. Yuan, 2015, "Arbitrage asymmetry and the idiosyncratic volatility puzzle," *The Journal of Finance*.

Stambaugh, R. F., and Y. Yuan, 2016, "Mispricing factors," *The Review of Financial Studies*, 30(4), 1270–1315.

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

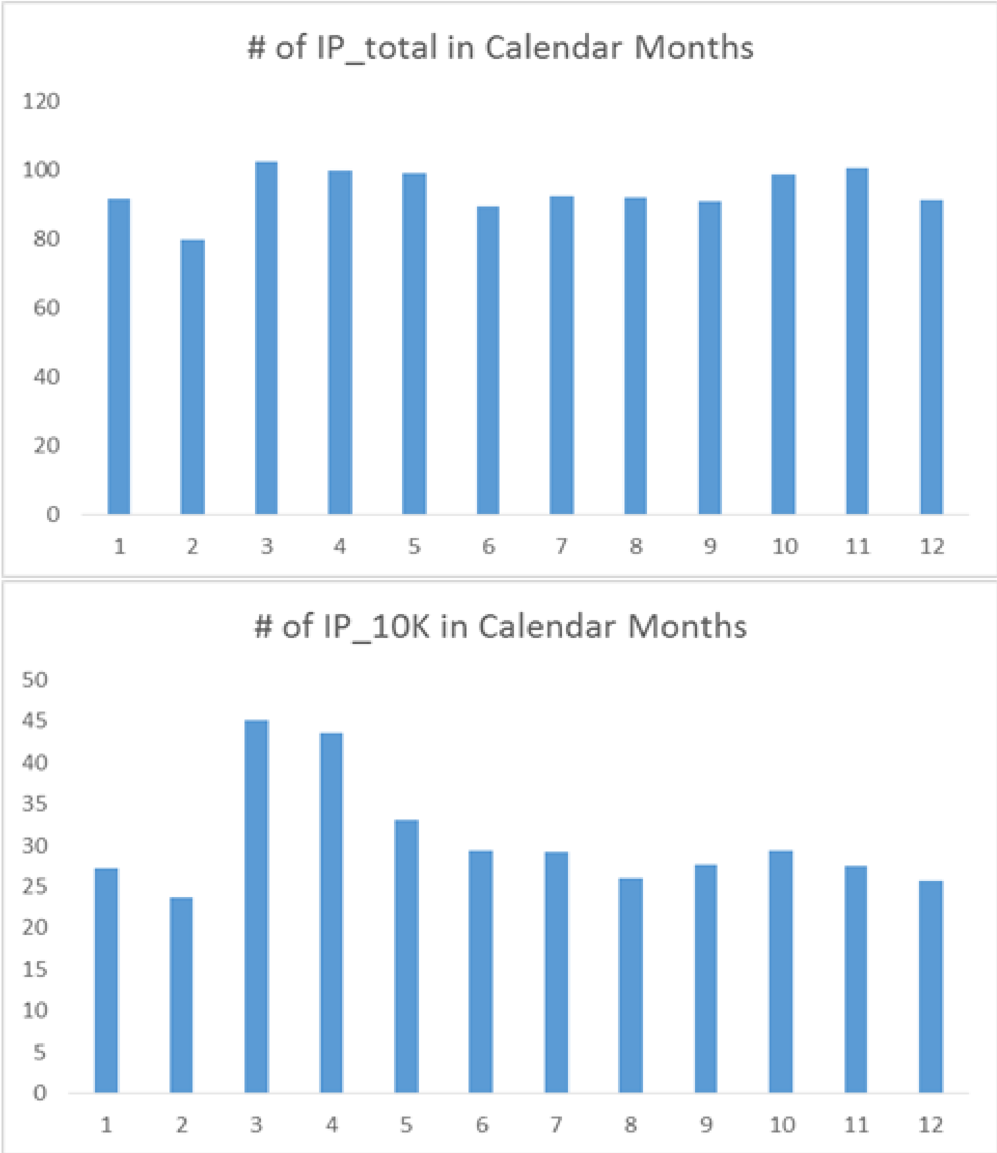
Verrecchia, R. E., 1982, "Information acquisition in a noisy rational expectations economy," *Econometrica: Journal of the Econometric Society*, pp. 1415–1430.

Womack, K. L., 1996, "Do brokerage analysts' recommendations have investment value?," *The journal of finance*, 51(1), 137–167.

You, H., and X.-j. Zhang, 2009, "Financial reporting complexity and investor underreaction to 10-K information," *Review of Accounting Studies*, 14(4), 559–586.

Zhang, X., 2006, "Information uncertainty and stock returns," *The Journal of Finance*, 61(1), 105–137.

Figure 1: **Averge Number of IPs in Calendar Months**




This figure plots the average number of IPs searching for EDGAR filings in each calendar month. The average is first calculated across stocks within a particular year-month and then averaged across years. IP_total is the total number of unique IP addresses searching for all six types of EDGAR filings. IP_10K is the total number of unique IP addresses searching for 10-K files. The sample period is from January 2003 to December 2014.

Table 1: **Stock-Level Descriptive Statistics**

This table presents the descriptive statistics of our variables. Panel A reports the summary statistics for the full sample. Panel B reports the pairwise rank correlation between our variables where they overlap. Panel C reports the characteristics of portfolios sorted by the abnormal number of IPs searching for 10-K filings in the SEC's EDGAR system (AIP_10K). IP_total is the total number of unique IP addresses searching for all six types of SEC filings. IP_funtl is the total number of unique IP addresses searching for 10-K, 10-Q, and 8-K filings. AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system. For each month, we sort all stocks into deciles based on their AIP_10K. We first calculate the mean of each variable for each decile in each month, and then calculate the time-series average of cross-sectional means. LnME is the natural log of a firm's market capitalization at the end of June of each year in millions of US dollars. Coverage is log one plus analyst coverage. Turnover12 is the monthly turnover ratio averaged over the past 12 months. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. Lendable supply is the shares held and made available to lend by Markit lenders divided by total shares outstanding. DCBS is a score from 1 to 10 created by Markit using their proprietary information and is intended to capture the cost of borrowing the stock. Outflows is calculated following Edmans, Goldstein, and Jiang (2012), which reflects fund outflow expressed as a percentage of stock's total dollar trading volume within a quarter. The overall sample period is from January 2003 to December 2014.

| Panel A: Summary Statistics | | | | | |
|---|---|---|---|---|---|
| Variable | Mean | Median | STD | P25 | P75 |
| *Number of IP searching for EDGAR filings* | | | | | |
| IP_total | 155 | 94 | 317 | 56 | 159 |
| IP_funtl | 107 | 64 | 213 | 37 | 111 |
| IP_10K | 60 | 32 | 135 | 17 | 60 |
| IP_10Q | 37 | 24 | 61 | 13 | 42 |
| IP_8K | 33 | 19 | 79 | 10 | 36 |
| *Stock-level characteristics* | | | | | |
| LnME | 6.16 | 6.08 | 1.98 | 4.74 | 7.47 |
| LnBM | -0.66 | -0.56 | 0.84 | -1.11 | -0.12 |
| Mom | 16.67% | 7.64% | 57.57% | -12.06% | 31.78% |
| Coverage | 1.49 | 1.59 | 1.01 | 0.59 | 2.30 |
| IVOL | 0.02 | 0.02 | 0.02 | 0.01 | 0.03 |
| Turnover12 | 0.17 | 0.12 | 0.19 | 0.05 | 0.21 |
| IO | 55.30% | 59.15% | 31.41% | 28.92% | 80.58% |
| dROA (%) | 0.032 | -0.018 | 4.844 | -0.684 | 0.599 |
| FREV (%) | -0.106 | -0.001 | 22.185 | -0.070 | 0.052 |
| Outflows | -0.10% | -0.05% | 0.19% | -0.11% | -0.02% |
| Lendable Supply | 13.96% | 14.46% | 8.98% | 5.85% | 20.89% |
| DCBS | 1.48 | 1.00 | 1.22 | 1.00 | 1.17 |

## Table 1 Continued

| Panel B: Rank Correlations | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | IP_total | IP_funtl | IP_10K | LnME | Cov | Turnover12 | Ivol | LnBM | Mom | IO |
| IP_total | 1.000 | | | | | | | | | |
| IP_funtl | 0.918 | 1.000 | | | | | | | | |
| IP_10K | 0.812 | 0.897 | 1.000 | | | | | | | |
| LnME | 0.671 | 0.664 | 0.672 | 1.000 | | | | | | |
| Cov | 0.594 | 0.605 | 0.603 | 0.832 | 1.000 | | | | | |
| Turnover12 | 0.588 | 0.579 | 0.539 | 0.544 | 0.621 | 1.000 | | | | |
| Ivol | -0.134 | -0.149 | -0.212 | -0.523 | -0.360 | -0.016 | 1.000 | | | |
| LnBM | -0.239 | -0.229 | -0.224 | -0.319 | -0.326 | -0.303 | 0.051 | 1.000 | | |
| Mom | 0.031 | 0.023 | 0.044 | 0.112 | 0.051 | 0.049 | -0.117 | 0.008 | 1.000 | |
| IO | 0.469 | 0.494 | 0.514 | 0.650 | 0.647 | 0.615 | -0.306 | -0.193 | 0.095 | 1.000 |

## Table 1 Continued

| Panel C: Descriptive statistics by AIP_10K deciles | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Obs | AIP_10K | IP_total | IP_funtl | IP_10K | LnME | Cov | Turnover12 | Ivol | LnBM | Mom | IO |
| 1(Low) | 330 | -1.25 | 59 | 35 | 12 | 5.977 | 1.369 | 0.154 | 0.025 | -0.590 | 0.150 | 45.53% |
| 2 | 330 | -0.60 | 76 | 51 | 22 | 6.074 | 1.513 | 0.163 | 0.024 | -0.719 | 0.164 | 53.38% |
| 3 | 330 | -0.38 | 91 | 63 | 30 | 6.166 | 1.573 | 0.166 | 0.024 | -0.742 | 0.163 | 57.21% |
| 4 | 330 | -0.21 | 104 | 72 | 36 | 6.248 | 1.611 | 0.170 | 0.024 | -0.741 | 0.172 | 59.23% |
| 5 | 330 | -0.07 | 116 | 82 | 42 | 6.270 | 1.623 | 0.171 | 0.024 | -0.711 | 0.176 | 60.20% |
| 6 | 330 | 0.07 | 128 | 91 | 48 | 6.284 | 1.634 | 0.170 | 0.024 | -0.700 | 0.173 | 60.79% |
| 7 | 330 | 0.22 | 141 | 101 | 55 | 6.218 | 1.594 | 0.165 | 0.024 | -0.662 | 0.174 | 60.19% |
| 8 | 330 | 0.39 | 160 | 116 | 66 | 6.118 | 1.526 | 0.164 | 0.024 | -0.623 | 0.164 | 58.91% |
| 9 | 330 | 0.62 | 201 | 147 | 87 | 6.032 | 1.454 | 0.158 | 0.025 | -0.563 | 0.162 | 56.09% |
| 10(High) | 330 | 1.14 | 464 | 342 | 226 | 6.257 | 1.483 | 0.163 | 0.025 | -0.537 | 0.168 | 53.28% |

Table 2: **Cross-Sectional Determinants of Number of IPs Searching EDGAR Filings**

This table presents the Fama-MacBeth regression of log number of IPs searching for SEC filings through EDGAR system. In Panel A, the dependent variable is log one plus the number of unique IP addresses searching for SEC filings in a month. In Panel B, the dependent variable is log one plus the number of unique IP addresses searching for 10-K, 10-Q and 8-K filings in a month. In Panel C, the dependent variable is log one plus the number of unique IP addresses searching for 10-K filings in a month. LnME is the natural log of a firm's market capitalization at the end of June of each year in millions of US dollars. Coverage is log one plus analyst coverage. Turnover12 is the average monthly turnover ratio over the past 12 months. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. SP500 is a dummy equal to one if the stock belongs to S&P500 index. EAM is a dummy variable that equals one when a given firm announces quarterly earnings in the month. The overall sample period is from January 2003 to December 2014.

Panel A: Dependent Variable is log(1+# of unique IP adresses searching all EDGAR filings)

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| LnME | 0.2713*** | 0.2356*** | 0.2475*** | 0.2943*** | 0.2992*** | 0.3015*** | 0.3026*** | 0.2608*** | 0.2628*** |
| | (69.44) | (71.54) | (73.46) | (75.60) | (76.94) | (77.29) | (77.58) | (75.05) | (74.98) |
| Coverage | | 0.1310*** | 0.0422*** | 0.0382*** | 0.0321*** | 0.0332*** | 0.0360*** | 0.0337*** | 0.0399*** |
| | | (32.65) | (14.39) | (14.36) | (12.17) | (12.56) | (14.17) | (13.99) | (16.86) |
| Turnover12 | | | 1.0083*** | 0.7934*** | 0.7862*** | 0.7912*** | 0.7877*** | 0.8175*** | 0.8113*** |
| | | | (30.21) | (29.08) | (30.04) | (29.75) | (30.52) | (30.68) | (30.92) |
| IVOL | | | | 9.1266*** | 9.0159*** | 9.0510*** | 9.0215*** | 8.5748*** | 8.0871*** |
| | | | | (34.65) | (33.38) | (33.16) | (32.36) | (31.55) | (31.65) |
| MOM | | | | | -0.0518*** | -0.0529*** | -0.0507*** | -0.0508*** | -0.0513*** |
| | | | | | (-6.00) | (-6.19) | (-5.99) | (-6.38) | (-6.43) |
| LnBM | | | | | | 0.0171*** | 0.0158*** | 0.0087*** | 0.0108*** |
| | | | | | | (8.19) | (7.25) | (4.06) | (5.16) |
| IO | | | | | | | -0.0299** | 0.0657*** | 0.0575*** |
| | | | | | | | (-1.99) | (4.80) | (4.37) |
| SP500 | | | | | | | | 0.3634*** | 0.3591*** |
| | | | | | | | | (58.81) | (58.60) |
| EAM | | | | | | | | | 0.1587*** |
| | | | | | | | | | (9.62) |
| Constant | 2.5352*** | 2.6342*** | 2.5357*** | 2.0730*** | 2.0483*** | 2.0408*** | 2.0449*** | 2.2164*** | 2.1892*** |
| | (39.20) | (40.68) | (40.19) | (33.45) | (33.37) | (33.32) | (32.62) | (34.26) | (34.03) |
| Ave.R-sq | 0.404 | 0.483 | 0.520 | 0.554 | 0.558 | 0.559 | 0.563 | 0.574 | 0.582 |
| N.of Obs. | 610651 | 488129 | 488129 | 488123 | 488123 | 488123 | 484835 | 484835 | 484835 |

## Table 2 Continued

Panel B: Dependent Variable is log(1+# of unique IP adresses searching 10-K, 10-Q and 8-K filings)

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| LnME | 0.2723*** | 0.2355*** | 0.2468*** | 0.2931*** | 0.2984*** | 0.3015*** | 0.3005*** | 0.2563*** | 0.2578*** |
| | (64.05) | (61.21) | (60.97) | (65.17) | (66.87) | (67.27) | (67.10) | (63.63) | (63.29) |
| Coverage | | 0.1405*** | 0.0530*** | 0.0492*** | 0.0421*** | 0.0436*** | 0.0369*** | 0.0343*** | 0.0414*** |
| | | (35.59) | (15.60) | (15.87) | (13.86) | (14.48) | (15.57) | (15.24) | (18.17) |
| Turnover12 | | | 0.9833*** | 0.7702*** | 0.7708*** | 0.7787*** | 0.7560*** | 0.7878*** | 0.7856*** |
| | | | (29.18) | (26.62) | (27.23) | (26.95) | (27.32) | (27.66) | (28.78) |
| IVOL | | | | 9.0866*** | 8.9652*** | 9.0334*** | 9.0934*** | 8.6203*** | 7.9829*** |
| | | | | (36.40) | (34.66) | (34.19) | (33.54) | (32.67) | (32.41) |
| MOM | | | | | -0.0684*** | -0.0698*** | -0.0685*** | -0.0687*** | -0.0696*** |
| | | | | | (-7.72) | (-8.00) | (-7.95) | (-8.50) | (-8.56) |
| LnBM | | | | | | 0.0251*** | 0.0223*** | 0.0148*** | 0.0172*** |
| | | | | | | (10.23) | (9.01) | (6.09) | (7.28) |
| IO | | | | | | | 0.0411*** | 0.1421*** | 0.1303*** |
| | | | | | | | (2.76) | (10.31) | (9.80) |
| SP500 | | | | | | | | 0.3863*** | 0.3814*** |
| | | | | | | | | (62.83) | (62.17) |
| EAM | | | | | | | | | 0.2092*** |
| | | | | | | | | | (10.96) |
| Constant | 2.2017*** | 2.2804*** | 2.1866*** | 1.7281*** | 1.7033*** | 1.6943*** | 1.6868*** | 1.8686*** | 1.8366*** |
| | (34.86) | (36.21) | (35.81) | (28.85) | (28.72) | (28.66) | (27.97) | (30.13) | (29.99) |
| Ave.R-sq | 0.386 | 0.458 | 0.491 | 0.522 | 0.526 | 0.527 | 0.533 | 0.543 | 0.554 |
| N.of Obs. | 610651 | 488129 | 488129 | 488123 | 488123 | 488123 | 484835 | 484835 | 484835 |

## Table 2 Continued

Panel C: Dependent Variable is log(1+# of unique IP adresses searching 10-K filings)

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| LnME | 0.2979*** | 0.2674*** | 0.2765*** | 0.3120*** | 0.3169*** | 0.3201*** | 0.3155*** | 0.2648*** | 0.2678*** |
| | (61.33) | (60.72) | (59.37) | (61.48) | (62.64) | (63.24) | (62.55) | (58.19) | (58.65) |
| Coverage | | 0.1453*** | 0.0729*** | 0.0698*** | 0.0637*** | 0.0649*** | 0.0431*** | 0.0401*** | 0.0482*** |
| | | (35.85) | (23.41) | (23.28) | (21.42) | (21.49) | (16.48) | (15.66) | (18.95) |
| Turnover12 | | | 0.8122*** | 0.6461*** | 0.6415*** | 0.6522*** | 0.5924*** | 0.6288*** | 0.6188*** |
| | | | (30.68) | (28.59) | (28.59) | (28.74) | (28.38) | (28.75) | (29.59) |
| IVOL | | | | 6.9981*** | 6.9145*** | 7.0130*** | 7.2542*** | 6.7143*** | 6.2019*** |
| | | | | (30.56) | (29.46) | (28.94) | (29.41) | (28.15) | (27.44) |
| MOM | | | | | -0.0484*** | -0.0510*** | -0.0521*** | -0.0517*** | -0.0517*** |
| | | | | | (-5.54) | (-5.93) | (-6.09) | (-6.52) | (-6.60) |
| LnBM | | | | | | 0.0267*** | 0.0213*** | 0.0127*** | 0.0159*** |
| | | | | | | (9.03) | (7.48) | (4.39) | (5.84) |
| IO | | | | | | | 0.1600*** | 0.2765*** | 0.2654*** |
| | | | | | | | (10.36) | (18.94) | (18.82) |
| SP500 | | | | | | | | 0.4416*** | 0.4358*** |
| | | | | | | | | (53.84) | (53.85) |
| EAM | | | | | | | | | 0.1730*** |
| | | | | | | | | | (7.36) |
| Constant | 1.3873*** | 1.4159*** | 1.3396*** | 0.9886*** | 0.9639*** | 0.9554*** | 0.9267*** | 1.1349*** | 1.1097*** |
| | (25.17) | (25.47) | (24.67) | (18.65) | (18.51) | (18.43) | (17.62) | (20.88) | (20.57) |
| Ave.R-sq | 0.388 | 0.467 | 0.486 | 0.501 | 0.504 | 0.506 | 0.511 | 0.522 | 0.532 |
| N.of Obs. | 610651 | 488129 | 488129 | 488123 | 488123 | 488123 | 484835 | 484835 | 484835 |

Table 3: **Portfolio Excess Returns Sorted by Abnormal Number of IPs**

This table reports the monthly average excess returns (in percentage) for each of the decile portfolios, as well as the long-short portfolio (High-Low). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all types of SEC filings in the EDGAR system on a set of firm characteristics (equation (1)). Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. In the end of each month, all stocks are sorted into deciles based on their abnormal numbers of IPs, and a long-short portfolio is formed by buying the highest decile and shorting the lowest decile portfolio. Portfolio returns are computed over the next month. Panel A reports the results for equally weighted portfolios and Panel B shows the results for value-weighted portfolios. The sample runs from January 2003 to December 2014.

Panel A: Equal-weighted Decile Portfolio Excess Return

| | AIP_10K | t-stat | AIP_funtl | t-stat | AIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | 0.47 | 1.22 | 0.50 | 1.29 | 0.46 | 1.20 |
| 2 | 0.63 | 1.40 | 0.76 | 1.73 | 0.78 | 1.78 |
| 3 | 0.75 | 1.68 | 0.80 | 1.79 | 0.81 | 1.83 |
| 4 | 0.85 | 1.81 | 1.04 | 2.27 | 1.08 | 2.33 |
| 5 | 0.93 | 1.99 | 1.00 | 2.13 | 1.00 | 2.15 |
| 6 | 1.02 | 2.11 | 0.99 | 2.07 | 1.07 | 2.24 |
| 7 | 1.11 | 2.28 | 1.14 | 2.34 | 1.19 | 2.40 |
| 8 | 1.26 | 2.51 | 1.06 | 2.05 | 1.12 | 2.19 |
| 9 | 1.32 | 2.54 | 1.24 | 2.35 | 1.14 | 2.21 |
| High | 1.48 | 2.98 | 1.29 | 2.55 | 1.18 | 2.29 |
| High - Low | **1.00** | 4.70 | **0.79** | 3.61 | **0.71** | 3.18 |

Panel B: Value-weighted Decile Portfolio Excess Return

| | AIP_10K | t-stat | AIP_funtl | t-stat | AIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | 0.48 | 1.42 | 0.57 | 1.60 | 0.40 | 1.01 |
| 2 | 0.59 | 1.39 | 0.72 | 1.72 | 0.80 | 2.01 |
| 3 | 0.68 | 1.61 | 0.86 | 2.15 | 0.76 | 1.93 |
| 4 | 0.83 | 2.03 | 0.97 | 2.35 | 1.04 | 2.58 |
| 5 | 0.99 | 2.54 | 0.92 | 2.20 | 0.85 | 2.09 |
| 6 | 0.75 | 1.83 | 0.89 | 2.23 | 0.80 | 2.03 |
| 7 | 0.88 | 2.18 | 0.90 | 2.38 | 1.00 | 2.62 |
| 8 | 1.01 | 2.70 | 0.84 | 2.13 | 0.89 | 2.26 |
| 9 | 0.74 | 2.04 | 0.87 | 2.43 | 0.94 | 2.60 |
| High | 0.75 | 2.28 | 0.66 | 2.01 | 0.71 | 2.13 |
| High - Low | 0.26 | 1.32 | 0.09 | 0.44 | 0.31 | 1.23 |

Table 4: **Portfolios Alphas Sorted by Abnormal Number of IPs**

This table reports the monthly Carhart (1997) four-factor alphas (in percentage) for each of the 10 decile portfolios, as well as the long-short portfolio (High-Low). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. In the end of each month, all stocks are sorted into deciles based on their abnormal numbers of IPs, and a long-short portfolio is formed by buying the highest decile and shorting the lowest decile portfolio. Portfolio returns are computed over the next month. Panel A reports the results for equally weighted portfolios and Panel B shows the results for value-weighted portfolios. The sample runs from January 2003 to December 2014.

Panel A: Equal-weighted Decile Portfolio four-factor alpha

| | AIP_10K | t-stat | AIP_funtl | t-stat | AIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | -0.28 | -2.33 | -0.29 | -2.58 | -0.36 | -3.28 |
| 2 | -0.24 | -2.78 | -0.12 | -1.31 | -0.08 | -0.89 |
| 3 | -0.13 | -1.39 | -0.06 | -0.63 | -0.10 | -1.26 |
| 4 | -0.05 | -0.58 | 0.11 | 1.10 | 0.13 | 1.35 |
| 5 | -0.03 | -0.36 | 0.06 | 0.64 | 0.05 | 0.59 |
| 6 | 0.07 | 0.66 | -0.01 | -0.09 | 0.11 | 1.15 |
| 7 | 0.08 | 0.56 | 0.17 | 1.28 | 0.20 | 1.60 |
| 8 | 0.26 | 1.88 | 0.08 | 0.44 | 0.11 | 0.66 |
| 9 | 0.27 | 1.32 | 0.20 | 1.12 | 0.17 | 1.03 |
| High | 0.52 | 2.92 | 0.34 | 1.91 | 0.23 | 1.13 |
| High - Low | **0.80** | 3.90 | **0.63** | 2.96 | **0.59** | 2.77 |

Panel B: Value-weighted Decile Portfolio four-factor alpha

| | AIP_10K | t-stat | AIP_funtl | t-stat | AIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | -0.23 | -1.40 | -0.18 | -1.05 | -0.42 | -2.18 |
| 2 | -0.28 | -2.12 | -0.14 | -1.08 | -0.05 | -0.37 |
| 3 | -0.17 | -1.37 | 0.10 | 0.77 | -0.04 | -0.33 |
| 4 | 0.01 | 0.14 | 0.10 | 0.79 | 0.20 | 1.76 |
| 5 | 0.14 | 1.23 | 0.02 | 0.15 | -0.04 | -0.39 |
| 6 | -0.14 | -1.37 | 0.01 | 0.10 | -0.10 | -0.90 |
| 7 | 0.00 | -0.03 | 0.06 | 0.59 | 0.17 | 1.68 |
| 8 | 0.23 | 2.64 | -0.03 | -0.24 | 0.06 | 0.45 |
| 9 | -0.10 | -1.05 | 0.08 | 0.94 | 0.13 | 1.32 |
| High | 0.02 | 0.18 | -0.06 | -0.68 | -0.01 | -0.11 |
| High - Low | 0.25 | 1.19 | 0.12 | 0.52 | 0.41 | 1.68 |

Table 5: **Variation in the Limits to Arbitrage and Short-Sales Constraints**

This table reports the return predictability results for variation in the limits to arbitrage. We sort stocks into terciles based on each limits-to-arbitrage variable X, including idiosyncratic volatility (IVOL) (Panel A), institutional ownership (IO) (Panel B), analyst coverage (Coverage) (Panel C) and lendable supply (Panel D). For lending fee measure (Panel E), we sort stocks into two groups based on whether a stock's DCBS score is above or below 2. We then independently sort stocks into quintiles based on the abnormal number of IPs searching for 10-K (AIP_10K). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. We report the Carhart (1997) four-factor alpha of the lowest and highest AIP portfolios in the lowest and highest X groups. The "High-Low" column reports the Carhart (1997) four-factor alpha (in percentage) of the high-AIP minus low-AIP portfolios. In the bottom row of each panel, we report the difference of four-factor alphas between the high and low limits-to-arbitrage groups. T-statistics are in brackets. The sample runs from January 2003 to December 2014.

| | Low AIP_10K | High AIP_10K | High-Low |
|---|---|---|---|
| Panel A: Double sort on IVOL and AIP_10K | | | |
| High IVOL | -0.76 | 0.48 | 1.24 |
| | (-3.27) | (1.95) | (4.44) |
| Low IVOL | 0.03 | 0.27 | 0.23 |
| | (0.30) | (3.34) | (1.76) |
| High IVOL Sample - | | | 1.01 |
| Low IVOL Sample | | | (3.74) |
| Panel B: Double sort on IO and AIP_10K | | | |
| High IO | -0.17 | 0.23 | 0.40 |
| | (-1.61) | (1.75) | (2.36) |
| Low IO | -0.56 | 0.48 | 1.03 |
| | (-3.53) | (1.91) | (4.41) |
| Low IO Sample - | | | 0.63 |
| High IO Sample | | | (2.52) |
| Panel C: Double sort on analyst coverage and AIP_10K | | | |
| High Coverage | -0.33 | 0.18 | 0.51 |
| | (-3.08) | (1.54) | (3.07) |
| Low Coverage | -0.41 | 0.68 | 1.10 |
| | (-2.59) | (3.23) | (5.77) |
| Low Coverage Sample - | | | 0.58 |
| High Coverage Sample | | | (2.58) |
| Panel D: Double sort on lendable supply and AIP_10K | | | |
| High Lendable Supply | -0.28 | 0.09 | 0.37 |
| | (-2.55) | (0.68) | (2.05) |
| Low Lendable Supply | -0.52 | 0.43 | 0.95 |
| | (-2.59) | (2.03) | (3.53) |
| Low Supply Sample - | | | 0.58 |
| High Supply Sample | | | (1.88) |
| Panel E: Double sort on lending fee and AIP_10K | | | |
| High Lending Fee | -0.66 | 0.49 | 1.14 |
| | (-2.62) | (1.33) | (2.85) |
| Low Lending Fee | -0.27 | -0.01 | 0.26 |
| | (-2.03) | (-0.11) | (1.39) |
| High Fee Sample - | | | 0.88 |
| Low Fee Sample | | | (2.07) |

Table 6: **Variation in the Complexity of 10-K Filings**

This table reports the return predictability results for variation in the complexity of 10-K filings. For each month, we run cross-sectional regression of the log of filing size and number of words on the log of a firm's market capitalization, and use the regression residual as our proxy for filing complexity. We sort stocks into terciles based on the residual size or residual number of words of the most recent 10-K filing. We then independently sort stocks into quintiles based on the abnormal number of IPs searching for 10-K filings (AIP_10K). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. We report the Carhart (1997) four-factor alpha (in percentage) of the lowest and highest AIP portfolios in the lowest and highest filing complexity groups. The "High-Low" column reports the Carhart (1997) four-factor alpha of the high-AIP minus low-AIP portfolios. In the bottom row of each panel, we report the difference of four-factor alphas between the high and low filing complexity groups. T-statistics are in brackets. The sample runs from January 2003 to December 2014.

Panel A: Double sort on residual file size and AIP_10K

| | Low AIP_10K | High AIP_10K | High-Low |
|---|---|---|---|
| Large Filing Size | -0.65 | 0.59 | 1.24 |
| | (-4.22) | (3.23) | (4.88) |
| Small Filing Size | -0.27 | 0.38 | 0.66 |
| | (-1.84) | (2.68) | (3.30) |
| Large Filing Size - | | | 0.58 |
| Small Filing Size | | | (2.29) |

Panel B: Double sort on word count and AIP_10K

| | Low AIP_10K | High AIP_10K | High-Low |
|---|---|---|---|
| More word count | -0.48 | 0.52 | 1.00 |
| | (-3.29) | (2.39) | (5.06) |
| Lesser word count | -0.36 | 0.20 | 0.56 |
| | (-3.02) | (1.35) | (2.93) |
| More word count - | | | 0.44 |
| Lesser word count | | | (1.99) |

Table 7: **Cross-sectional Variation at IP level**

This table reports the return predictability results for IPs searching for 10-K filings. In Panel A, for each stock-month, we compute the number of unique IPs that searched only the current 10-K filings and both the current and historical filings, where current (historical) 10-K is defined as 10-Ks filed after (before) the most recent 10-K filing date. In Panel B, for each stock-month, we compute the number of unique IPs that searched the firm's 10-Ks only in night time (6pm of day $t$ to 6am of day $t+1$) and day time (6am of day $t$ to 6pm of day $t$). We then sort stocks into deciles based on the abnormal number of IPs within each category (AIP_10K). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. We report the Carhart (1997) four-factor alpha (in percentage) of the lowest and highest AIP decile portfolios. The "High-Low" column reports the Carhart (1997) four-factor alpha of the high-AIP minus low-AIP portfolios. In the bottom row of each panel, we report the difference of four-factor alphas between the two categories. T-statistics are in brackets. The sample runs from January 2003 to December 2014.

Panel A: EDGAR searching for current and historical filings

| | Low AIP_10K | High AIP_10K | High-Low |
|---|---|---|---|
| Current filings | -0.41 | 0.21 | 0.61 |
| | (-2.29) | (1.29) | (3.08) |
| Both current and historical filings | -0.45 | 0.55 | 1.00 |
| | (-4.54) | (3.63) | (5.28) |
| Both current and historical filings - | | | 0.39 |
| Current filings | | | (2.53) |

Panel B: Daytime and Nighttime searches

| | Low AIP_10K | High AIP_10K | High-Low |
|---|---|---|---|
| Nighttime search | -0.39 | 0.43 | 0.82 |
| | (-3.25) | (2.92) | (4.71) |
| Daytime search | -0.35 | 0.45 | 0.79 |
| | (-3.23) | (3.15) | (4.70) |
| Nighttime search - | | | 0.03 |
| Daytime search | | | (0.25) |

Table 8: **Fama-MacBeth Regression**

This table reports the results of the Fama and MacBeth (1973) regression of monthly stock returns on the abnormal number of IPs searching for SEC filings through the EDGAR system (AIP). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | Dependent Variable: One-month-ahead stock return | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| AIP_total | 0.0060*** | | | 0.0053*** | | | 0.0050*** | | |
| | (2.68) | | | (2.64) | | | (2.88) | | |
| AIP_funtl | | 0.0047*** | | | 0.0041*** | | | 0.0042*** | |
| | | (2.70) | | | (2.78) | | | (2.94) | |
| AIP_10K | | | 0.0051*** | | | 0.0046*** | | | 0.0044*** |
| | | | (3.73) | | | (3.81) | | | (3.74) |
| REV | | | | -0.0247*** | -0.0245*** | -0.0247*** | -0.0283*** | -0.0281*** | -0.0284*** |
| | | | | (-3.18) | (-3.16) | (-3.19) | (-3.74) | (-3.72) | (-3.75) |
| LnME | | | | -0.0006 | -0.0006 | -0.0006 | -0.0014** | -0.0014** | -0.0014** |
| | | | | (-0.89) | (-0.92) | (-0.93) | (-2.59) | (-2.60) | (-2.58) |
| LnBM | | | | 0.0019 | 0.0019 | 0.0019 | 0.0014 | 0.0013 | 0.0013 |
| | | | | (1.64) | (1.59) | (1.58) | (1.29) | (1.24) | (1.24) |
| MOM | | | | -0.0058 | -0.0057 | -0.0058 | -0.0048 | -0.0047 | -0.0048 |
| | | | | (-0.95) | (-0.94) | (-0.94) | (-0.88) | (-0.86) | (-0.86) |
| IVOL | | | | | | | -0.0015 | -0.0025 | -0.0007 |
| | | | | | | | (-0.02) | (-0.04) | (-0.01) |
| Turnover12 | | | | | | | -0.0094 | -0.0091 | -0.0089 |
| | | | | | | | (-1.37) | (-1.32) | (-1.28) |
| IO | | | | | | | 0.0122*** | 0.0119*** | 0.0114*** |
| | | | | | | | (4.00) | (3.94) | (3.86) |
| Constant | 0.0123** | 0.0122** | 0.0122** | 0.0122 | 0.0122* | 0.0123* | 0.0119** | 0.0120** | 0.0119** |
| | (2.18) | (2.18) | (2.18) | (1.65) | (1.66) | (1.67) | (2.33) | (2.36) | (2.35) |
| Ave.R-sq | 0.003 | 0.003 | 0.003 | 0.030 | 0.030 | 0.030 | 0.046 | 0.046 | 0.046 |
| N.of Obs. | 483667 | 483667 | 483667 | 483667 | 483667 | 483667 | 480793 | 480793 | 480793 |

## Table 9: **Abnormal Number of IPs and Firm Fundamentals**

This table reports the results of the panel regression of future change in firm fundamentals on the abnormal number of IPs searching for SEC filings in the EDGAR system in month $t$. AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. The dependent variable in Columns (1) to (3) is the change of quarterly Return-on-Assets from four quarters ago. In Column (4) to (6), the dependent variable is the standardized unexpected earnings (SUE), defined as the change of quarterly EPS from four quarters ago divided by stock prices 12 months ago. The dependent variable in Columns (7) to (9) is the monthly revision of analysts consensus forecast for annual EPS. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. Coverage is log one plus analyst coverage. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). We control for the year-quarter fixed effects in Columns (1) to (6) and the year-month fixed effects in Columns (7) to (9). Turnover12 is the monthly turnover ratio averaged over the past 12 months. Standard errors are double clustered at both firm and time level. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | Change of ROA | | | SUE | | | Forecast Revision | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| AIP_total | 0.0017* | | | 0.0013 | | | 0.0007*** | | |
| | (1.96) | | | (1.42) | | | (2.78) | | |
| AIP_fundl | | 0.0026** | | | 0.0026** | | | 0.0016*** | |
| | | (2.51) | | | (2.22) | | | (6.19) | |
| AIP_10K | | | 0.0028*** | | | 0.0043*** | | | 0.0019*** |
| | | | (2.92) | | | (3.57) | | | (5.28) |
| LROA | -0.3425*** | -0.3428*** | -0.3430*** | | | | | | |
| | (-4.71) | (-4.73) | (-4.74) | | | | | | |
| LnME | 0.0008 | 0.0008 | 0.0008 | -0.0022*** | -0.0021*** | -0.0021*** | -0.0005 | -0.0005 | -0.0005 |
| | (1.27) | (1.31) | (1.33) | (-3.73) | (-3.68) | (-3.63) | (-1.51) | (-1.55) | (-1.63) |
| LnBM | -0.0013 | -0.0012 | -0.0012 | -0.0009 | -0.0009 | -0.0008 | -0.0008** | -0.0008** | -0.0009** |
| | (-0.87) | (-0.84) | (-0.83) | (-0.50) | (-0.48) | (-0.44) | (-2.37) | (-2.39) | (-2.47) |
| MOM | 0.0100*** | 0.0099*** | 0.0100*** | 0.0220*** | 0.0220*** | 0.0219*** | 0.0025*** | 0.0025*** | 0.0025*** |
| | (3.55) | (3.56) | (3.57) | (8.19) | (8.17) | (8.18) | (5.20) | (5.14) | (5.18) |
| Coverage | 0.0004 | 0.0004 | 0.0005 | 0.0002 | 0.0003 | 0.0003 | 0.0021*** | 0.0021*** | 0.0021*** |
| | (0.29) | (0.31) | (0.32) | (0.23) | (0.27) | (0.29) | (3.30) | (3.29) | (3.28) |
| Turnover12 | -0.0118** | -0.0117** | -0.0117** | 0.0316*** | 0.0318*** | 0.0319*** | -0.0082*** | -0.0082*** | -0.0082*** |
| | (-2.43) | (-2.42) | (-2.43) | (3.45) | (3.47) | (3.47) | (-3.15) | (-3.16) | (-3.17) |
| IO | -0.0010 | -0.0011 | -0.0013 | -0.0093*** | -0.0095*** | -0.0096*** | 0.0049*** | 0.0051*** | 0.0052*** |
| | (-0.48) | (-0.51) | (-0.61) | (-3.75) | (-3.87) | (-3.97) | (5.47) | (5.61) | (5.73) |
| IVOL | -0.0777 | -0.0773 | -0.0775 | 0.2287** | 0.2330** | 0.2365** | -0.1111** | -0.1115** | -0.1138** |
| | (-1.41) | (-1.41) | (-1.42) | (2.29) | (2.32) | (2.34) | (-2.35) | (-2.37) | (-2.41) |
| Time FE | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| Adj.R-sq | 0.056 | 0.056 | 0.056 | 0.023 | 0.023 | 0.023 | 0.002 | 0.002 | 0.002 |
| N.of Obs. | 128504 | 128504 | 128504 | 150712 | 150712 | 150712 | 348130 | 348130 | 348130 |

Table 10: **Identifying Mispricing using Mutual Fund Outflows**

This table reports the results of the Fama and MacBeth (1973) regression of the quarterly change in the abnormal number of IPs searching for SEC filings on quarterly mutual fund outflows. Outflows is calculated following Edmans, Goldstein, and Jiang (2012). In Columns (1) and (2), the dependent variable is the quarterly change in AIP_total in the quarter in which mutual fund outflows occur. In Columns (3) and (4), the dependent variable is the quarterly change in AIP_funtl. In Columns (5) and (6), the dependent variable is the quarterly change in AIP_10K. LnME is the natural log of a firm's market capitalization at the end of June of each year in millions of US dollars. Coverage is log one plus analyst coverage. Turnover12 is the monthly turnover ratio averaged over the past 12 months. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | dAIP_total | | dAIP_funtl | | dAIP_10K | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Outflows | -1.9303** | -1.5459** | -1.9145*** | -1.3527*** | -2.4242*** | -1.7256*** |
| | (-2.06) | (-2.31) | (-3.36) | (-3.27) | (-4.02) | (-4.92) |
| LnME | | -0.0091*** | | -0.0094*** | | -0.0093*** |
| | | (-6.03) | | (-5.68) | | (-5.81) |
| LnBM | | 0.0013 | | -0.0014 | | -0.0017 |
| | | (0.56) | | (-0.57) | | (-0.75) |
| Coverage | | 0.0080*** | | 0.0076*** | | 0.0087*** |
| | | (4.50) | | (4.28) | | (3.70) |
| IVOL | | -1.8233*** | | -1.9963*** | | -1.8354*** |
| | | (-6.48) | | (-7.68) | | (-6.19) |
| Turnover12 | | -0.0015 | | 0.0158 | | 0.0203 |
| | | (-0.09) | | (1.13) | | (1.56) |
| IO | | -0.0023 | | -0.0141** | | -0.0143** |
| | | (-0.36) | | (-2.54) | | (-2.28) |
| MOM | | -0.0336*** | | -0.0370*** | | -0.0398*** |
| | | (-5.17) | | (-5.70) | | (-7.68) |
| Constant | 0.0007 | 0.0901*** | 0.0050** | 0.1036*** | 0.0049** | 0.0967*** |
| | (0.29) | (7.79) | (2.09) | (8.54) | (2.06) | (6.54) |
| Ave.R-sq | 0.001 | 0.031 | 0.001 | 0.034 | 0.001 | 0.026 |
| N.of Obs. | 131863 | 131041 | 131863 | 131041 | 131863 | 131041 |

Table 11: **Anomaly-based Mispricing Measure and Abnormal Number of IPs**

Panel A of this table reports the average abnormal number of IPs for quintile portfolios sorted on composite mispricing score (CMS). The composite mispricing measure is the average of the ranking percentiles produced by 11 anomaly variables following Stambaugh, Yu, and Yuan (2015). AIP is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for SEC filings on a set of firm characteristics. Panel B and C report the equal-weighted monthly Carhart (1997) four-factor alphas (in percentages) and the average composite mispricing score of portfolios double sorted by stock's composite mispricing score and the abnormal number of IPs searching 10-K filings (AIP_10K), respectively. In the end of each month, all the stocks are sorted into quintiles based on composite mispricing score. We then independently sort the stocks into quintiles based on their AIP_10K. We also report, for each mispricing quintile, the high-AIP minus low-AIP portfolio alpha and CMS. The sample runs from January 2003 to December 2014.

Panel A: Abnormal Number of IPs across Composite Mispricing Measure Sorted Portfolios

| | AIP_10K | AIP_funtl | AIP_total |
|---|---|---|---|
| Most Undervalued | 0.23 | 0.16 | 0.13 |
| 2 | 0.13 | 0.08 | 0.06 |
| 3 | 0.07 | 0.04 | 0.02 |
| 4 | 0.01 | 0.00 | -0.01 |
| Most Overvalued | -0.04 | 0.00 | 0.01 |
| Most Undervalued - Most Overvalued | 0.27 | 0.17 | 0.12 |
| t-stat | (32.78) | (24.75) | (19.48) |

Panel B: Two-way sorts on AIP and Composite Mispricing Measure (alpha)

| | Most Undervalued | 2 | 3 | 4 | Most Overvalued |
|---|---|---|---|---|---|
| Low AIP | -0.05 | -0.24 | -0.20 | -0.23 | -0.45 |
| 2 | 0.06 | 0.19 | 0.11 | 0.15 | -0.28 |
| 3 | 0.32 | 0.44 | 0.31 | 0.20 | -0.40 |
| 4 | 0.43 | 0.43 | 0.57 | 0.38 | -0.28 |
| High AIP | 1.05 | 0.69 | 0.52 | 0.59 | -0.08 |
| High - Low | **1.10** | **0.93** | **0.72** | **0.82** | 0.37 |
| t-stat | (4.42) | (5.28) | (3.45) | (3.77) | (1.42) |

Panel C: Two-way sorts on AIP and Composite Mispricing Measure (CMS)

| | Most Undervalued | 2 | 3 | 4 | Most Overvalued |
|---|---|---|---|---|---|
| Low AIP | 0.362 | 0.445 | 0.501 | 0.562 | 0.675 |
| 2 | 0.361 | 0.445 | 0.501 | 0.562 | 0.673 |
| 3 | 0.359 | 0.444 | 0.501 | 0.561 | 0.671 |
| 4 | 0.356 | 0.444 | 0.500 | 0.561 | 0.669 |
| High AIP | 0.353 | 0.444 | 0.500 | 0.560 | 0.668 |
| High - Low | -0.009 | -0.001 | -0.001 | -0.001 | -0.007 |
| t-stat | (-10.90) | (-4.70) | (-5.75) | (-5.55) | (-7.27) |

Table 12: **Abnormal Number of IPs and Investor Trading**

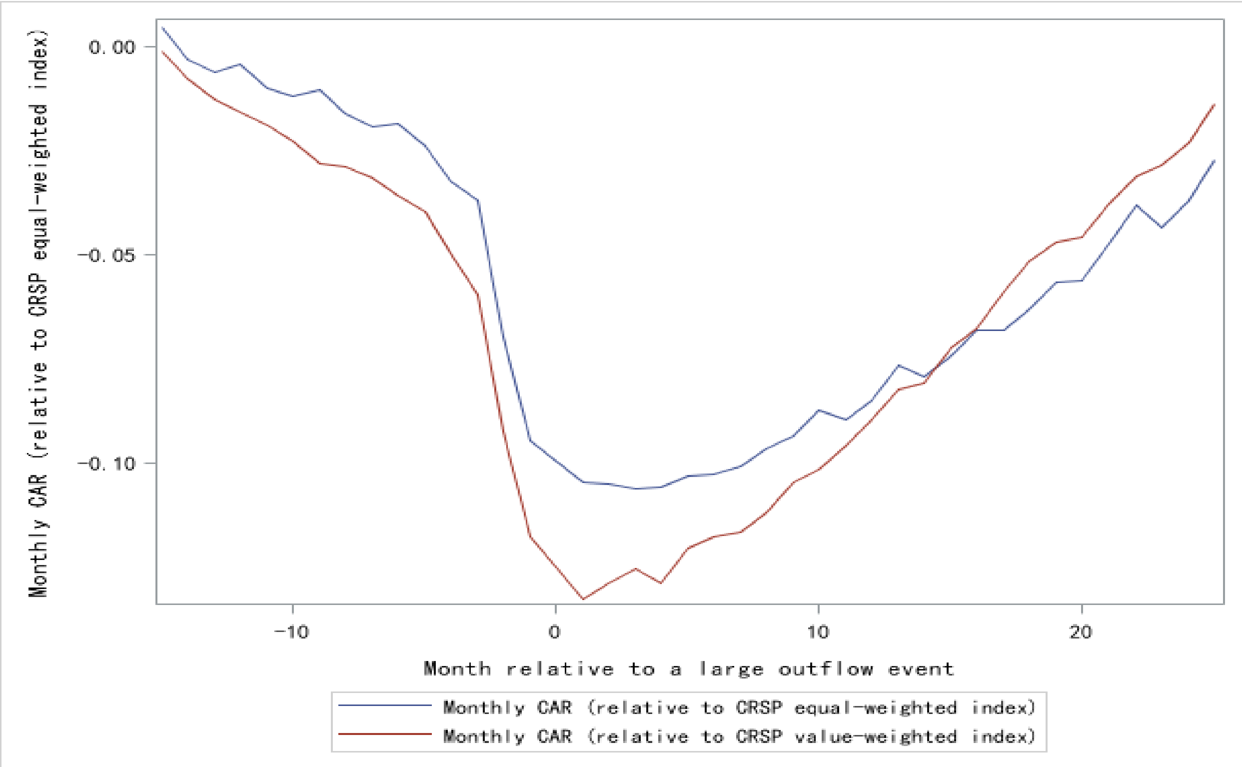
This table reports the results from the Fama and MacBeth (1973) regression of investor trading on lagged abnormal number of IPs searching for SEC filings in the EDGAR system. In Columns (1) to (3), the dependent variable is quarterly net purchases by mutual funds. Net purchase is measured as the quarterly change in mutual fund holding on a stock, with holding expressed as a fraction of a firm's shares outstanding. In Columns (4) to (6), the dependent variable is monthly retail order imbalance. Retail order imbalance is defined as the difference between daily retail buy volume and retail sell volume, scaled by total daily retail trading volume, and then aggregated to monthly level. Retail buys and sells are classified as in Boehmer, Jones, and Zhang (2017). Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. Coverage is log one plus analyst coverage. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively. The sample in Columns (1) to (3) runs from January 2003 to December 2014. The sample in Columns (4) to (6) runs from January 2010 to December 2014.

| | Net Purchases by Mutual Funds | | | Retail Order Imbalance | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| AIP_total | 0.0030 | | | 0.0090*** | | |
| | (0.52) | | | (7.16) | | |
| AIP_funtl | | 0.0046 | | | 0.0079*** | |
| | | (0.77) | | | (6.89) | |
| AIP_10K | | | 0.0065 | | | 0.0076*** |
| | | | (1.00) | | | (7.81) |
| LnME | -0.0003 | -0.0003 | -0.0003 | -0.0002 | -0.0002 | -0.0002 |
| | (-1.07) | (-1.08) | (-1.06) | (-0.20) | (-0.22) | (-0.25) |
| LnBM | 0.0003 | 0.0004 | 0.0003 | 0.0045*** | 0.0045*** | 0.0044*** |
| | (0.32) | (0.38) | (0.34) | (4.95) | (4.96) | (4.84) |
| Coverage | -0.0008 | -0.0007 | -0.0005 | -0.0025*** | -0.0026*** | -0.0027*** |
| | (-0.30) | (-0.26) | (-0.22) | (-3.46) | (-3.61) | (-3.76) |
| IVOL | -0.1965* | -0.1975* | -0.2012* | -0.3268*** | -0.3257*** | -0.3243*** |
| | (-1.96) | (-1.94) | (-1.95) | (-6.41) | (-6.40) | (-6.37) |
| Turnover12 | -0.0069** | -0.0068** | -0.0064** | -0.0350*** | -0.0351*** | -0.0352*** |
| | (-2.04) | (-2.22) | (-2.38) | (-12.83) | (-12.79) | (-12.75) |
| IO | 0.0739*** | 0.0736*** | 0.0728*** | 0.0109*** | 0.0117*** | 0.0124*** |
| | (2.86) | (2.90) | (2.92) | (3.01) | (3.26) | (3.42) |
| MOM | 0.0066*** | 0.0067*** | 0.0065*** | -0.0030 | -0.0031* | -0.0031* |
| | (6.44) | (6.29) | (7.74) | (-1.67) | (-1.73) | (-1.70) |
| Constant | 0.0048* | 0.0050* | 0.0053* | 0.0469*** | 0.0467*** | 0.0465*** |
| | (1.95) | (1.91) | (1.96) | (6.63) | (6.61) | (6.58) |
| Ave.R-sq | 0.113 | 0.113 | 0.113 | 0.010 | 0.010 | 0.010 |
| N.of Obs. | 131795 | 131795 | 131795 | 184715 | 184715 | 184715 |

# Appendices

## Online Appendix to "Information Acquisition and Stock Returns: Evidence from EDGAR Search Traffic"

Figure A1: **Effect of Mutual Funds Hypothetical Sales on Stock Prices**




This figure plots the monthly cumulative average abnormal returns (CAR) of stocks around the event months, where an event is defined as a firm-quarter observation in which mutual fund fire sale induced outflows falls below the 10th percentile value of the full sample. Outflows is calculated following Edmans, Goldstein, and Jiang (2012). CAR is computed over the benchmark of the CRSP equal-weighted (blue line) or value-weighted index (red line) from 15 months before the event to 24 months after.

Table A1: **Returns and Alphas of Portfolios Sorted by Raw Number of IPs**

This table reports the monthly excess returns and Carhart (1997) four-factor alphas (in percentage) for decile portfolios sorted by the raw number of IPs searching for SEC filings. At the end of each month, all stocks are sorted into deciles based on their raw numbers of IPs, and a long-short portfolio is formed by buying the highest decile and shorting the lowest decile portfolio. Portfolio returns are computed over the next month. Panel A reports the results for equally weighted excess return and Panel B shows the results Carhart (1997) four-factor alphas. The sample runs from January 2003 to December 2014.

Panel A: Equal-weighted Decile Portfolio Excess Return

| | IP_10K | t-stat | IP_funtl | t-stat | IP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | 0.73 | 2.04 | 0.87 | 2.62 | 0.73 | 2.17 |
| 2 | 0.80 | 1.87 | 0.80 | 1.90 | 0.92 | 2.17 |
| 3 | 0.63 | 1.32 | 0.91 | 1.89 | 1.01 | 2.19 |
| 4 | 0.95 | 1.86 | 1.12 | 2.28 | 1.12 | 2.22 |
| 5 | 1.05 | 2.01 | 0.89 | 1.73 | 1.12 | 2.19 |
| 6 | 1.12 | 2.10 | 1.17 | 2.23 | 1.07 | 2.08 |
| 7 | 1.12 | 2.07 | 1.12 | 2.11 | 1.01 | 1.92 |
| 8 | 1.22 | 2.25 | 1.05 | 1.91 | 1.14 | 2.06 |
| 9 | 1.19 | 2.26 | 1.04 | 1.96 | 0.99 | 1.84 |
| High | 1.10 | 2.31 | 1.09 | 2.20 | 0.98 | 1.99 |
| High - Low | 0.37 | 1.58 | 0.22 | 0.68 | 0.26 | 1.19 |

Panel B: Equal-weighted Decile Portfolio 4-factor alpha

| | IP_10K | t-stat | IP_funtl | t-stat | IP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | 0.04 | 0.23 | 0.18 | 1.18 | 0.05 | 0.30 |
| 2 | -0.12 | -0.78 | -0.05 | -0.39 | 0.06 | 0.44 |
| 3 | -0.26 | -1.96 | -0.07 | -0.54 | 0.08 | 0.68 |
| 4 | -0.08 | -0.59 | 0.05 | 0.35 | 0.00 | 0.00 |
| 5 | -0.08 | -0.70 | -0.11 | -0.94 | 0.01 | 0.08 |
| 6 | -0.01 | -0.12 | -0.02 | -0.17 | -0.09 | -0.83 |
| 7 | 0.01 | 0.15 | -0.08 | -0.96 | -0.09 | -0.94 |
| 8 | 0.06 | 0.77 | -0.08 | -0.73 | -0.11 | -1.17 |
| 9 | 0.05 | 0.49 | -0.05 | -0.49 | -0.13 | -1.33 |
| High | 0.13 | 1.49 | -0.02 | -0.20 | -0.05 | -0.50 |
| High - Low | 0.09 | 0.47 | -0.20 | -1.15 | -0.09 | -0.56 |

Table A2: **Robustness of Decile Portfolio Sorts**

This table reports the results of several robustness tests for a long/short portfolio based on the abnormal number of IPs searching for 10-K filings in the EDGAR system (AIP_10K). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. For the first robustness test, we report the gross return-weighted portfolio returns, for which the weights are 1 + the stock's lagged monthly return, following Asparouhova, Bessembinder, and Kalcheva (2013). The second robustness test shows the portfolio returns adjusted using the DGTW method. The third set of robustness tests shows the Fama-French 48 industry-adjusted excess return. The fourth row shows the alpha using the Pástor and Stambaugh (2003) liquidity factor augmented with the Fama-French factors and the momentum factor. For the fifth set of tests, we report the alphas using the Fama and French (2015) Five Factor model. For the sixth and seventh sets of tests, we report the alphas using the Stambaugh and Yuan (2016) Mispricing Factors model and the Hou, Xue, and Zhang (2015) Q-factor model. For the eighth set of analyses, we exclude stocks whose market capitalizations are in the bottom quintile based on NYSE size breakpoints. In the ninth row, we skip six months between the moment an abnormal IP is constructed and the moment at which we start measuring returns. In the tenth and eleventh rows, we report the four-factor alpha for two sub-sample periods, one from 2003 to 2008 and the another from 2009 to 2014. The last row report the four-factor alpha after removing the financial crisis period (year 2008 and 2009). T-statistics are in brackets. Returns and alphas are reported in percentage.

| | EW | VW |
|---|---|---|
| Gross return-weighed portfolio | 1.096 | NA |
| | (5.16) | |
| DGTW adjusted | 0.910 | 0.410 |
| | (4.51) | (2.22) |
| FF48 Industry-adjusted | 0.739 | 0.155 |
| | (3.26) | (1.16) |
| FF + Cahart + PS Factor | 0.800 | 0.348 |
| | (4.23) | (1.78) |
| FF five factor (2015) | 0.685 | 0.248 |
| | (3.36) | (1.19) |
| Mispricing factors (Stambaugh and Yuan 2017) | 0.892 | 0.276 |
| | (4.42) | (1.35) |
| Q-factor (Hou, Xue and Zhang 2015) | 0.897 | 0.183 |
| | (4.66) | (0.87) |
| Remove microcap stocks | 0.518 | 0.276 |
| | (2.58) | (1.35) |
| Skip six months | 0.532 | 0.266 |
| | (2.23) | (1.28) |
| 2003-2008 | 0.620 | 0.261 |
| | (2.41) | (0.89) |
| 2009-2014 | 1.073 | 0.121 |
| | (3.74) | (0.45) |
| Remove financial crisis period | 0.733 | 0.116 |
| | (3.87) | (0.56) |

Table A3: **Alternative Implementations of AIP**

This table reports several alternative implementations of AIP_10K when calculating the long/short portfolio Carhart (1997) four-factor alpha (in percentage). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. In the first row, we calculate AIP_10K using model (9) of equation (1). In the second row, we also include the square term of the four firm characteristics when calculating AIP. In the third row, we include the lagged number of IPs in the expected IP regression. Column (1) reports the results for the equal-weighted portfolio, and Column (2) reports for the value-weighted portfolio. T-statistics are in brackets. The sample runs from January 2003 to December 2014.

| | EW | VW |
|---|---|---|
| Model (9) of Expected IP Regression | 0.672 | 0.082 |
| | (3.92) | (0.42) |
| Nonlinear functional form of Expected IP Regression | 0.689 | 0.552 |
| | (4.30) | (2.39) |
| Control for lagged # of IPs in Expected IP Regression | 0.698 | 0.508 |
| | (5.44) | (2.03) |

Table A4: **Alphas of Portfolios Sorted by Within-Firm Changes of AIP**

This table reports the monthly Carhart (1997) four-factor alphas (in percentage) for decile portfolios sorted by changes in AIP relative to its 12-month moving average (dAIP). In the end of each month, all stocks are sorted into deciles based on their dAIP, and a long-short portfolio is formed by buying the highest decile and shorting the lowest decile portfolio. Portfolio returns are computed over the next month. Panel A reports the results for equally-weighted portfolios and Panel B shows the results for value-weighted portfolios. The sample runs from January 2004 to December 2014.

Panel A: Equal-weighted Decile Portfolio 4-factor alpha

| | dAIP_10K | t-stat | dAIP_funtl | t-stat | dAIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | -0.45 | -2.77 | -0.36 | -2.19 | -0.38 | -2.62 |
| 2 | -0.08 | -0.82 | -0.03 | -0.24 | 0.00 | 0.01 |
| 3 | 0.22 | 1.94 | 0.02 | 0.18 | 0.19 | 1.42 |
| 4 | 0.21 | 2.15 | 0.20 | 0.99 | 0.18 | 1.55 |
| 5 | 0.19 | 2.04 | 0.23 | 2.53 | 0.21 | 1.64 |
| 6 | 0.16 | 0.90 | 0.27 | 2.09 | 0.21 | 1.24 |
| 7 | 0.22 | 1.49 | 0.22 | 1.77 | 0.34 | 2.63 |
| 8 | 0.23 | 2.14 | 0.19 | 1.48 | 0.28 | 2.91 |
| 9 | 0.42 | 3.72 | 0.23 | 1.79 | 0.32 | 2.67 |
| High | 0.43 | 2.36 | 0.27 | 1.81 | 0.36 | 2.74 |
| High - Low | **0.88** | 4.82 | **0.63** | 3.27 | **0.74** | 3.65 |

Panel B: Value-weighted Decile Portfolio 4-factor alpha

| | dAIP_10K | t-stat | dAIP_funtl | t-stat | dAIP_total | t-stat |
|---|---|---|---|---|---|---|
| Low | -0.24 | -1.30 | 0.05 | 0.24 | -0.10 | -0.46 |
| 2 | -0.20 | -1.15 | 0.00 | 0.03 | -0.18 | -1.38 |
| 3 | 0.23 | 1.52 | 0.25 | 1.38 | 0.18 | 1.35 |
| 4 | 0.26 | 1.41 | 0.13 | 0.89 | 0.08 | 0.56 |
| 5 | 0.39 | 2.30 | 0.21 | 1.69 | -0.01 | -0.07 |
| 6 | 0.15 | 1.65 | 0.04 | 0.33 | 0.22 | 1.37 |
| 7 | 0.14 | 0.97 | 0.11 | 0.84 | 0.11 | 0.77 |
| 8 | -0.14 | -0.96 | 0.18 | 1.17 | 0.17 | 1.05 |
| 9 | 0.19 | 0.80 | 0.07 | 0.35 | 0.06 | 0.32 |
| High | 0.15 | 0.87 | -0.09 | -0.50 | 0.37 | 1.94 |
| High - Low | 0.39 | 1.44 | -0.14 | -0.46 | **0.47** | 1.73 |

Table A5: **Portfolio Sorts Within Industry**

This table reports the Carhart (1997) four-factor alpha of the long/short portfolio (in percentage) sorted on AIP within each industry of Fama-French 12 industry classification. In the end of each month, all stocks within each industry are sorted into quintiles based on their AIP_10K, and a long-short portfolio is formed by buying the highest quintile and shorting the lowest quintile portfolio. AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K files in the EDGAR database on a set of firm characteristics. The sample runs from January 2003 to December 2014.

| Group | Industry | four-factor alpha | t-stat |
|---|---|---|---|
| 1 | Consumer NonDurables | **0.69** | 2.50 |
| 2 | Consumer Durables | 0.82 | 1.59 |
| 3 | Manufacturing | **0.66** | 2.24 |
| 4 | Energy | **1.06** | 3.31 |
| 5 | Chemicals | **0.78** | 1.81 |
| 6 | Business Equipment | **0.71** | 3.71 |
| 7 | Telecommunications | **0.94** | 2.05 |
| 8 | Utilities | 0.21 | 0.91 |
| 9 | Shops | **0.50** | 1.99 |
| 10 | Health | **0.77** | 2.24 |
| 11 | Financials | **0.48** | 2.39 |
| 12 | Other | **0.65** | 2.75 |

Table A6: **Two-way Sorts by Firm Size and Abnormal Number of IPs**

This table reports the monthly Carhart (1997) four-factor alphas (in percentages) sorted by stock's market capitalization and the abnormal number of IPs searching 10-K filings (AIP_10K). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. In the end of each month, all the stocks are sorted into quintiles based on NYSE size breakpoints. We then independently sort the stocks into quintiles based on their AIP_10K. We also report, for each size quintile, the high-AIP minus low-AIP portfolio alpha. Panel A reports the resuls on an equal-weighted basis and Panel B reports the results on a value-weighted basis. T-statistics are in brackets. The sample runs from January 2003 to December 2014.

Panel A: Equal-weighted four-factor alpha

| | Small firms | 2 | 3 | 4 | Large firms |
|---|---|---|---|---|---|
| Low AIP | -0.51 | -0.14 | -0.27 | -0.17 | -0.19 |
| 2 | -0.19 | -0.17 | -0.22 | -0.04 | -0.23 |
| 3 | -0.13 | 0.09 | -0.04 | 0.01 | -0.02 |
| 4 | 0.17 | 0.11 | 0.10 | 0.16 | 0.20 |
| High AIP | 0.64 | 0.22 | 0.16 | 0.20 | -0.26 |
| High-Low | **1.14** | **0.36** | **0.43** | **0.37** | -0.07 |
| t-stat | (5.38) | (1.72) | (2.01) | (1.68) | (-0.26) |

Panel B: Value-weighted four-factor alpha

| | Small firms | 2 | 3 | 4 | Large firms |
|---|---|---|---|---|---|
| Low AIP | -0.57 | -0.20 | -0.27 | -0.19 | -0.20 |
| 2 | -0.28 | -0.17 | -0.21 | -0.04 | -0.19 |
| 3 | -0.15 | -0.04 | -0.02 | -0.01 | -0.02 |
| 4 | -0.03 | 0.09 | 0.11 | 0.15 | 0.23 |
| High AIP | 0.41 | 0.02 | 0.19 | 0.21 | -0.30 |
| High-Low | **0.98** | 0.22 | **0.46** | **0.40** | -0.10 |
| t-stat | (4.80) | (0.97) | (2.18) | (1.78) | (-0.37) |

Table A7: **Abnormal Number of IPs and Earnings Announcement Returns**

This table reports the results of the Fama and MacBeth (1973) regression of a three-day cumulative abnormal return CAR on the abnormal number of IPs searching for SEC filings through EDGAR system (AIP). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of unique IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. In Columns (1) to (3), abnormal return is calculated as daily stock return minus return on the CRSP value-weighted portfolio return. In Columns (4) to (6), abnormal return is calculated as daily stock return minus the return on the characteristics-matched portfolio following Daniel, Grinblatt, Titman, and Wermers (1997). Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | Market-adjusted CAR(-1,+1) | | | DGTW-adjusted CAR(-1,+1) | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| AIP_total | 0.0020 | | | 0.0019 | | |
| | (1.39) | | | (1.45) | | |
| AIP_fundl | | 0.0025* | | | 0.0024* | |
| | | (1.90) | | | (1.93) | |
| AIP_10K | | | 0.0036*** | | | 0.0033*** |
| | | | (2.74) | | | (2.93) |
| Rev | -0.0001 | 0.0003 | 0.0002 | 0.0000 | 0.0004 | 0.0004 |
| | (-0.02) | (0.13) | (0.08) | (0.01) | (0.17) | (0.19) |
| LnME | 0.0001 | 0.0000 | 0.0001 | 0.0003 | 0.0003 | 0.0003 |
| | (0.17) | (0.05) | (0.16) | (0.54) | (0.46) | (0.52) |
| LnBM | 0.0025** | 0.0024** | 0.0022*** | 0.0023** | 0.0022** | 0.0021** |
| | (2.56) | (2.61) | (2.71) | (2.55) | (2.61) | (2.67) |
| Mom | -0.0021 | -0.0020 | -0.0019 | -0.0013 | -0.0013 | -0.0012 |
| | (-1.54) | (-1.48) | (-1.46) | (-1.25) | (-1.18) | (-1.13) |
| Turnover12 | -0.0188*** | -0.0193*** | -0.0203*** | -0.0208*** | -0.0211*** | -0.0220*** |
| | (-2.68) | (-2.95) | (-3.65) | (-3.83) | (-4.10) | (-5.12) |
| Ivol | -0.0395 | -0.0420 | -0.0402 | -0.0219 | -0.0244 | -0.0228 |
| | (-1.17) | (-1.30) | (-1.11) | (-0.54) | (-0.63) | (-0.53) |
| IO | 0.0153*** | 0.0157*** | 0.0158*** | 0.0147*** | 0.0150*** | 0.0151*** |
| | (6.75) | (6.98) | (7.27) | (6.53) | (6.66) | (6.93) |
| Constant | -0.0041 | -0.0037 | -0.0049 | -0.0051 | -0.0048 | -0.0058 |
| | (-1.37) | (-1.32) | (-1.36) | (-1.39) | (-1.36) | (-1.36) |
| Ave.R-sq | 0.051 | 0.051 | 0.051 | 0.050 | 0.050 | 0.050 |
| N.of Obs. | 121929 | 121929 | 121929 | 121530 | 121530 | 121530 |

## Table A8: **Controlling for Firm Events, Change of Breadth of Ownership and Extreme Returns**

This table reports the results of the Fama and MacBeth (1973) regression of monthly stock returns on the abnormal number of IPs searching for EDGAR filings (AIP). AIP is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all types of files in the EDGAR site on a set of firm characteristics. Columns (1), (4) and (7) show the results for IPs searching for all types of EDGAR filings. Columns (2), (5) and (8) show the results for IPs searching for 10-K, 10-Q, and 8-K files. Columns (3), (6) and (9) show the results for IPs searching for 10-K files. SUE is a firm's standardized unexplained earnings, defined as the realized earnings per share (EPS) minus EPS from four quarters prior, divided by the standard deviation of this difference over the prior eight quarters. EAM is a dummy variable that equals one when a given firm announces quarterly earnings in the month. Upgrade is a dummy equals one when there is an analyst recommendation upgrade in the previous month. Downgrade is a dummy equals one when there is an analyst recommendation downgrade in the previous month. DM is a dummy variable that equals one when there is an ex-dividend event in the previous month. num_8K is the natural log of one plus number of 8-K filings in the previous month. dBreadth is the percentage change of breadth of 13F institutional ownership, following Chen, Hong, and Stein (2002). Following Bali, Cakici, and Whitelaw (2011), the stock's extreme positive return (Maxret) is defined as its maximum daily return in the prior month. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| | AIP_total | AIP_fundl | AIP_10K | AIP_total | AIP_fundl | AIP_10K | AIP_total | AIP_fundl | AIP_10K |
| AIP | 0.0041** | 0.0045*** | 0.0043*** | 0.0042** | 0.0047*** | 0.0043*** | 0.0053*** | 0.0047*** | 0.0046*** |
| | (2.45) | (3.09) | (3.81) | (2.49) | (3.10) | (3.94) | (3.38) | (3.14) | (4.20) |
| REV | -0.0312*** | -0.0309*** | -0.0312*** | -0.0316*** | -0.0312*** | -0.0315*** | -0.0352*** | -0.0351*** | -0.0358*** |
| | (-4.26) | (-4.23) | (-4.27) | (-4.34) | (-4.29) | (-4.34) | (-4.46) | (-4.48) | (-4.54) |
| LnME | -0.0018*** | -0.0018*** | -0.0018*** | -0.0018*** | -0.0018*** | -0.0018*** | -0.0018*** | -0.0018*** | -0.0017*** |
| | (-3.69) | (-3.74) | (-3.72) | (-3.69) | (-3.72) | (-3.72) | (-3.67) | (-3.71) | (-3.73) |
| LnBM | 0.0016 | 0.0015 | 0.0015 | 0.0016 | 0.0015 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| | (1.50) | (1.44) | (1.46) | (1.52) | (1.46) | (1.47) | (1.48) | (1.43) | (1.41) |
| MOM | -0.0065 | -0.0064 | -0.0064 | -0.0065 | -0.0064 | -0.0063 | -0.0065 | -0.0065 | -0.0064 |
| | (-1.15) | (-1.14) | (-1.12) | (-1.14) | (-1.12) | (-1.11) | (-1.16) | (-1.16) | (-1.14) |
| IVOL | 0.0169 | 0.0131 | 0.0174 | 0.0240 | 0.0209 | 0.0220 | -0.0636 | -0.0692 | -0.0768 |
| | (0.24) | (0.18) | (0.24) | (0.34) | (0.29) | (0.31) | (-0.69) | (-0.74) | (-0.78) |
| Turnover12 | -0.0087 | -0.0082 | -0.0084 | -0.0091 | -0.0086 | -0.0089 | -0.0085 | -0.0079 | -0.0080 |
| | (-1.25) | (-1.18) | (-1.19) | (-1.29) | (-1.22) | (-1.24) | (-1.22) | (-1.13) | (-1.14) |
| IO | 0.0118*** | 0.0113*** | 0.0110*** | 0.0120*** | 0.0115*** | 0.0112*** | 0.0120*** | 0.0114*** | 0.0111*** |
| | (3.58) | (3.52) | (3.40) | (3.55) | (3.50) | (3.37) | (3.56) | (3.51) | (3.38) |
| SUE | 0.0028*** | 0.0028*** | 0.0027*** | 0.0028*** | 0.0028*** | 0.0027*** | 0.0027*** | 0.0028*** | 0.0027*** |
| | (8.48) | (8.52) | (8.57) | (8.57) | (8.62) | (8.64) | (8.49) | (8.53) | (8.54) |
| EAM | 0.0033*** | 0.0035*** | 0.0028** | 0.0031** | 0.0033** | 0.0028** | 0.0031** | 0.0032** | 0.0027** |
| | (2.61) | (2.69) | (2.33) | (2.55) | (2.60) | (2.31) | (2.51) | (2.56) | (2.27) |
| Upgrade | 0.0023*** | 0.0023*** | 0.0025*** | 0.0023*** | 0.0023*** | 0.0024*** | 0.0024*** | 0.0024*** | 0.0025*** |
| | (2.76) | (2.76) | (2.95) | (2.79) | (2.77) | (2.94) | (2.89) | (2.90) | (3.03) |
| Downgrade | -0.0010 | -0.0011 | -0.0013 | -0.0009 | -0.0010 | -0.0012 | -0.0013 | -0.0012 | -0.0015* |
| | (-1.00) | (-1.16) | (-1.38) | (-0.90) | (-1.03) | (-1.29) | (-1.54) | (-1.36) | (-1.78) |
| DM | 0.0030*** | 0.0031*** | 0.0031*** | 0.0031*** | 0.0032*** | 0.0031*** | 0.0031*** | 0.0031*** | 0.0031*** |
| | (2.78) | (2.77) | (2.75) | (2.95) | (2.96) | (2.86) | (2.87) | (2.89) | (2.83) |
| num_8K | | | | -0.0010 | -0.0012* | -0.0004 | -0.0010 | -0.0012* | -0.0004 |
| | | | | (-1.55) | (-1.80) | (-0.64) | (-1.51) | (-1.76) | (-0.63) |
| dBreadth | | | | | | | 0.0722 | 0.0825 | 0.0836 |
| | | | | | | | (0.94) | (1.06) | (1.11) |
| Maxret | | | | | | | -0.0308 | -0.0317 | -0.0346 |
| | | | | | | | (-1.52) | (-1.60) | (-1.53) |
| Constant | 0.0121** | 0.0124** | 0.0123** | 0.0125** | 0.0128** | 0.0125** | 0.0123** | 0.0127** | 0.0124** |
| | (2.46) | (2.52) | (2.50) | (2.50) | (2.56) | (2.53) | (2.41) | (2.48) | (2.46) |
| Ave.R-sq | 0.053 | 0.053 | 0.053 | 0.054 | 0.054 | 0.053 | 0.057 | 0.057 | 0.057 |
| N.of Obs. | 443261 | 443261 | 443261 | 443261 | 443261 | 443261 | 442698 | 442698 | 442698 |

Table A9: **Controlling for News Coverage and News Sentiment**

This table reports the results of the Fama and MacBeth (1973) regression of monthly stock returns on the abnormal number of IPs searching for SEC filings (AIP). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. News coverage is the natural logarithm of the number of news article covering the company in a given month in the RavenPack database. News sentiment is the event sentiment score from RavenPack, which indicates how firm-specific news events are categorized and rated as having a positive or negative effect on stock prices by experts with extensive experience and backgrounds in linguistics, finance, and economics. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| AIP_total | 0.0043** | | | 0.0041** | | |
| | (2.33) | | | (2.40) | | |
| AIP_funtl | | 0.0040** | | | 0.0044*** | |
| | | (2.13) | | | (2.86) | |
| AIP_10K | | | 0.0050*** | | | 0.0052*** |
| | | | (3.65) | | | (3.95) |
| REV | -0.0250*** | -0.0232*** | -0.0252*** | -0.0270*** | -0.0267*** | -0.0276*** |
| | (-2.95) | (-2.83) | (-2.96) | (-3.21) | (-3.18) | (-3.24) |
| LnME | -0.0013*** | -0.0013** | -0.0012** | -0.0015*** | -0.0016*** | -0.0015*** |
| | (-2.62) | (-2.60) | (-2.27) | (-2.67) | (-2.74) | (-2.72) |
| LnBM | 0.0011 | 0.0012 | 0.0012 | 0.0006 | 0.0005 | 0.0003 |
| | (1.03) | (1.05) | (1.05) | (0.59) | (0.45) | (0.27) |
| MOM | -0.0053 | -0.0051 | -0.0052 | -0.0050 | -0.0049 | -0.0046 |
| | (-0.84) | (-0.81) | (-0.81) | (-0.75) | (-0.74) | (-0.69) |
| IVOL | 0.1334* | 0.1171 | 0.1363* | 0.1338* | 0.1290* | 0.1396* |
| | (1.66) | (1.52) | (1.68) | (1.73) | (1.69) | (1.74) |
| Turnover12 | -0.0083 | -0.0093 | -0.0070 | -0.0116 | -0.0110 | -0.0105 |
| | (-0.96) | (-1.01) | (-0.82) | (-1.11) | (-1.08) | (-1.05) |
| IO | 0.0082** | 0.0084*** | 0.0070** | 0.0110*** | 0.0111*** | 0.0105*** |
| | (2.56) | (2.82) | (2.12) | (3.74) | (3.71) | (3.61) |
| News Coverage | -0.0004 | -0.0005 | -0.0005 | | | |
| | (-0.66) | (-0.71) | (-0.72) | | | |
| News Sentiment | | | | 0.0161*** | 0.0161*** | 0.0171*** |
| | | | | (3.88) | (3.86) | (3.48) |
| Constant | 0.0153*** | 0.0155*** | 0.0153*** | 0.0122** | 0.0122** | 0.0115** |
| | (2.88) | (2.92) | (2.91) | (2.26) | (2.27) | (2.10) |
| Ave.R-sq | 0.055 | 0.055 | 0.055 | 0.056 | 0.055 | 0.056 |
| N.of Obs. | 264816 | 264816 | 264816 | 264816 | 264816 | 264816 |

Table A10: **Abnormal Number of IPs and Long-horizon Returns**

This table reports the results from the Fama and MacBeth (1973) regression of cumulative returns from month $t+j$ to $t+k$ (Cumret(j,k)) on the abnormal number of IPs searching for 10-K filings in the EDGAR system (AIP_10K) in month $t$. The dependent variable is next quarter return (skipping the immediate month) in Column (1), the second quarter return in Column (2), the second half-year return in Column (3), and the second year return in Column (4). AIP_10K is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for 10-K filings in the EDGAR system on a set of firm characteristics. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | Cumret(2,4) | Cumret(5,7) | Cumret(8,13) | Cumret(14,25) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| AIP_10K | 0.0102*** | 0.0068** | 0.0150 | 0.0175 |
| | (2.95) | (2.05) | (1.57) | (0.64) |
| REV | -0.0072 | 0.0037 | 0.0033 | -0.0451 |
| | (-0.53) | (0.21) | (0.11) | (-0.93) |
| LnME | -0.0023 | -0.0013 | -0.0015 | -0.0048 |
| | (-1.64) | (-1.03) | (-0.61) | (-1.11) |
| LnBM | 0.0046* | 0.0041 | 0.0118** | 0.0197* |
| | (1.72) | (1.57) | (2.36) | (1.79) |
| MOM | -0.0193 | -0.0117 | -0.0300* | -0.0421 |
| | (-1.24) | (-0.88) | (-1.75) | (-1.26) |
| IVOL | 0.0407 | -0.0184 | 0.2652 | 0.5759 |
| | (0.20) | (-0.10) | (0.73) | (0.84) |
| Turnover12 | -0.0165 | -0.0312* | -0.0451 | -0.0488 |
| | (-0.92) | (-1.95) | (-1.53) | (-1.08) |
| IO | 0.0116 | 0.0152** | 0.0414** | 0.0956** |
| | (1.63) | (2.18) | (2.42) | (2.47) |
| Constant | 0.0370** | 0.0281* | 0.0451 | 0.0947 |
| | (2.41) | (1.72) | (1.53) | (1.51) |
| Ave.R-sq | 0.051 | 0.044 | 0.036 | 0.035 |
| N.of Obs. | 469185 | 456068 | 425505 | 360584 |

Table A11: **Which Types of SEC Filings?**

This table reports the results of the Fama and MacBeth (1973) regression of monthly stock returns on the abnormal number of IPs searching for SEC filings (AIP). AIP_total is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for all type of SEC filings in the EDGAR system on a set of firm characteristics. Similarly, AIP_funtl (AIP_10K) is constructed using the number of IPs searching for 10-K, 10-Q, and 8-K (10-K) filings in the EDGAR system. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| Dependent variable: One-month ahead stock returns | | |
|---|---|---|
| | (1) | (2) |
| AIP_total | -0.0014 | -0.0003 |
| | (-0.63) | (-0.17) |
| AIP_fundl | 0.0022 | 0.0012 |
| | (1.11) | (0.70) |
| AIP_10K | 0.0049*** | 0.0043*** |
| | (3.96) | (4.02) |
| REV | | -0.0287*** |
| | | (-3.80) |
| LnME | | -0.0014** |
| | | (-2.52) |
| LnBM | | 0.0013 |
| | | (1.24) |
| MOM | | -0.0048 |
| | | (-0.88) |
| IVOL | | -0.0027 |
| | | (-0.04) |
| Turnover12 | | -0.0088 |
| | | (-1.27) |
| IO | | 0.0112*** |
| | | (3.84) |
| Constant | 0.0122** | 0.0120** |
| | (2.18) | (2.34) |
| Ave.R-sq | 0.005 | 0.048 |
| N.of Obs. | 483667 | 480793 |

Table A12: **Abnormal Number of IPs or Abnormal Number of Searches?**

This table reports the results of the Fama and MacBeth (1973) regression. Asearch is the residual from a monthly regression of log one plus the total number of EDGAR requests for SEC filings. AIP is the residual from a monthly regression of log one plus the total number of unique IP addresses searching for SEC filings on a set of firm characteristics. Columns (1) and (2) show the results for searching for all types of SEC filings. Columns (3) and (4) show the results for searching activities for 10-K, 10-Q, and 8-K filings. Columns (5) and (6) show the results for searching activities for 10-K filings. Size (LnME) is the natural log of a firm's market capitalization at the end of June of each year. Book-to-market (LnBM) is the natural log of the book-to-market ratio. The cases with negative book value are deleted. Momentum (MOM) is defined as the cumulative returns from month t-12 to t-2. The short term reversal measure (REV) is the lagged monthly return. Institutional ownership (IO) is the sum of shares held by institutions from 13F filings in each quarter divided by the total shares outstanding. IVOL is the idiosyncratic volatility, calculated following Ang, Hodrick, Xing, and Zhang (2006). Turnover12 is the monthly turnover ratio averaged over the past 12 months. All t-statistics are Newey-West adjusted with four lags to control for heteroskedasticity and autocorrelation. ***, **, and * represent significance levels of 1%, 5%, and 10%, respectively.

| | All EDGAR Filings | | 10-K, 10-Q, 8-K | | 10-K | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Asearch | 0.0014 | -0.0004 | 0.0020* | -0.0024 | 0.0033*** | -0.0039 |
| | (1.54) | (-0.42) | (1.90) | (-1.49) | (3.93) | (-1.57) |
| AIP | | 0.0055** | | 0.0062*** | | 0.0084*** |
| | | (2.45) | | (2.83) | | (2.90) |
| REV | -0.0283*** | -0.0284*** | -0.0283*** | -0.0284*** | -0.0284*** | -0.0289*** |
| | (-3.73) | (-3.76) | (-3.74) | (-3.77) | (-3.75) | (-3.75) |
| LnME | -0.0014** | -0.0014*** | -0.0014** | -0.0014** | -0.0014*** | -0.0013*** |
| | (-2.59) | (-2.63) | (-2.61) | (-2.52) | (-2.64) | (-3.11) |
| LnBM | 0.0013 | 0.0014 | 0.0014 | 0.0014 | 0.0012 | 0.0015* |
| | (1.26) | (1.31) | (1.34) | (1.36) | (1.13) | (1.71) |
| MOM | -0.0049 | -0.0048 | -0.0048 | -0.0049 | -0.0048 | -0.0049 |
| | (-0.89) | (-0.88) | (-0.87) | (-0.89) | (-0.86) | (-1.15) |
| IVOL | 0.0048 | -0.0014 | 0.0065 | -0.0033 | 0.0039 | -0.0021 |
| | (0.07) | (-0.02) | (0.09) | (-0.05) | (0.05) | (-0.03) |
| Turnover12 | -0.0100 | -0.0096 | -0.0095 | -0.0091 | -0.0095 | -0.0088 |
| | (-1.46) | (-1.39) | (-1.38) | (-1.33) | (-1.37) | (-1.33) |
| IO | 0.0127*** | 0.0123*** | 0.0122*** | 0.0115*** | 0.0120*** | 0.0109*** |
| | (4.10) | (4.04) | (4.06) | (3.86) | (4.03) | (3.57) |
| Constant | 0.0115** | 0.0120** | 0.0116** | 0.0119** | 0.0117** | 0.0120*** |
| | (2.26) | (2.35) | (2.29) | (2.33) | (2.32) | (3.19) |
| Ave.R-sq | 0.046 | 0.047 | 0.046 | 0.048 | 0.046 | 0.049 |
| N.of Obs. | 480793 | 480793 | 480793 | 480793 | 480793 | 480793 |