HARVARD | BUSINESS | SCHOOL

Search Based Peer Firms: Aggregating Investor Perceptions through Internet Co-Searches

Charles M.C. Lee
Paul Ma
Charles C.Y. Wang

Working Paper

13-048

July 2, 2014

Search-Based Peer Firms: Aggregating Investor Perceptions through Internet Co-Searches*

Charles M.C. Lee
Stanford University
Graduate School of Business

Paul Ma
University of Minnesota
Carlson School of Management

Charles C.Y. Wang
Harvard Business School

July 2014
Forthcoming in the *Journal of Financial Economics*

Abstract

Applying a "co-search" algorithm to Internet traffic at the SEC's EDGAR website, we develop a novel method for identifying economically-related peer firms and for measuring their relative importance. Our results show that firms appearing in chronologically adjacent searches by the same individual (Search-Based Peers or SBPs) are fundamentally similar on multiple dimensions. In direct tests, SBPs dominate GICS6 industry peers in explaining cross-sectional variations in base firms' out-of-sample: (a) stock returns, (b) valuation multiples, (c) growth rates, (d) R&D expenditures, (e) leverage, and (f) profitability ratios. We show that SBPs are not constrained by standard industry classification, and are more dynamic, pliable, and concentrated. We also show that co-search intensity captures the degree of similarity between firms. Our results highlight the potential of the collective wisdom of investors — extracted from co-search patterns — in addressing long-standing benchmarking problems in finance.

JEL: D83, G0, M2
Keywords: peer firm, EDGAR search traffic, revealed preference, co-search, industry classification

*The authors can be contacted at professor.lee@gmail.com, paulma@umn.edu, and charles.cy.wang@hbs.edu. We are extremely grateful to Scott Bauguess, as well as others, at the U.S. Securities and Exchange Commission for helpful comments, suggestions, and assistance with the data. We also benefited from advice and suggestions from Ryan Buell, Paul Dunmore, Dirk Jenter, Cristi Gleason, Blake Grossman, Joseph Grundfest, Jerry Hoberg, Stephannie Larocque, Ivan Marinovic, Michael Minnis, Michael Roberts, Tatiana Sandino, John Shoven, Pian Shu, Frank Zhang and seminar participants at the National University of Singapore, University of Sydney, HKUST, University of Oregon, University of Southern California, Penn State University, University of Iowa, the American Accounting Association annual meeting, the American Accounting Association FARS meeting, the Conference on Financial Economics and Accounting at UNC, the Minnesota Empirical Accounting Conference, the HBS IMO Conference, and the NCCU/NTU Accounting Symposium. We also thank an anonymous referee for helping us improve and polish the paper. An earlier version of this paper was previously circulated as "Identifying Peer Firms: Evidence from EDGAR Search Traffic." All views remain our own.

1 Introduction

In various contexts requiring performance attribution, financial researchers have long wrestled with the need to identify peer firms for benchmarking purposes. Since at least King (1966), most of this effort has relied on benchmarking to "industry", an amorphous notion pertaining to a group of economically-related firms. Over the years, the earlier Standard Industry Classification (SIC) scheme and the subsequent North American Industry Classification System (NAICS) industry groups were reorganized (Fama and French, 1997) or supplanted by alternative classification schemes such as the Global Industry Classification Scheme (GICS).

Though the idea of grouping economically-related firms has simple conceptual appeal, it can be difficult to implement in practice, perhaps contributing to the proliferation of alternative classification schemes. For every obvious example of a natural industry benchmark (e.g., Kellogg and General Mills) there are many more pairings (or non-pairings) that are far less intuitive. For example, although Apple may reasonably be grouped with Dell and Hewlett-Packard, it is less clear how closely Astro-Med, which belongs to the same six-digit GICS grouping, is economically related to these firms. Conversely, while Walmart and Target may seem like natural benchmarks for each other, they actually belong to different two-digit GICS sectors.[1] These examples illustrate that benchmarking through standard industry classification schemes can often be an imprecise approximation.

Ultimately, the choice of the appropriate "industry" benchmark relies on some degree of subjective judgment. Indeed, the popular GICS classification was created at least in part by incorporating investors' perceptions (their subjective judgments) of firms' main lines of business, in a departure from the more traditional approach taken by the SIC

[1] Walmart's six-digit GICS code is 301010 (Food, Staples, and Retailing) while Target's is 255030 (Multiline Retail).

and NAICS that group firms on the basis of similarities in production processes or end-products. The fact that GICS "outperforms" both the SIC and the NAICS suggests the potential usefulness of incorporating investor perceptions when developing economically-related benchmarks.[2]

In this paper, we identify and study firms' economic benchmarks as collectively perceived by the users of the Security Exchange Commission's (SEC) Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) website. Since the middle of 1996, the SEC has required all public domestic companies to submit their filings electronically via EDGAR. With an average of over 3,000 filings per day and hundreds of thousands of document views per day by users, the EDGAR site is central to the SEC's mission of "protecting investors, maintaining fair, orderly, and efficient markets, and facilitating capital formation." Despite its importance as a central clearinghouse in the gathering and dissemination of firm-level financial information, little is known about EDGAR users' information acquisition patterns and the latent information embedded therein.[3]

Our maintained assumption is that the population of EDGAR users is collectively searching for firm fundamental information to aid in their investment decisions, and that an important part of this process involves the comparison of these fundamentals to economically related or similar firms. Under this assumption, EDGAR users acquiring the financial information of one firm are also likely to acquire the financial information for benchmark firms. If so, we expect individual users' search sequences to reflect their perceptions of the relevant set of benchmark firms as well as their relative importance.[4]

Following this intuition, we extract investors' collective perceptions of a firm's eco-

[2]See Bhojraj and Lee (2002); Bhojraj et al. (2003); Chan et al. (2007); Lewellen and Metrick (2010).

[3]A recent study by Drake et al. (2012) shows that the volume of web traffic at the EDGAR site is tied to the release of news, particularly negative news that increases uncertainty about a firm. Beyond this study, little is known about how EDGAR data are used.

[4]Some of today's largest web-based service providers and vendors, such as Google, Facebook, Amazon, and Netflix, follow similar logic in making user-specific product recommendations based on customer usage patterns.

nomically related peers by applying a basic co-search algorithm to approximately 3.5 billion document requests at the EDGAR site that took place between 2008 and 2011, where each observation in the sample represents a time-stamped record of a specific "page view" originating from a given unique user in search of a particular SEC document (identified by the filing firm and type of SEC document). For each base firm to peer firm pairing, we compute a bilateral "search fraction" variable, which measures the "distance" between firms, or the degree to which the two firms are associated in the collective minds of EDGAR users. Specifically, for two firms i and j we compute the proportion of daily unique EDGAR visitors viewing firm i's fundamental information who then subsequently viewed firm j's information (i.e., *Annual Search Fraction*). We then define a firm's top "Search-Based Peers" (SBPs) as peer firms with the highest *Annual Search Fraction* — i.e., firms most likely to be "co-searched," or to appear in chronologically adjacent searches, with the base firm by EDGAR users.

We interpret a firm's top SBPs as representing its most relevant economic benchmarks in the collective minds of EDGAR users. Note that the SBPs need not reflect the views of any individual EDGAR user per se; rather, these peer firms represent a composite sketch formed by extracting the latent intelligence across multiple users, each searching according to his or her own views on the relevant benchmarks. To the extent that two firms that are frequently "co-searched" by multiple investors are closely associated in investors' minds, we conjecture that they will exhibit similarities that might not be easily identified by other traditional benchmarking techniques.

Our results confirm that EDGAR users' search sequences are non-random, and that they contain latent information about fundamental firm associations. First, we show that SBPs are related to, but not constrained by, traditional industry classification schemes. Among the 10 closest peers, as identified by search traffic patterns, 49% have the same six-digit GICS code as the base firm and 68% have the same two-digit GICS code. Moreover,

SBPs also exhibit strikingly similar *operational performance and business characteristics* to their base firms. For example, relative to other GICS2 firms, SBPs are much closer to base firms in terms of their past (as well as forecasted) growth rates in sales and earnings. They also have much more similar measures of returns-on-capital (ROEs and ROAs), asset turnover ratios, profit margins, leverage ratios, and even levels of R&D spending relative to sales.

Second, although SBPs share many common traits with firms identified by traditional industry classification schemes, we show across multiple dimensions of fundamental firm characteristics that SBPs transcend the level of comparability commonly associated with traditional industry groupings. Specifically, we find that SBPs perform better than GICS6 peers in explaining the out-of-sample cross-sectional variation in (a) stock returns, (b) valuation multiples, (c) growth rates, (d) R&D expenditures, (e) leverage, and (f) profitability ratios. Compared to a portfolio of 10 random GICS6 peers originating from the S&P1500, an equal-weighted portfolio of the top 10 SBPs explains 64% more of the cross-sectional variation in the S&P1500 base firms' monthly returns as well as on average 84% more of the variation in their valuation multiples, growth rates, profitability, and other financial characteristics.

Our results show that the *Annual Search Fraction* variable has incremental explanatory power, consistent with the variable capturing the degree of similarity (i.e., "distance") between the base firm and its benchmarking peer firm. A portfolio of top 10 SBPs weighted by their relative *Annual Search Fraction* consistently outperforms its equal-weighted counterpart. Compared to a portfolio of random 10 GICS6 peers from the S&P1500, the search-fraction-weighted SBP portfolios explain 81% more of the cross-sectional variation in the S&P1500 base firms' monthly returns as well as on average 101% more of the variation in their valuation multiples, growth rates, profitability, and other financial characteristics. In further analyses, we show that SBPs' advantage is ro-

bust across major industrial sectors and size groupings, and this superior performance continues to hold even when we augment GICS6 peers with various size-based matching techniques. We also show that SBPs' outperformance is robust to various perturbations to the co-search algorithm.

The superior performance of the SBPs seems to stem from several sources. First, our algorithm aggregates information contained in the co-search patterns of 3.2 million users. Our findings suggest that EDGAR users may not solely be looking for similarities in terms of lines-of-business or production processes. Indeed their search patterns reveal a collective wisdom about firm operating performance, leverage, and growth prospects that consistently elude other benchmarking techniques. Second, SBPs reflect a more dynamic and pliable matching algorithm. Whereas the traditional industry classifications organize firms into mutually exclusive and collectively exhaustive groupings that change little over time, the set of closest SBPs can vary from year to year and is not constrained by a rigid structure.

Strikingly, we find that the *Annual Search Fraction* variable behaves in accordance with a power law distribution.[5] In other words, the top few SBPs are much closer in "distance" from the base firm, in terms of their search fraction ownership, than the rest. For example, while the median base firm in 2011 had over 500 SBPs, a quarter of total co-searches was concentrated in the top ten SBPs. The top 3 SBPs, on average, own 14.4% of the base firms' *Annual Search Fraction*.

We interpret these intriguing search statistics in two ways. First, our search fraction can be viewed as the degree of collective agreement across EDGAR users about any particular base-peer firm pairing. Thus, the power law distribution suggests that the agreement about a firm being an economically related peer tends to decay at an exponential rate. A second and stronger interpretation of this finding is that a base firm has a

[5]We test and show that our data is consistent with a power law distribution but we do not rule out other distributions such as a log-normal.

relatively small set of peer firms to which it is economically closely related. This suggests that perhaps traditional industry groupings might be too broad or too rigid to identify the best peer firms for benchmarking purposes.

Taken together, our results suggest a novel approach for identifying a firm's set of fundamentally similar peers as well as their relative importance, and highlight the opportunities presented by co-searching techniques applied to Internet search sequences in extracting the collective wisdom of investors. Along multiple dimensions, the SBPs identified by our algorithm are far more similar to the base firms than peers identified by traditional industry and size groupings. The fact that SBPs can explain, out of sample, an order-of-magnitude more of the cross-sectional variation in various firm attributes compared to GICS suggests that SBPs represent a powerful way of accounting for "industry" effects, which can be useful in standard asset pricing and other performance attribution tests.

We note, however, that our approach also has a number of attendant limitations. Most importantly, future applications of our approach for peer firm identification will depend on data availability. Our understanding is that the SEC plans to expand overall access to EDGAR traffic data due, in part, to this study. Nevertheless, the concept of extracting the collective wisdom of investors through their information acquisition patterns can be extended to a variety of other search traffic on the web. Overall, we believe the findings in our study should be viewed as a promising alternative, rather than a definitively superior approach, to more traditional peer identification techniques.[6]

Our work also contributes to the literature by highlighting the importance and usefulness of fundamental accounting data to market participants. To date, little is known about the role and usefulness of archived EDGAR data to investors. Our evidence sug-

[6] Hoberg and Phillips (2010, 2014) propose another interesting alternative approach to peer identification through textual analysis of firms' self-reported business descriptions as filed in the 10-K reports. We discuss and compare our approach with theirs in more detail in the next section.

gests that, collectively, the users of the financial data are extracting information in a highly sophisticated manner. These findings help shed light on investor information acquisition behavior that complements theoretical work on the subject (Van Nieuwerburgh and Veldkamp, 2010). Overall, our findings are consistent with the view that users access archived information on EDGAR at least in part for benchmarking purposes.

Finally, we contribute to the nascent but growing literature of empirical work on understanding the intelligence latent in web traffic data. While the existing body of work using data, such as Google Trends measures the nominal level of traffic at the firm level, our methodology and data move the analysis to the user level. The higher granularity of our data thus allows for the identification of investors' co-searching patterns not previously documented. In being the first to analyze traffic patterns of this scope, we also highlight several challenges and methodological approaches in analyzing raw search traffic, with particular emphasis on distinguishing between informed and uninformed searches, which should benefit future researchers of this dataset as well as other search traffic datasets. In the context of our study, we provide a simple classification rule to distinguish between human (informed) and robotic (uninformed) searches. To validate our classification rule, we show that our algorithm produces very little intelligence when applied to the robotic searches.

The remainder of the paper is organized as follows. Section 2 provides background on alternative industry classification schemes used in the literature. Section 3 describes our data and empirical strategy. Section 4 reports our empirical results. Section 5 presents robustness checks, and, finally, Section 6 concludes the study.

2 Background and Literature

Four major classifications — the SIC, NAICS, GICS, and Fama French — are commonly used by practitioners as well as researchers in accounting, economics, and finance.[7] The SIC system was first completed in 1937 by the Central Statistical Board of the United States and is the oldest of the four systems. The last change to the SIC hierarchical structure was in 1987. As a replacement, the U.S. Office of Management and Budget (OMB) organized the Economic Classification Policy Committee which subsequently developed the NAICS, officially adopted by the OMB for use by all governmental agencies in 1997.[8] Both NAICS and SIC groupings, however, are developed through the collaboration of several governmental agencies based on production and supply chain processes, which may have little bearing on the approrpriateness of economic benchmarks. In contrast, the GICS scheme is provided through a collaboration between Standard & Poor's (S&P) and Morgan Stanley Capital International. It is based on the judgment of a team of financial analysts who group firms on the basis of their principal business activity, determined either by observable sources of revenues and earnings or by market perceptions.

Earlier work such as Bhojraj et al. (2003) and Chan et al. (2007) provide evidence that GICS outperforms alternative industry classification schemes, suggesting, potentially, the usefulness of exploiting market perceptions of related firms. Taking the cue from GICS, our paper investigates and exploits the usefulness of investor perceptions in benchmarking by extracting perceptions of related firms latent in EDGAR users' information acquisition patterns.

[7]The widespread adoption of an industry classification system can be profitable for its owner, as hedge funds, exchange traded funds, and other financial instruments are charged (often substantial) fees for its usage. In recent years, the commercial success of GICS, launched by MSCI, has led to several other industry grouping schemes. Among these, the best known are probably the Industry Classification Benchmark (ICB) owned by FTSE and launched in 2005, and the Dow Jones Industry Classification System (DJICS), launched in 2011. Our review of these alternatives suggests they are based on very similar classification algorithms to those employed by MSCI GICS. For parsimony, we have not included them in our tests.

[8]For example, the U.S. Census classifies business establishments based on the NAICS.

Our paper also relates to the literature that analyzes the efficacy of industry peer groupings. Earlier work, such as that of Guenther and Rosman (1994) and Kahle and Walkling (1996), compared two-digit SIC codes between Compustat and Center for Research in Security Prices (CRSP) sources and found that they disagreed, on average, 38% of the time; however, using Compustat SIC codes yielded higher intra-industry correlations in stock returns.[9] Other work, such as Fan and Lang (2000), proposes using an input–output table based on commodity flow data to capture relatedness based on whether firms share the same inputs or receive as input the other's output. Their approach, like ours, has the appealing ability to measure the degree of relatedness between firms. One drawback of their methodology is the need for well-specified production processes, which can be particularly problematic for firms in certain industries, such as biotechnology, software, or service industries.

Studies that share the spirit of leveraging the choices of informed parties include that of Ramnath (2002), who used analysts' choice of firm coverage in determining peer firms. The author categorizes firms as peers if at least five analysts covered the same group of firms. De Franco et al. (2011) expand upon this method using a hand-collected sample of analyst reports disclosing the peer firms used by each analyst. The authors find that analysts strategically select peers to bias in favor of the investment thesis of their research report. Perhaps not surprisingly, their results show that 92% of the peers chosen by the analysts came from the same two-digit GICS (GICS2). In contrast, our closest 10 firms have a 68% correspondence with GICS2, which may suggest that EDGAR investors de-bias, to some degree, the choice of peers in analyst recommendations.

Finally our paper complements recent studies which also challenge existing conventions of industry classification. Hoberg and Phillips (2010, 2014) analyze textual similarity in firms' self-reported business descriptions in their 10-K filings to infer product

[9] All of our industry codings, including those for the SIC, are sourced from Compustat.

market-based peers.[10] Like our study, the authors show that it is possible to use additional information to improve on traditional industry-based peer identification schemes. However, the two methods are quite distinct in many other respects. While they nominate an innovative and direct measure of peers based on firms' self-descriptions, we nominate peer firms through perceptions of EDGAR users (investors) as inferred from their search choices. Although our SBPs could partially reflect users' perceptions about similarities in firms' product markets, they are likely to also incorporate other, possibly more nuanced, dimensions of fundamental similarity.

3 Data Description and Defining Search-Based Peers

3.1 Data and Empirical Strategy

Our data come from the SEC, which logs all search traffic on its EDGAR system. The raw dataset comprises of all visits to the EDGAR website from January 2008 to December 2011, totaling 3.5 billion visitor page views. Each observation in the raw data contains information on the visitor's IP address, timestamp, CIK, and accession numbers that uniquely match a particular company's specific SEC filing.[11] Table A1 shows a sample of the EDGAR traffic.

In generating our baseline results, we pre-process the raw "page view" data in six major steps, which are summarized in Table 1. In a pre-step, we discard observations for which there is no match to a particular SEC filing based on the SEC master filing index.[12] Then, in our first step, we restrict searches to base firms in the S&P 1500 universe, leaving

[10] Building upon the same SEC disclosure requirement, and also with a view to capture firms in similar product spaces, Rauh and Sufi (2012) and Lewellen (2013) both identify peers through firms' self-reported named competitors in the business description of the 10-K filings.

[11] We do not have the original IP but, rather, an ID that allows us to differentiate between different IPs.

[12] This occurrence is extremely rare and is likely a data error or a deleted filing.

811 million visitor page view observations, or roughly a quarter of the entire sample.[13]

In the second data processing step, we restrict to daily IP addresses which search for at least two unique firms on a given day, which, as we explain below, is required to apply our co-search algorithm. This step reduces our overall sample to 746 million visitor page view observations.

In the third data processing step, we attempt to eliminate search traffic generated by robots or by automated scripts written to download massive numbers of filings. Robot-generated downloads are removed from the sample because we believe (and later test and show) they would be uninformative for our analysis.

We devise a simple heuristic cutoff rule for identifying robot traffic based the number of unique firms (CIK-based) that a given IP address downloads on a given day. To motivate a cutoff threshold, we plot in the left-hand-side panel of Figure 1 the distribution of unique firm downloads by IP addresses on a daily basis. Each point on the curve represents the percentage of daily IP addresses downloading fewer than a particular number of firms' fundamental information. This figure shows that a great majority of daily IP addresses downloads a small number of unique firms' SEC filings. Specifically, 95% (99%) of all daily IP's download no more than 50 (500) unique firms' filings. Moreover, the right-hand-side panel of this figure shows that those IP addresses downloading fewer than 50 (500) in a day account for only 14.82% (33.1%) of the total EDGAR search traffic (visitor page views), suggesting that a small percentage of daily IP addresses downloading massive amounts of filings account for a great majority of EDGAR's search traffic, consistent with robot behavior as well as patterns observed on the Internet in general.[14] We choose a conservative cutoff at 50 firms – i.e., those daily IP addresses downloading more than 50 unique firms' filings are classified as robots and excluded from our analysis – resulting in

[13]This loss in sample can be explained by the fact that EDGAR hosts a variety of non-firm based filings.

[14]For example, http://techland.time.com/2013/12/13/robots-have-taken-over-the-internet/ reports that in 2013 60% of website visits are robot-generated.

a sample of 224.7 million visitor page view observations.[15]

In the fourth data processing step, we restrict our analysis to searches for filings containing fundamental information, namely, 10-K, 10-Q, 8-K, S-1, and 14A filings, leaving 136.5 million daily user page views.[16] These documents contain fundamental information about firm operating activities and collectively account for three-quarters of the underlying traffic flow.[17]

To infer investor perceptions of economic benchmarks, we define *Annual Search Fraction*,

$$f_{ij}^{t} = \frac{\Sigma_{d=1}^{365}(\text{Daily Searches for } j \text{ after } i)_d}{\Sigma_{d=1}^{365}(\text{Daily Searches for } i \text{ and then any firm } j \neq i)_d} \tag{1}$$

between base firm i and peer firm j in calendar year t, or the unique number of daily unique visitor searches for firm j's information after searching for firm i's information. Equation (1) is a version of a measure known as "confidence" (Agrawal et al., 1993) and has several desirable properties. First, it is a scaled measure that sums to one for each firm. Second, it allows the relationship between two firms i and j to be asymmetric in the sense that firm j may be an important peer of firm i but not vice versa.[18] Third,

[15]We allow for the same IP to be classified as robot or human on different days since IP address assignments change over time. It is, of course, possible that this simple rule excludes some intelligent traffic. In devising such a rule we attempt to maximize the signal to noise ratio while maintaining a reasonable sample size. In later robustness tests we show that the excluded robot traffic is indeed far less informative and robot-like.

[16]Forms 10-K and 10-Q are a company's annual and quarterly reports, respectively. Form 8-K is the "current report" that companies must file with the SEC to announce major events that shareholders should know about. Form S-1 is the "Registration Statement by the Securities Act of 1933" that a public company files to register their security with the SEC. Finally, Form DEF 14A, known as the "Definitive Proxy Statement," is filed by a company when a shareholder vote is required and it typically contains information on self-reported peer firms used for compensation purposes.

[17]The principal group of excluded traffic relates to Ownership Forms 3, 4, and 5. Companies are required to file these forms when there are material changes in the equity holdings of their corporate insiders. Our algorithm also includes filings belonging to the following categories in the SEC's form index: "10-K405", "10-KSB-A", "10-KSB", "10KSB","10KSB/A" "10-KT", "10-KSB40", "10-QSB", "10QSB" and "10QSB-A". However, searches for these forms occur very infrequently: of the traffic that we categorize in the baseline sample, approximately 1% of searches are for these forms. Moreover, given our focus on S&P1500 firms, the influence of these filings are trivial.

[18]In the networks literature, we are allowing for directed edges between firms. This may be sensible,

it is easy to interpret. For example, $f^{2011}_{GOOG,AAPL}$=.106 means that 10.6% of the daily IP addresses co-searching for Google's fundamental information and at least one other firm in calendar year 2011, subsequently searched for Apple's information. Note that this measure's construction implies that we are not using any information from users who only search for a singular firm's filings before leaving the EDGAR website.

In the final filtering step, we remove consecutive searches for firm i by a given user per day; to avoid double-counting users who repeatedly click back and forth between two companies' filings, we count a search click linking two firms only once per unique IP per day. This sixth and last data processing step results in a final dataset that consists of 10.4 million instances of search linkages between firm pairs by the same user within our firm and year sampling periods. To illustrate our key data processing steps and the computation of the *Annual Search Fraction* measure, Table A1 provides a hypothetical raw data sample and how the measure is computed for each ordered firm pair.

The *Annual Search Fraction* can be thought of as a distance metric between two firms, reflecting of the degree of fundamental connectedness or similarity between two firms in the collective minds of EDGAR users. The intuition follows a revealed preference logic: if individuals tend to search for the fundamental filings of two firms together, then these individuals must view these firms as economically related. Indeed, many web-based companies employ product recommendation systems following this logic. For example, Amazon.com publishes for each product listing the set of related products under "Customers Who Bought This Item Also Bought." Similarly, Google Knowledge Graph lists for a given object or person searched the set of most related objects or people under "People also searched for." Other sites like Netflix, Youtube, and SSRN employ similar methodologies. In each case, firms are generating recommendations based on

for example, when a firm like AOL is benchmarked to a substantially larger firm like Yahoo, but not vice versa. The idea that "industry" benchmarks could be firm-specific has been raised in recent research such as Hoberg and Phillips (2010, 2014), Rauh and Sufi (2012), Lewellen (2013).

users' aggregate co-search patterns.[19] Another interpretation of *Annual Search Fraction* is that it captures the degree of agreement across the population of EDGAR users on any base-peer firm pairing: e.g., 10.6% of users agree that Apple is an economic benchmark to Google.

3.2 Descriptive Statistics on Search-Based Peers

Using the cleaned final dataset, we identify a given firm's top 10 Search Traffic Peers (SBPs) by ranking its *Annual Search Fraction*s over prior calendar year. For example, a base firm's top 10 SBPs in 2009 are identified using 2008 EDGAR search traffic data. The top SBPs can be thought of as those firms most fundamentally connected to the base firm in the minds of investors.

In all of our analyses, we use the base sample of firms from the S&P1500 universe as of January 1, 2008. We also restrict SBPs to come from the same S&P 1500 universe, but show in the online appendix that our main results are robust to relaxing this restriction. Figure 2 provides four examples of the closest SBPs generated by our algorithm. Panel (a) of Figure 2 reports the top 10 SBPs for Walmart using search traffic from 2008. Perhaps unsurprisingly, Target is the top SBP for Walmart, even though GICS would place them in separate 2-digit groupings: while Walmart belongs to the Food, Staples, and Retailing industry (301010), Target is classified as Multiline Retail firm (255030).[20] This example

[19] In the recommendation systems literature, these techniques are often generalized under the term collaborative filtering (Resnick et al., 1994). As applied in the context of our paper, collaborative filtering would refer to algorithms that make recommendations of the form "people who have similar search histories as you have also searched for the following firms". Our approach can be thought of as a very simple form of this collaborative filtering technique that treats all users who have searched for a particular firm as being equally similar. That is, our algorithm makes "recommendations" of the following form: "people who have searched for this firm have also searched for these other firms."

[20] This example illustrates further that not only are SBPs not constrained by standard industry classification schemes, but that the *Annual Search Fraction* measure is useful in identifying similarity among firms. Note that while Exxon Mobil can be considered, in some sense, as a benchmark for Walmart, it could also be possible that larger or more salient firms could have a greater baseline likelihood of being searched by EDGAR users. If so, these firms may be more likely to appear as any base firm's top SBP using the current definition of *Annual Search Fraction*. In the Robustness section of our paper

illustrates that SBPs may not be constrained by standard industry classification.

Panel (b) of Figure 2 reports the top 10 SBPs, using 2008 search traffic, for Time Warner, a media conglomerate in a multitude of businesses such as film, cable, TV, radio, publishing, and Internet access. Time Warner's 2008 SBPs capture these different aspects of its business: Viacom, News Corp, and CBS represent the more traditional part of Time Warner's business while Google, Yahoo, and Microsoft represent the Internet-related part of its business. This example illustrating how one can, through co-search patterns, pick peers from across different industries, a feat which Cohen and Lou (2012) argue can be especially difficult for investors in analyzing conglomerate firms. Finally, Panels (c) and (d) of Figure 2 display the top 10 SBP firms for Google using search traffic from 2008 and 2011, respectively, as well as their *Annual Search Fraction* over the same calendar year. Interestingly, while in 2008 Apple was deemed the third closest peer, with an *Annual Search Fraction* of under 0.05, by 2011 it had surpassed Microsoft and Yahoo as Google's closest peer and nearly doubled its *Annual Search Fraction*. This example reflects the flexibility in SBPs in that they can adapt to changing economic conditions or competitive environments, unlike the more rigid structures offered by traditional industry classification schemes.

These examples also reveal that SBPs' *Annual Search Fraction*s are highly concentrated. In other words, the firms that score highest in terms of their EDGAR traffic ranking own a disproportionately large share of the adjacent searches to any given base firm. This led us to wonder if the *Annual Search Fraction* between base firms and their peers may follow a power law. Indeed, the *Annual Search Fraction* as a function of the ordinal peer ranking, plotted in Panel (a) of Figure 3, appears to follow a power-law-like trend. Moreover, in Panel (b) of Figure 3, we find a strong linear fit (99.9% R^2) in the the relationship between log *Annual Search Fraction* and log peer rank, consistent with

we consider an alternative definition of *Annual Search Fraction* that addresses the possibility that the baseline likelihood of being searched may differ for some firms.

a power law.

If we view the search fraction measure as the degree of collective agreement across EDGAR users, the power law distribution suggests that the agreement about a firm being an economically related peer tends to decay at an exponential rate. A second and stronger interpretation of this result is that a base firm has a relatively small set of most economically relevant benchmarks, suggesting that the traditional industry groupings might be too broad or too rigid to identify the best peer firms for benchmarking purposes.[21]

Panel A of Table 2 summarizes the *Annual Search Fraction* of the top 10 SBPs to our sample of base firms and the degree of correspondence in the standard industry classifications of the SBPs and their base firms. This panel shows a high degree of concordance between standard industry classifications of base firms and the classifications of their top 10 SBPs, with the level of concordance highest among the top SBPs: 83% (70%) of the firms classified as the top SBP belong to the same GICS2 (GICS6) industry as the base firm, while only 59% (37%) of firms classified as the 10th closest peer by EDGAR traffic data belong to the same GICS2 (GICS6) industry.[22]

Panel B of Table 2 further clarifies how our peer firms differ from GICS firms. This panel groups base firms according to their GICS2 designation. For base firms in the financial sector we find the highest level of correspondence: 83% of the top 10 SBPs also belong to the same GICS2 sector. However, we find a substantially lower level of correspondence for most other sectors. In the industrials and materials sectors, for example, we find only a 54% and 51% concordance, respectively, at the GICS2 sector level and only 34% and 44% at the GICS6 industry level, respectively.

These summary statistics have two important implications. First, they strongly sup-

[21]We note that identifying the ideal number of benchmark firms in performance attribution exercises is an important topic, and analyzed in the recent work by Lewellen and Metrick (2010), but is outside the scope of the present paper.

[22]Note that all classification codes are as of 2011 because of limited data. Bhojraj et al. (2003) show that the GICS rarely changes for a given firm and the amount of bias it introduces should therefore be minimal.

port our hypotheses that EDGAR search patterns are non-random and could be informative of firm similarities. Second, any outperformance of our SBPs over benchmarks from standard classifications reflects the novel information captured in the search patterns of EDGAR users, who are together tapping into additional sources of information not contained in standard industry groupings.

4 Empirical Analysis and Results

In this Section we investigate how well top 10 SBPs explain the cross-sectional variation in base firms' monthly stock returns as well as financial ratio, valuation multiples, and other fundamental information, and compare the performance of SBPs to that of GICS6. We focus on GICS6 as the main alternative peer identification scheme because it has been shown to outperform other standard classification schemes along multiple dimensions (e.g., Bhojraj et al., 2003; Lewellen and Metrick, 2010), and because it is widely used among practitioners and increasingly so among academics.

4.1 Monthly Returns

Our first test compares GICS6 and SBPs in their abilities to explain the cross-sectional variation in base firms' monthly stock returns. We estimate the following cross-sectional regression, for every month from 2009 to 2011:

$$R_{i,t} = \alpha_t + \beta_t R_{p_i,t} + \epsilon_{i,t}, \tag{2}$$

where $R_{i,t}$ is the CRSP monthly cum-dividend return for each base firm (i), taken from the CRSP monthly files, and $R_{p_i,t}$ is the average monthly returns for a portfolio of benchmark firms specific to base firm i. To assess performance, we compare the average R^2 produced by monthly regressions using benchmark portfolios of 10 firms randomly selected from

the base firms' GICS6 industries versus the average R^2 produced by portfolios of the base firms' top 10 SBPs. To avoid look-ahead bias, our SBPs are always identified using search traffic from the prior calendar year. For example, the *Annual Search Fraction* measure f_{ij}^t used to identify SBPs in calendar year 2009 are computed using calendar year 2008 data. Thus, we estimate cross-sectional regressions of Equation (2) for every month from 2009 to 2011 and obtain an average R^2 based on the 36 regressions.

The intuition for this test is that peer firms that are more economically related to their base firms should exhibit greater contemporaneous correlation with them in returns. To see this intuition more clearly, it is useful to consider the following decomposition of returns for firm i in period t into its idiosyncratic, $\epsilon_{i,t}$, and non-idiosyncratic, $\delta_t(\theta_i)$, components:

$$R_{i,t} = \delta_t(\theta_i) + \epsilon_{i,t}, \text{ and} \tag{3}$$

$$R_{p_{i(N)}} = \frac{1}{N}\sum_j R_j = \frac{1}{N}\sum_j \delta_t(\theta_j) + \frac{1}{N}\sum_j \epsilon_{jt}, \tag{4}$$

where $p_{i(N)}$ denotes base firm i's benchmark portfolio consisting of N peer firms. The non-idiosyncratic component represents a firm's response to common (market or industry level) shocks over the period, which is assumed to depend on some economic or fundamental characteristics of the firm indexed by θ. For example, in a market factor model of returns, $\theta_i = \beta_i$ and $\delta_t(\theta_i) = \beta_i(r_m - r_f)$.

In the univariate regression tests of Equation (2), the R^2 measure is equivalent to the squared correlation coefficient, and differences in R^2 between candidate benchmark portfolios are driven by differences in

$$\frac{\hat{Cov}(R_{i,t}, R_{p_i})}{\hat{Std}(R_{p_i})} \approx \hat{Cov}\left(\delta_t(\theta_i), \frac{1}{N}\sum_j \delta_t(\theta_j)\right) \Big/ \hat{Std}\left(\frac{1}{N}\sum_j (\delta_t(\theta_j) + \epsilon_{j,t})\right). \tag{5}$$

The numerator in Equation (5) highlights the intuition that benchmark portfolios will

produce higher R^2 if, all else equal, their responses to common shocks covary with base firms' responses more strongly. In other words, higher R^2 will reflect greater economic similarity between base firm and peer firms (i.e., as reflected in their "type" θ). Our *Annual Search Fraction* measure can be interpreted as capturing this degree of economic relatedness.

The denominator in Equation (5) also suggests that, all else equal, R^2 is higher when the benchmark portfolio minimizes the influence of idiosyncratic shocks (lowering the denominator), a function of the portfolio size. Assuming that firms are unique and live in a continuum of θ, increasing the portfolio size trades off the denominator effect with the numerator effect, as less economically relevant firms are added to the portfolio.[23]

Since choosing the optimal number of peers is beyond the scope of this paper, in our baseline tests we standardize each base firm's benchmark portfolio to consist of 10 firms to mitigate the effects of differential portfolio sizes. In particular, we compare the performance of base firms' top 10 SBPs to an equal-weighted portfolio of 10 random firms from the base firm's GICS6 industry — but exclude the base firm.[24] As a result, this process precludes us from including those base firms with fewer than 10 GICS6 peer firms. However, in later robustness tests we show that the relative performance of SBPs is robust to the number of firms included in the benchmark portfolio.

We consider two types of peer portfolios using SBPs. The first type of peer portfolio, denoted "SBP EW," takes the closest 10 peer firms as implied by our *Annual Search Fraction* measure and forms an equally weighted portfolio. The second type of peer portfolio, denoted "SBP TW" (traffic-weighted), takes the closest 10 firms as implied through our *Annual Search Fraction* measure but forms a weighted average portfolio, where a firm's portfolio weight is the *Annual Search Fraction* measures rescaled to sum to one. To fa-

[23] See Lewellen and Metrick (2010) for a detailed discussion of optimal benchmark portfolio size.

[24] To reduce estimation noise in our random sampling of GICS6 firms, we simulate 500 draws of random firms and all results are based on the average of these draws.

cilitate comparisons, all the regressions are conducted using the same underlying set of base firms, so that our analyses include only base firms with at least 10 GICS6 peers and 10 SBPs.

Table 3 reports the average R^2 values from monthly regressions of Equation (2), using base firms from the S&P1500 and the S&P500. We restrict the benchmark portfolio firms to come from the S&P1500, a requirement that implicitly induces some size-matching with the base firms.

Our results reveal that SBP portfolios significantly outperform GICS6 peer portfolios. For the group of S&P1500 base firms, their GICS6 peer portfolios explain, on average, 6.9% of the cross-sectional variation in monthly returns, which is significantly lower compared to the 11.4% explained by their SBP EW portfolios and the 12.5% explained by their SBP TW portfolios. Similarly, for the set of S&P500 base firms, their GICS6 peer portfolios explain, on average, 11.6% of the cross-sectional variation in monthly returns, which is again significantly lower than the 18.9% explained by their SBP EW portfolios and the 21.4% explained by their SBP TW portfolios.[25]

In results tabulated in our online appendix, we perform the baseline return tests using different types of benchmark portfolios, depending on whether one restricts the portfolio firms to come from the S&P1500 or the CRSP universe, and whether one restricts the benchmark portfolio firms to come from the same GICS2 sector as the base firm. We find that SBP portfolios consistently outperform GICS6 in these alternative tests. Overall,

[25]We note that the higher R^2 could, in theory, result from either positive or negative co-movements in the cross section — i.e., positive or negative numerator in Equation (5) — since benchmark portfolio firms' prices can co-move with the base firm either positively (due to common responses to shocks) or negatively (due to opposing responses to shocks, such as substitution effects between competitors). If the cross section of peer firms co-move with base firms positively, we would observe *positive* regression coefficients and relatively high R^2; if the cross section of peer firms co-move with base firms negatively, we would observe *negative* regression coefficients and relatively high R^2. However, if the two effects occur with similar frequency across the entire cross section of base firms, we should find flat slope coefficients and relatively low R^2. We find (and report in the online appendix) that the average slope coefficients for our Table 3 regressions are strongly positive, and suggest that the cross-sectional relation between base and peer firm returns is driven by common responses to macro shocks shocks rather than substitution effects.

SBP EW portfolios explain between 63% to 298% (or 170% on average) more of the variation in base firms' monthly returns, while SBP TW portfolios explain between 81% to 343% (or 194% on average) more of the variation in base firms' monthly returns. Interestingly, in variations of the test we find that the SBP TW portfolios significantly outperform the SBP EW portfolios, explaining between 4% to 15% (or 10% on average) more of the variation in base firms' monthly returns. This evidence strongly suggests that our *Annual Search Fraction* measure captures the fundamental similarities between the base and peer firms.

We note that our baseline specification in Table 3 — focusing on S&P1500 base firms and benchmark portfolio firms — is meant to be conservative. Compared to the alternative specifications, the baseline compares least favorably relative to the GICS6. For example, when we allow the benchmark portfolios to consist of any firm in CRSP (i.e., remove the S&P1500 requirement), we effectively remove the size-matching between base and peer firms implicit in the prior set of tests. While this specification tends to significantly deteriorate the performance of GICS6 peer portfolios, it actually weakly increases the performance of SBPs. Another variation of our tests restricts the benchmark portfolios to consist of firms in the same GICS2 sector as the base firm. Given the high level of concordance between base firms' and their SBPs' GICS2 sectors, we motivate this restricted model as a way to test for the incremental insights EDGAR users bring to bear when selecting benchmarks from within a particular GICS2 group. While this restriction does not impact the performance of GICS6 peer portfolios, we find a small improvement in our SBPs, thus improving SBPs' relative performance. For tractability, we focus all our following analyses and robustness tests using the baseline specification of Table 3, focusing on the base and peer firms from the S&P1500, where the performance of SBP portfolios is the weakest.

Figure 4 Panels (a) and (b) depict the time series of the monthly R^2 values from

monthly return regressions of Table 3 rows 1 and 2, respectively. Both SBP EW and TW portfolios outperform GICS6 peer portfolios in all but one month. Moreover, SBP TW portfolios outperform SBP EW portfolios for a great majority of months. These figures suggest that the results of Table 3 — i.e., the superior performance of SBPs and the informativeness of *Annual Search Fraction*s — are systematic and not driven by subperiods, and support the hypothesis that *Annual Search Fraction*s are non-random and contain information about fundamental connectedness between firms not readily captured in standard classification schemes.

4.2 Valuation Multiples, Financial Ratios, and Other Characteristics

Another way to evaluate SBPs is to assess the extent to which they explain the cross section of base firms' valuation multiples, financial ratios, and key accounting measures. The intuition of these additional tests is similar to the above – benchmark portfolios that are more economically related to their base firms should exhibit greater contemporaneous correlation with them in these fundamental measures.

To perform these additional tests, we gather quarterly data from Compustat and Institutional Brokers' Estimate System (I/B/E/S) on a range of valuation multiples, financial ratios, and other fundamental characteristics, including the price-to-book multiples (pb), enterprise value-to-sales multiples (evs), price-to-earnings multiples (pe), returns on net operating assets ($rnoa$), returns on equity (roe), asset turnover (at), profit margins (pm), leverage (lev), long-term analyst growth forecasts ($ltgrowth$), one-year-ahead realized sales growth ($salesgrowth$), and research and development expenses scaled by net sales ($rdpersales$). The exact computation of these variables (as well as all others used in this paper) are detailed in Table A2.

With each of these variables, we run the analogous cross-sectional regression,

$$Variable_{i,t} = a_t + \beta_t Variable_{p_i,t} + \epsilon_{i,t}, \tag{6}$$

where $Variable_{i,t}$ is the variable of interest for each base firm (i) and the regressor $Variable_{p,t}$ is the mean value of 10 firms, based on either the same GICS6 group or one of our two traffic-based measures (SBP EW and SBP TW). We estimate these regressions on a quarterly basis, at the end of March, June, September, and December of each calendar year from 2009 to 2011. The relevant variables are computed using financials that are available from Compustat at the end of each quarter. Similarly, we obtain the most up-to-date median long-term analyst forecasts from the Institutional Brokers' Estimate System at the end of each calendar quarter.

For the entire firm quarter–year sample, we drop observations that are missing data on total assets, long-term debt, net income before extraordinary items, debt in current liabilities, or operating income after depreciation. We also drop observations with negative common or total equity and keep only observations with net sales exceeding $100 million and a share price greater than $3 at the end of the fiscal quarter. Finally, to mitigate the influence of outliers, we truncate observations at the first and 99th percentiles for each of the variables for each regression equation.[26] In addition, we require net income before extraordinary items to be positive and require non-missing values for current liabilities, current assets, and property, plants, and equipment in computing *rnoa*. To facilitate comparisons, all the regressions are conducted using the same underlying set of base firms, so that our analyses include only base firms with at least 10 GICS6 peers and 10 SBPs.

Table 4 compares GICS6 and SBP portfolios and shows that SBP portfolios outperform the GICS6 peer portfolios for most of the variables tested. Within the S&P1500

[26]This is done on an equation by equation basis to avoid losing observations unnecessarily.

base firm sample (similar results are reported for the S&P500 sample in the Internet Appendix), the SBP EW portfolios explain a significantly greater proportion of the cross-sectional variations than the GICS6 peer portfolios for a great majority of variables. These results are significant at the 5% level for all valuation multiples except *pe*, all financial statement ratios, and all other financial information except for *rdpersales*. When we turn to the SBP TW portfolios, the outperformance over GICS6 achieve 1% significance for all variables except *pe*, which is significant at the 10% level.

In unreported results, we check that restricting the SBPs within GICS2 produces nearly identical results as those of Table 4, though the unrestricted portfolios results are generally stronger both in terms of magnitudes as well as statistical significance. Together, these findings show that EDGAR users in the aggregate are able to identify peer firms that substantially outperform the GICS6 in terms of their ability to explain cross-sectional variations in key valuation multiples, financial statement ratios, and other fundamental characteristics.

4.3 Multivariate Analysis of Traffic Peer Choice

We complement the above analyses by examining the fundamental characteristics of top SBPs in a multivariate setting. In particular, we examine whether firms that are more similar to the base firm in fundamental characteristics are more likely to be a top SBP. To do so, we match each base firm in the S&P1500 sample to its top 10 SBPs and all GICS2 peers not included in the top 10 SBPs. Then we compute, for each base to peer firm pair a difference metric for each of the following fourteen characteristics: market capitalization (*size*), *pb*, *pe*, *rnoa*, *roe*, *at*, *evs*, *lev*, *salesgrowth*, *rdpersales*, number of analysts covering the firm (*coverage*), standard deviation in analysts' EPS forecasts (*eps spread*), *ltgrowth*, and standard deviation in analysts' long-term growth forecasts (*ltgrowth spread*).

To reduce the effect of outliers, we discretize the difference metrics into three values. The difference metric takes on the value of 1 if the absolute percentage difference in the firm characteristic between the base firm and the peer firm is less than 25% (i.e., $|\frac{Var_{peer}}{Var_{base}} - 1| \leq 25\%$ for some variable "Var"); the metric takes on the value of 2 if the absolute percentage difference in the firm characteristic between the base firm and the peer firm is between 25% and 50%; finally, the metric takes on the value of 3 if the absolute percentage difference in the firm characteristic between the base firm and the peer firm is greater than 50%.

We estimate using pooled probit models the likelihood that a potential peer firm is a top 10 SBP as a function of differences in fundamental characteristics. Table 5 reports our estimation results; the odd columns report probit coefficient estimates and the even columns report marginal effects, which are evaluated at 1.[27] All else equal, our results suggest that, relative to other potential peer firms in the same GICS2 sector, a peer firm is more likely to be a base firm's top 10 SBP if it is more similar to the base firm in fundamental characteristics. With the exception of *pb* (no significance) and *roe* (10% significance), all explanatory variables are negative and significant at the 1% level in all specifications.

Similarities to the base firm in *pe*, *at*, *rdpersales*, and GICS4 industry grouping have the most economically significant association with the likelihood of being a top 10 SBP. Interpreting column 4 of Table 5, a specification which includes both Compustat-based financial variables as well as analysts-based variables from IBES, an increase from 1 to 2 in the difference metric in *pe* (*at*) [*rdpersales*] is associated with a 5.35% (3.83%) [4.09%] decline in the likelihood of being a top 10 SBP, while being in a different GICS4 industry grouping than the base firm is associated with a 10.57% decline in the likelihood of being a top 10 SBP.

[27] The variable "Different GICS4," an indicator that the peer and base firms belong to different 4-digit GICS industry groups, and year fixed effects are evaluated at 0.

In sum, the multivariate evidence suggests that absolute differences in each of these fundamental characteristics are significantly associated with the likelihood of being a top SBP. The overall evidence from this analysis is consistent with EDGAR investors, in the aggregate, searching for fundamentally similar firms and help to explain the superior performance of SBPs over GICS in prior tests.

5 Robustness

In this Section we examine the robustness of our primary results in Table 3. In particular, we consider alternative peer portfolio constructions, alterations to our "co-search" algorithm, and alternations to our sample.

5.1 Number of Firms in Peer Portfolio

Our results above are produced on the basis of comparing the 10 closest traffic-based peers to a randomly selected group of 10 peers from the GICS6 groupings. In this subsection we conduct robustness tests to examine the effect of the choice of peer size on the relative performance of peer groups in explaining the cross-sectional variation in the returns of base firms by varying the size of the peer portfolio from 1 to 15.[28] For a peer group of size N, we re-estimate rows 1 and 2 of Table 3 by matching to each base firm an equal-weighted portfolio of N firms (excluding the base firm) randomly selected from its GICS6 industry. As before, this process necessitates that we exclude base firms with fewer than N GICS6 peer firms. We compare the performance of these GICS6 peers to the performance of equally-weighted and traffic-weighted portfolios of each base firm's top N SBPs.

Panel (a) of Figure 5 compare, across the three types of peer portfolios, the 36-month

[28] Dikolli et al. (2012) provide evidence that the number of peers used can affect hypothesis testing for relative performance evaluation analysis.

average R^2 values from the estimation of Equation (2) as a function of peer portfolio size for the S&P1500 base firm sample. The graphs show that the magnitude of the difference in R^2 generated by equal-weighted SBP portfolios from those generated by GICS6 peer portfolios is declining as the portfolio size increases. This is intuitive since as the peer portfolio increases in size, there is a greater overlap between the GICS6 portfolio and the SBP portfolio. In contrast, the magnitude of the difference in R^2 values generated by traffic-weighted SBP portfolios persists even as the size of the peer group increases, again suggesting that traffic weights are a good proxy for fundamental similarity between firms. These dynamics are highlighted in Panel (b) of Figure 5, which plot the differences in R^2 values between the different types of peer portfolios.

Overall, our robustness test finds that the excess ability of SBP EW portfolios in explaining base firms' returns variations (i.e., in excess of GICS6 peer portfolios) diminishes with the size of peer groups. However, the excess ability of SBP TW portfolios in explaining base firms' returns variations remains stable even as the size of peer groups increases. This divergence between equal- and traffic-weighted peers is expected since, by construction, the marginal N^{th} peer as ranked by the *Annual Search Fraction* is less informative than the set of peers preceding it. In untabulated tests, we find that our baseline specification of SBP EW and TW portfolios consisting of 10 firms in fact outperform the portfolios of all GICS6 peers (i.e., removing the 10 firm constraint) in explaining the cross section of base firm returns by significant margins – 24% and 36%, respectively. Consistent with the above, this finding suggests out our SBPs represent a more powerful way of accounting for the cross-sectional variation in base firm returns.

5.2 Base Firm's Industry Classification and Size

We also examine whether our main findings on returns are driven by certain subpopulations of base firms. In particular, we test whether our baseline results on monthly

returns hold for base firms in different 2-digit GICS industry sectors and for firms in different in-sample size deciles.

To determine how our baseline results vary by base firms' industry grouping, we classify each base firm in the S&P1500 universe by its 2-digit GICS code as of 2011 and rerun the same specification in Equation (2). Our results, reported in Panel A of Table 6, show that in all 2-digit GICS base firm groupings the SBP EW portfolios significantly outperformed GICS6 peers; moreover, in the majority of industries weighting by *Annual Search Fraction* significantly increased the explanatory power of SBPs.[29]

To determine how our baseline results vary by size of the base firm, we again start with the universe of base firms in the S&P1500 and classify them according to within-sample market cap deciles for each calendar year based on their end of December market cap in the previous calendar year. Our results, reported in Panel B of Table 6, show that SBPs again significantly outperform GICS6 across the size distribution, with nearly double the R^2 at both the bottom and top decile. In the majority of the size deciles, as in Panel A, weighting by *Annual Search Fraction* significantly increased the explanatory power of SBPs.

5.3 Comparison Against Sized-Adjusted GICS6 Peers

We also compare the performance of SBPs against an alternative, perhaps more refined, set of benchmark peers. In lieu of comparing against a random set of GICS6 peers, we refine the benchmark peer group of firms to be those size-matched firms in the same industry, a peer identification scheme that has been used in the prior literature (e.g., Albuquerque, 2009). We segment base firms into within-sample market capitalization quartiles using the end of the prior calendar year's market capitalization as breakpoints. To find a benchmark peer firm in our tests, we draw random firms who share both the

[29]The telecommunications group is excluded since it has only 12 base firms.

same 6-digit GICS groupings and the same market capitalization quartile. Because this restriction reduces the potential number of peers available (and hence the number of base firms in our exercise), we re-run our baseline returns tests with only five peers per base firm. Our results, reported in row 2 of Table 7, find that even compared to these more refined size-adjusted GICS6 benchmark group of peers, SBPs continue to outperform in terms of explaining the cross-sectional variation in base firms' returns. The R^2 differences between traffic-weighted and equal-weighted SBP portfolios remain significant and large.

While the above size-matched GICS6 benchmark seems appropriate given the empirical literature, we also consider an alternative matching process in row 3 of Table 7. Instead of randomly selecting 5 GICS6 peer firms belonging to the same size quartiles as the base firm, we select the 5 GICS6 peers whose market capitalizations (again measured at the end of the previous calendar year) are the closest to the market capitalization of the base firm. This matching could potentially be more accurate by identifying those industry peers that are closest in size rather than randomly selecting those industry peers who belong to the same size quartile.[30] Row 3 of Table 7 shows that, using this algorithm, size-matched GICS6 peers produce average R^2s that are slightly larger than those produced using the algorithm in row 2 of the same table.[31] SBP portfolios, however, continue to outperform these size-matched GICS6 peer portfolios.

5.4 Alterations to Co-Search Algorithm

We consider the following three variations of our co-search algorithm.

[30] Of course, in cases where the closest firms in terms of size belong to different size quartiles, this algorithm may not perform well.

[31] Note that since the set of base firms differ across the two specifications, the R^2s across are not necessarily comparable.

5.4.1 Session Level

The first variation alters the time window over which we consider co-searches to be initiated by the same user from the daily level to the intraday session level, where a session is defined as a contiguous block of search activity within a particular day.[32] The intuition is that this allows us to sharpen the sequence of related searches by examining the sequence of contiguous searches, which may be more consistent with the users' workflows (e.g., a morning session vs an afternoon session). Row 4 of Table 7 report results of the exercise and finds virtually identical results compared to Table 3 row 1.

5.4.2 Upstream and Downstream Traffic

In lieu of defining peer firms based on downstream traffic, where the first and subsequent search defines the base and peer firm respectively, we examine a variation of the co-search algorithm in which identification of fundamentally connected firms includes both upstream and downstream traffic. This alternative algorithm relaxes the assumption that the first search is for the base firm, and allows for peer firm relations to be symmetric (i.e., if firm A is a peer of firm B, then firm B is a peer firm of firm A). Under this identification scheme, in a sequence of searches the *Annual Search Fraction* is defined as the percentage of unique users searching for firm j chronologically right after or right before searching for firm i on the same day. We re-run our returns tests using this more expansive definition of fundamentally related peers and report the results in row 5 of Table 7; our findings are similar to the baseline results, though the R^2 for both SBP EW and SBP TW portfolios are slightly larger than those in Table 3 row 1.

[32]Formally, a search at time t from user i is part of a new user session if user i has been inactive from t to $t-j$, where j is 60 minutes, otherwise the search is part of the current user session.

5.4.3 Baseline Probability Correction

The last variation follows from the observation that some firms may have a higher baseline likelihood of being searched, and therefore are more likely to appear in any search sequence. These firms are more likely to appear as a peer firm under our algorithm even if users' search sequence orderings are random and do not reflect their views about the connectedness between firms.[33] This phenomenon could explain why ExxonMobil appears as a top SBP to Walmart in Figure 2. A correction for the phenomenon can be implemented through redefining the *Annual Search Fraction* measure as follows:

$$\hat{f_{ij}} = \frac{f_{ij} \equiv \frac{i \text{ then } j}{i}}{\frac{!i \text{ then } j}{!i}} \tag{7}$$

where i is the base firm, and j is the peer firm; f_{ij} is our original *Annual Search Fraction* measure, $\hat{f_{ij}}$ is the corrected *Annual Search Fraction*, and ! is the negation operator. The denominator can be interpreted as measuring the average likelihood of firm j appearing as a peer to any base firm $\neq i$ in the sample.[34] In Row 6 of Table 7 we perform the correction and results are again similar to our baseline results.

5.5 Alterations to Sample Selection

We also consider the following partitions to our search sample to investigate whether particular types of searches are potentially more or less informative.

[33] In order to provide a framework understanding of this potential bias, first assume, without loss of generality, one representative agent who searches sequentially. At each time period, the agent makes an independent draw from the set of K firms with probabilities p_k such that $\Sigma p_k = 1$. The measurement of *Annual Search Fraction* of peer firm j of base firm i is simply p_j. In other words, large (or popular) firms are more likely to be peers of base firms simply because they show up more frequently at the baseline.

[34] Equation (7) is known as "lift" (Brin et al., 1997), and seeks to measure the incremental co-occurrence between firms i and j using the degree of co-occurrence under the assumption of statistical independence between i and j as the benchmark.

5.5.1 "Robot" Searches

We initially filtered the sample to reflect more human and implicitly "informed" searches using the argument that web crawlers are less likely to be intelligent and that a large number of unique firm downloads by a given IP is a good proxy for web crawlers. To provide evidence on this issue, we re-run our baseline returns test using the sample of searches we classified as robot-generated: i.e., generated by those daily IP addresses that searched for more than 50 unique firms' filings. Row 2 of Table 8 reports the results of this exercise and finds very little residual intelligence in these searches. Both SBP EW and SBP TW portfolios explain significantly less proportion of the cross-sectional variation in base firms' returns when compared to GICS6 peer portfolios (e.g., 1% for SBP EW compared to 7% for GICS6). Moreover, SBP TW portfolios perform worse than SBP EW, suggesting that the search sequences generated by these IP addresses are not informative. Together, these results are consistent with this subsample capturing robot-like searches.

5.5.2 HTML vs. TXT Searches

We also conjecture that there may be a distinction between HTML and TXT searches. On the SEC EDGAR website, each filing is generally viewable (downloadable) through one of two formats: 1) HTML or 2) the complete submission file which is TXT.[35] We hypothesize that TXT files are more likely to be associated with bulk (but potentially sophisticated) downloaders since they are meant to be parsed to be useful and not as viewable as HTML rendered pages.

Within our baseline sample, we partition based on whether the search was for HTML (90% of sample) or TXT (10%) and re-run our baseline returns tests. In rows 3 and 4 of

[35]Complete submission files show all HTML tags and all embedded graphics in binary form and are typically not human friendly to read. They are, however, the download format of choice for users who would like to parse the underlying data (presumably robots).

Table 8 we find that most of the explanatory power of SBPs comes from HTML-based searches, consistent with the notion that TXT downloads are likely associated with bulk unsophisticated downloading.

5.5.3 Filing Age

Finally, we explore whether the vintage of the information being searched is informative. We again partition our baseline sample, this time based on the filing age of the base firm using a cutoff of 6 months. Table A3 reports that 63% of the baseline sample comprise of searches for base firm's information that is under 6 months old (the difference between the click date and the filing date). Moreover, Figure 6 shows that in the vast majority of cases the ages of a base firm's filings downloaded are within 6 months of the ages of the peer firms' filings downloaded. These summary statistics suggest that investors are mostly looking for recent fundamental information and that their search sequences are consistent with benchmarking behavior.

In rows 5 and 6 of Table 8, we report results of the exercise and find both types of searches to be informative, although searches for newer information appears slightly more so. In the final row of Table 8, instead of partitioning, we use the entire baseline sample, but weigh each observation according to the base firm's filing age. The weighting scheme is such that filings with zero age (click on same day as filing) have a maximum weight of 1, while filings with age greater than 2 years (11% of sample) have a weight of 0 and filings in between are weighted linearly. This weighting scheme produces slightly stronger results compared to our baseline, consistent with the searches for newer information being more informative.

6 Conclusion

Academic researchers and practitioners such as financial analysts, auditors, corporate managers, regulators, and policy makers have long used industry classification schemes for benchmarking purposes. With the rise of a service and knowledge-based economy and a constantly changing competitive landscape, the exercise of benchmarking to standard industries — rigid groupings based on similar production processes or outputs — has become more dubious.

We propose a novel method for identifying economic benchmarks, by relying on the premise that EDGAR users are collectively searching for firm fundamentals to aid their investment decisions and that this process involves a comparison of related firms. Relative to GICS6 peer firms, firms that are frequently co-searched by multiple users (SBPs) not only explain a substantially greater proportion of the cross-sectional variations in base firms' (out-of-sample) monthly returns, but also variations in base firms' valuation multiples, financial ratios, and other fundamental characteristics. These results hold for both the S&P1500 firms and for the S&P500 firms across a variety of subsamples and specification checks. Our findings suggest that SBPs can be a useful input to performance attribution exercises.

While these results are promising, we note that this method of peer identification is limited by data availability in practical applications. We believe these limitations will be addressed in part through the SEC's plans to release this data more broadly. Nevertheless, as new databases on web traffic become available through search logs (including, for example, those from major search engines such as Bing, Google, and Yahoo), we are hopeful that the techniques for analyzing search traffic we develop here help guide future research in the area.

Looking ahead, the findings in this study highlight co-searching as a promising venue for future research. At a minimum, our results suggest that SBPs represent a powerful

alternative approach to account for industry effects in a multitude of firm performance measures and fundamentals. Moreover, we believe that opportunities exist in establishing wider applications of SBPs in a variety of decision contexts: for example, for portfolios construction purposes or for trading strategies (e.g., pairs trading). We look forward to these new research opportunities.

References

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, SIGMOD '93, pp. 207–216.

Albuquerque, A., 2009. Peer firms in relative performance evaluation. Journal of Accounting and Economics 48, 69–89.

Bhojraj, S., Lee, C. M. C., 2002. Who is my peer? a valuation-based approach to the selection of comparable firms. Journal of Accounting Research 40, 407–439.

Bhojraj, S., Lee, C. M. C., Oler, D. K., 2003. What's my line? a comparison of industry classification schemes for capital market research. Journal of Accounting Research 41, 745–774.

Brin, S., Motwani, R., Ullman, J. D., Tsur, S., 1997. Dynamic itemset counting and implication rules for market basket data. In: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, SIGMOD '97, pp. 255–264.

Chan, L. K. C., Lakonishok, J., Swaminathan, B., 2007. Industry classifications and return comovement. Financial Analysts Journal 63, 56–70.

Cohen, L., Lou, D., 2012. Complicated firms. Journal of Financial Economics 104, 383–400.

De Franco, G., Hope, O. K., Larocque, S., 2011. Analysts' choice of peer companies. SSRN Electronic Journal .

Dikolli, S., Hofmann, C., Pfeiffer, T., 2012. Relative performance evaluation and peer-performance summarization errors. Review of Accounting Studies pp. 1–32.

Drake, M. S., Roulstone, D. T., Thornock, J. R., 2012. What investors want: evidence from investors' use of the EDGAR database. SSRN eLibrary .

Fama, E. F., French, K. R., 1997. Industry costs of equity. Journal of Financial Economics 43, 153–193.

Fan, J. P. H., Lang, L. H. P., 2000. The measurement of relatedness: an application to corporate diversification. The Journal of Business 73, 629–660.

Gabaix, X., Ibragimov, R., 2011. Rank- 1/2: a simple way to improve the OLS estimation of tail exponents. Journal of Business & Economic Statistics 29.

Guenther, D. A., Rosman, A. J., 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. Journal of Accounting and Economics 18, 115–128.

Hoberg, G., Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. Review of Financial Studies 23, 3773–3811.

Hoberg, G., Phillips, G. M., 2014. Text-based network industries and endogenous product differentiation. Working Paper 15991, National Bureau of Economic Research.

Kahle, K. M., Walkling, R. A., 1996. The impact of industry classifications on financial research. Journal of Financial and Quantitative Analysis 31, 309–335.

King, B. F., 1966. Market and industry factors in stock price behavior. The Journal of Business 39, 139–190.

Lewellen, S., 2013. Executive compensation and peer effects. Tech. rep., Working Paper, Yale University.

Lewellen, S., Metrick, A., 2010. Corporate governance and equity prices: Are results robust to industry adjustments. Tech. rep., Working paper, Yale School of Management.

Ramnath, S., 2002. Investor and analyst reactions to earnings announcements of related firms: An empirical analysis. Journal of Accounting Research 40, 1351–1376.

Rauh, J. D., Sufi, A., 2012. Explaining corporate capital structure: Product markets, leases, and asset similarity. Review of Finance 16, 115–155.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. GroupLens: an open architecture for collaborative filtering of netnews. In: *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, ACM, New York, NY, USA, CSCW '94, pp. 175–186.

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

Table A1.
Traffic Example

This table illustrates our algorithm's procedure in generating our *Annual Search Fraction* measure on hypothetical data. The *Annual Search Fraction* is defined as the percentage of unique users searching for firm j after searching for firm i on the same day over a period of one calendar year. To apply the data processing rules in Table 1, we first exclude observation 1 because user 1.1.1.1 only searched 1 unique firm. We also exclude observation 2 because we do not use Form 4 traffic; consequently we only have one unique firm search for user 1.1.1.2, so we also drop observation 3. $f_{ab} = \frac{1}{2}$, $f_{ac} = \frac{1}{2}$, $f_{ba} = 1$ and $f_{ca} = 1$ are computed using observations $4 \sim 9$, highlighting the asymmetry of our approach. Note that we exclude traffic from observations 8 and 9 in computing f_{ab}, because we already counted the linked traffic between firms A and B by the same user 1.1.1.3 on the same day in observations 4 and 5.

Observation Number	IP Address	Timestamp	Form Type	Firm ID
1.	1.1.1.1	1-1-2008 6:32:51 AM	10-K	B
2.	1.1.1.2	1-1-2008 6:32:51 AM	Form 4	A
3.	1.1.1.2	1-1-2008 6:32:51 AM	10-K	B
4.	1.1.1.3	1-1-2008 6:32:51 AM	8-K	A
5.	1.1.1.3	1-1-2008 6:33:51 AM	10-K	B
6.	1.1.1.3	1-1-2008 6:34:51 AM	8-K	A
7.	1.1.1.3	1-1-2008 6:35:55 AM	10-K	C
8.	1.1.1.3	1-1-2008 7:34:51 AM	8-K	A
9.	1.1.1.3	1-1-2008 7:35:51 AM	10-K	B

Table A2.
Variable Description

This table reports the construction of variables used in our regressions. We use CRSP monthly stock returns and Compustat quarterly data for the sample period 2009–2011. CRSP variable names are in parentheses and Compustat variable names are in square brackets. After collecting the raw Compustat data, in accordance with Bhojraj, Lee, and Oler (2003), we drop all firm–quarter observations missing data on total assets [atq], total long term debt [dlttq], net income before extraordinary items [ibq], debt in current liabilities [lctq], or operating income after depreciation [oiadpq]. Further, we require the raw share price on the last day of each fiscal quarter to be greater than \$3, both total common equity [ceqq] and total shareholder equity [seqq] to be positive, and net sales [saleq] to be more than \$100 million.

Variable	Description	Calculation
returns	Monthly cum-dividend stock returns	(ret)
Valuation Multiples		
pb	Price-to-book ratio	Market cap / total common equity [ceqq]
evs	Enterprise value- to- sales ratio	(Market cap + long-term debt [dlttq]) / net sales [saleq]
pe	Price-to-earnings ratio	Market cap net income before extraordinary items [ibq]
Financial Statement Ratios		
rnoa	Return on net operating assets	Net operating income after depreciation [oiadpq] / (property, plant, and equipment [ppentq] + current assets [actq] - current liabilities [lctq])
roe	Return on equity	Net income before extraordinary items [ibq] / total common equity [ceqq]
at	(Inverse of) Asset turnover	Total assets [atq] / net sales [saleq]
pm	Profit margin	Net operating income after depreciation [oiadp] / net sales [saleq]
lev	Leverage	Long-term debt [dlttq] / total stockholder's equity [seqq]
Other Financial Information		
salesgrowth	One-year-ahead realized sales growth	(Net sale one year ahead in the future - current year net sales) / current year net sales [saleq]
rdpersales	R&D expense- to- sales ratio	R&D expense [xrdq] / net sales [saleq]
ltgrowth	Median analyst long-term growth forecast	
ltgrowth spread	Standard deviation in analyst long-term growth forecast	
eps spread	Standard deviation in analyst one-year-ahead EPS forecast	
coverage	Number of analysts covering firm	
size	Market capitalization	Price (prc) × shares outstanding (shrout)

Table A3.
Distribution of SEC Filing Age at Time of User Download: Base vs. Peer Firms

This table provides the distribution of the age of SEC filing at the time of user download (access date – filing date) for all base-peer firm pairs. We group age of SEC filings into five categories: fewer than 6 months [0–6 Months], between 6 to 12 months [6–12 Months], between 12 to 18 months [12–18 Months], between 18 to 24 months [18–24 Months], and more than 24 months [24+ Months]. Row (column) categories represent base (peer) firm filing age. Table values in Columns (1)–(6) of Panel A (B) represent the number (percentage) of user-day searches that accessed base and peer firm filings in the relevant filing age categories.

Panel A. Frequency Distribution

	(1)	(2)	(3)	(4)	(5)	(6)
	0 to 6	6 to 12	12 to 18	18 to 24	24+	Total
Base / Peer Firm	Months	Months	Months	Months	Months	
0–6 Months	5,354,077	757,681	138,597	61,381	244,888	6,556,624
6–12 Months	879,859	940,781	60,251	49,214	108,741	2,038,846
12–18 Months	190,812	68,835	80,544	18,566	47,361	406,118
18–24 Months	80,681	60,213	18,230	47,973	42,959	250,056
24+ Months	336,031	153,046	46,778	42,169	572,548	1,150,572
Total	6,841,460	1,980,556	344,400	219,303	1,016,497	10,402,216

Panel B. Percentage Distribution

	(1)	(2)	(3)	(4)	(5)	(6)
	0 to 6	6 to 12	12 to 18	18 to 24	24+	Total
Base / Peer Firm	Months	Months	Months	Months	Months	
0–6 Months	51.47%	7.28%	1.33%	0.59%	2.35%	63.03%
6–12 Months	8.46%	9.04%	0.58%	0.47%	1.05%	19.60%
12–18 Months	1.83%	0.66%	0.77%	0.18%	0.46%	3.90%
18–24 Months	0.78%	0.58%	0.18%	0.46%	0.41%	2.40%
24+ Months	3.23%	1.47%	0.45%	0.41%	5.50%	11.06%
Total	65.77%	19.04%	3.31%	2.11%	9.77%	100.00%

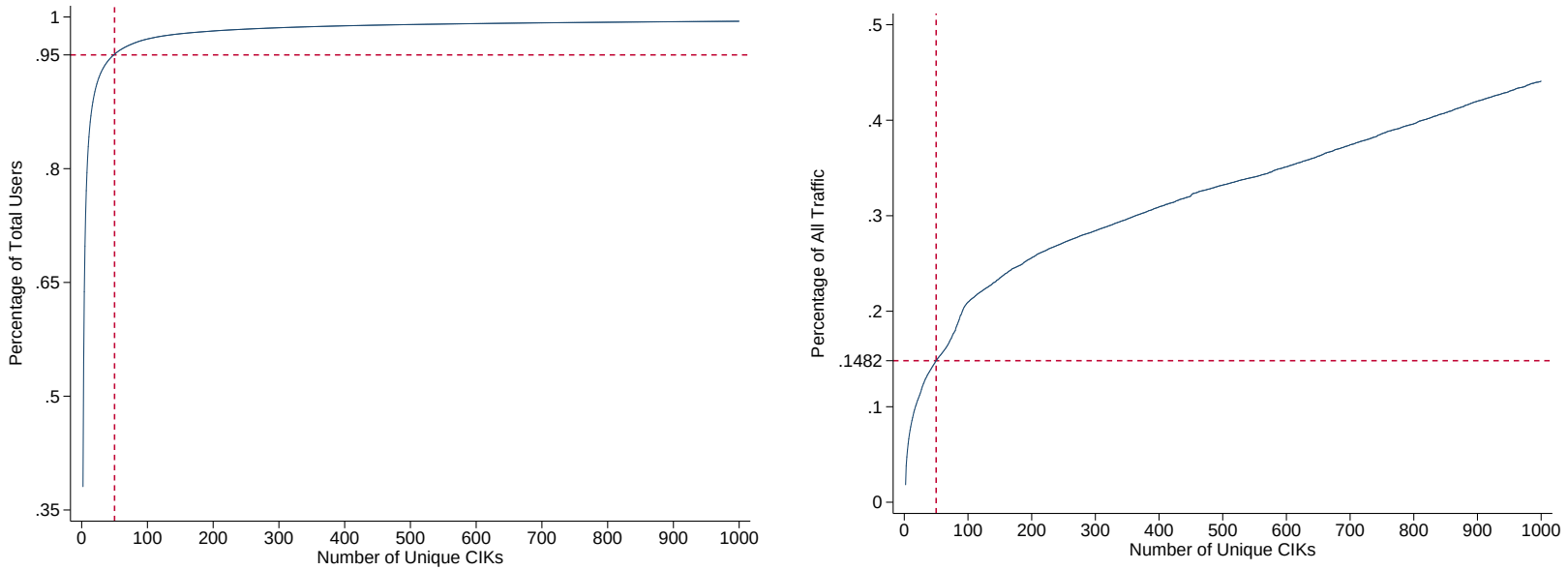
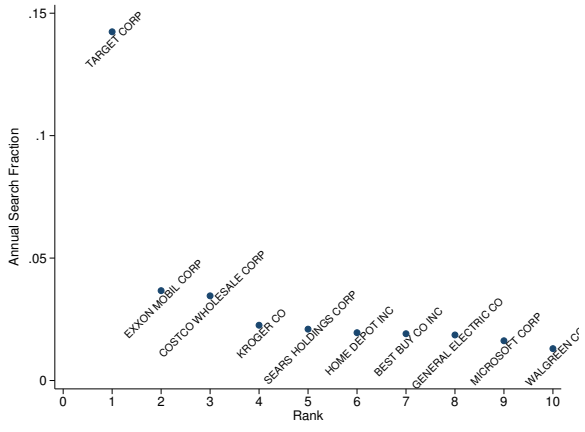
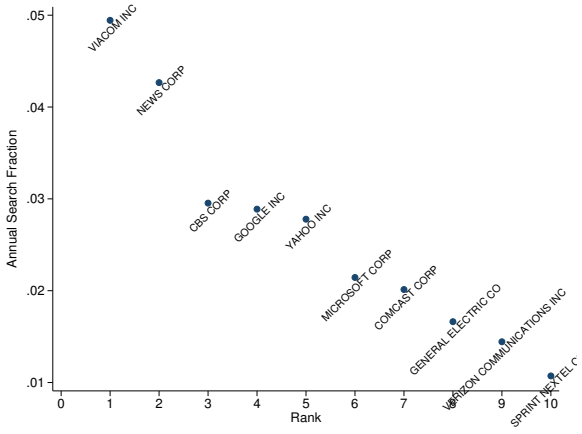


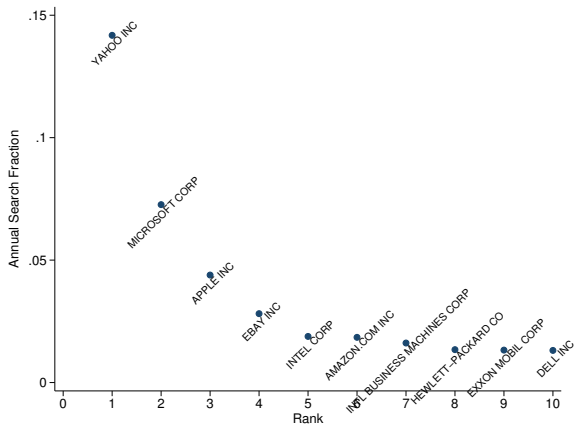
Figure 1. Distribution of Unique Firm Downloads. The left-hand-side graph plots the cumulative distribution function of the number of unique firm (CIK-based) downloads by daily-IP address observations: e.g., the percentage of daily-IP addresses downloading EDGAR filings of fewer than x number of unique CIKs between 2008-2011. The right-hand-side graph plots the percentage of total EDGAR search traffic generated by daily-IP addresses downloading fewer than x number of unique firm filings.



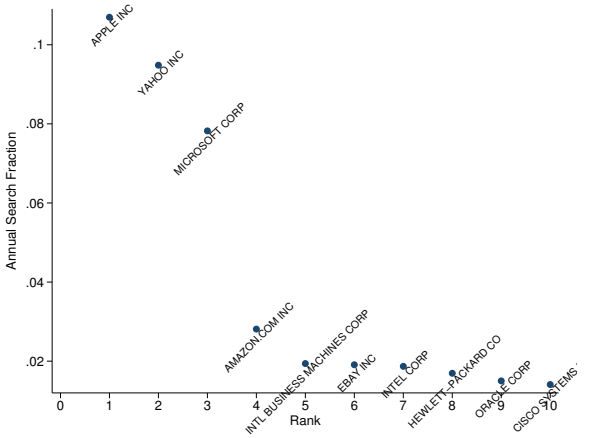
(a) Walmart in 2008



(b) Time Warner in 2008

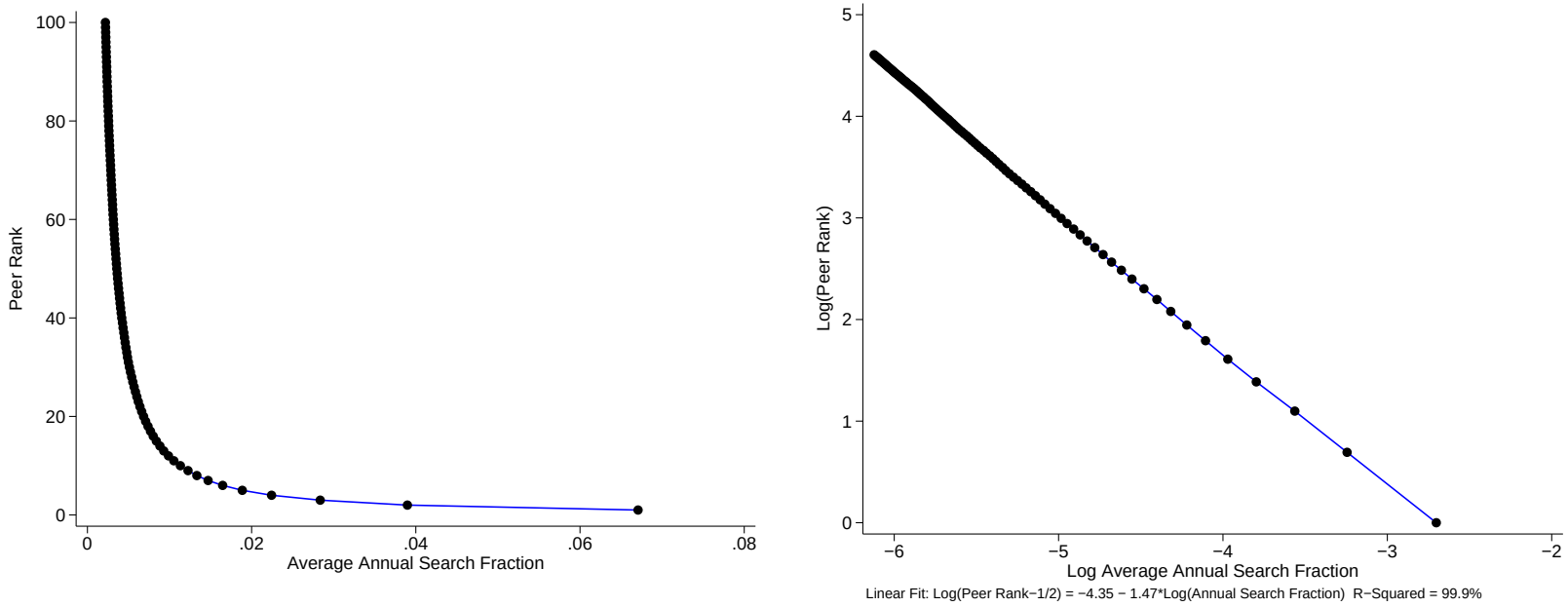


(c) Google in 2008



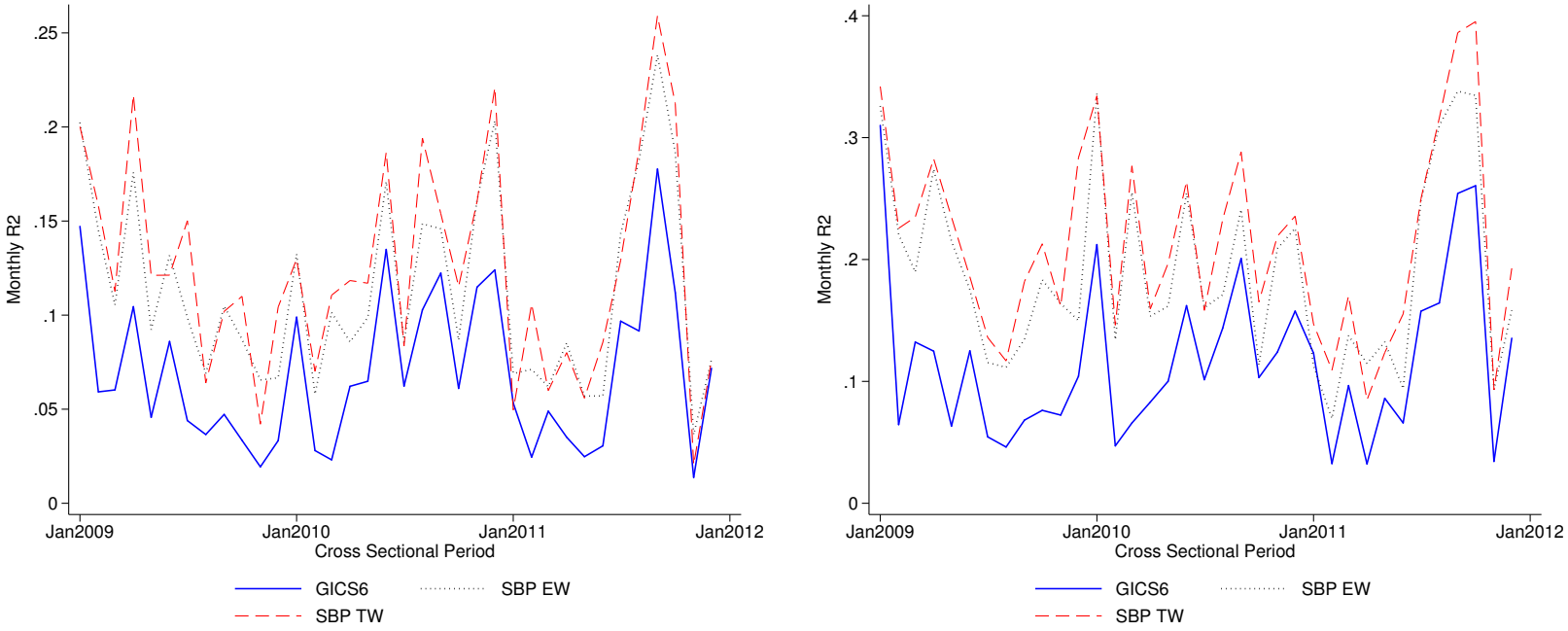
(d) Google in 2011

Figure 2. Examples of Peer Firms and Variation Across Time. These graphs illustrate examples of the 10 closest SBP firms for each base firm i, where the y-axis is the distance measure f_{ij}, which is the fraction of search traffic to firm j after searching firm i by the same user on a given day. Panels (a) and (b) illustrate Walmart and Time Warner in 2008. Panels (c) and (d) are depicted with Google as the base firm in different years to illustrate time series variation in SBPs.



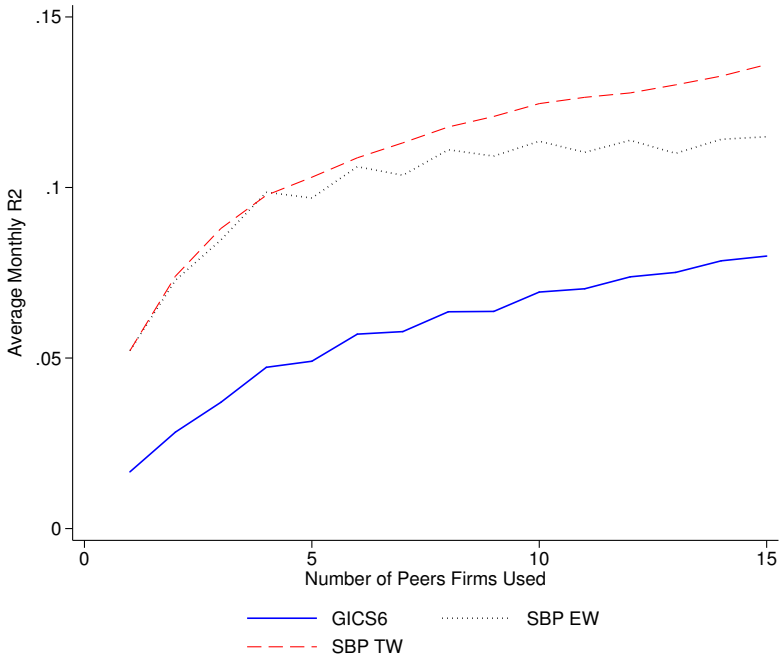
(a) Peer Rank vs. Average *Annual Search Fraction* **(b) Log Peer Rank vs. Log Average *Annual Search Fraction***

Figure 3. Distribution of Average *Annual Search Fraction*. The left-hand-side graph plots the average *Annual Search Fraction* for those firms identified as the "top 100" Search-Based Peers to a S&P1500 base firm in our data. For example, the last data point in the left-hand-side graph represents the average *Annual Search Fraction* for those firms with peer rank of 1. Similarly, the first data point on the left-hand-side graph represents the average *Annual Search Fraction* for those firms that with peer rank of 100. The right-hand-side graph plots the log of the peer rank vs. the log average *Annual Search Fraction*. The figure also reports results from a linear fit of the log-log relation with a $\frac{1}{2}$ adjustment factor for better finite sample properties (Gabaix and Ibragimov, 2011).

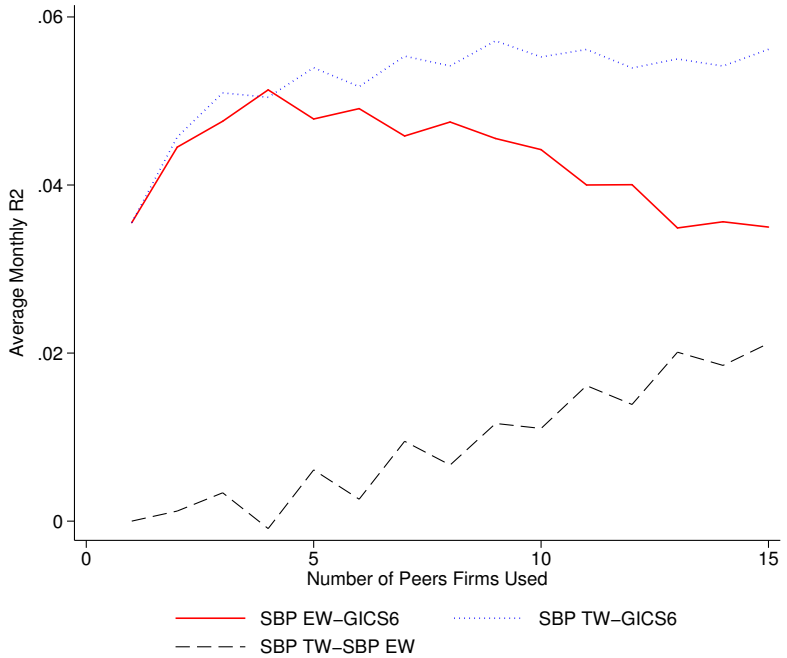


(a) S&P1500 Base Firms **(b) S&P500 Base Firms**

Figure 4. Time Series Variation in Cross-Sectional R^2 Values. These graphs depict the amount of cross-sectional variation in the base firm's stock returns explained by three different peer classification schemes. Each point represents the R^2 value from a cross-sectional regression of monthly firm i's returns on the average return of a portfolio of 10 similar firms, as defined through random selection based on either the GICS6 code, the average of the 10 closest equal weighted Search-Based Peers traffic measured firms (SBP EW), or the traffic-weighted average of the 10 closest peers firms (SBP TW). The specifications mirror the unrestricted S&P1500 peers results in Table 3. We report separate graphs for both the S&P1500 and S&P500 base firms in Panels (a) and (b). The monthly cross-sectional regressions span from January 2009 to December 2011.



(a) S&P1500 Base Firms



(b) S&P1500 Base Firms

Figure 5. Average Monthly R^2 Values and the Number of Peer Firms Used. These figures examine the sensitivity of the monthly return results to the number of peer firms used in the portfolio construction for the S&P1500 base firm sample. Each point in Panel (a) represents the average R^2 for 36 monthly cross-sectional return regressions, for each of the three measures: the GICS6, SBP EW, and SBP TW, as described in Figure 4. The relative differences between each benchmark for the same time series is plotted in Panel (b).

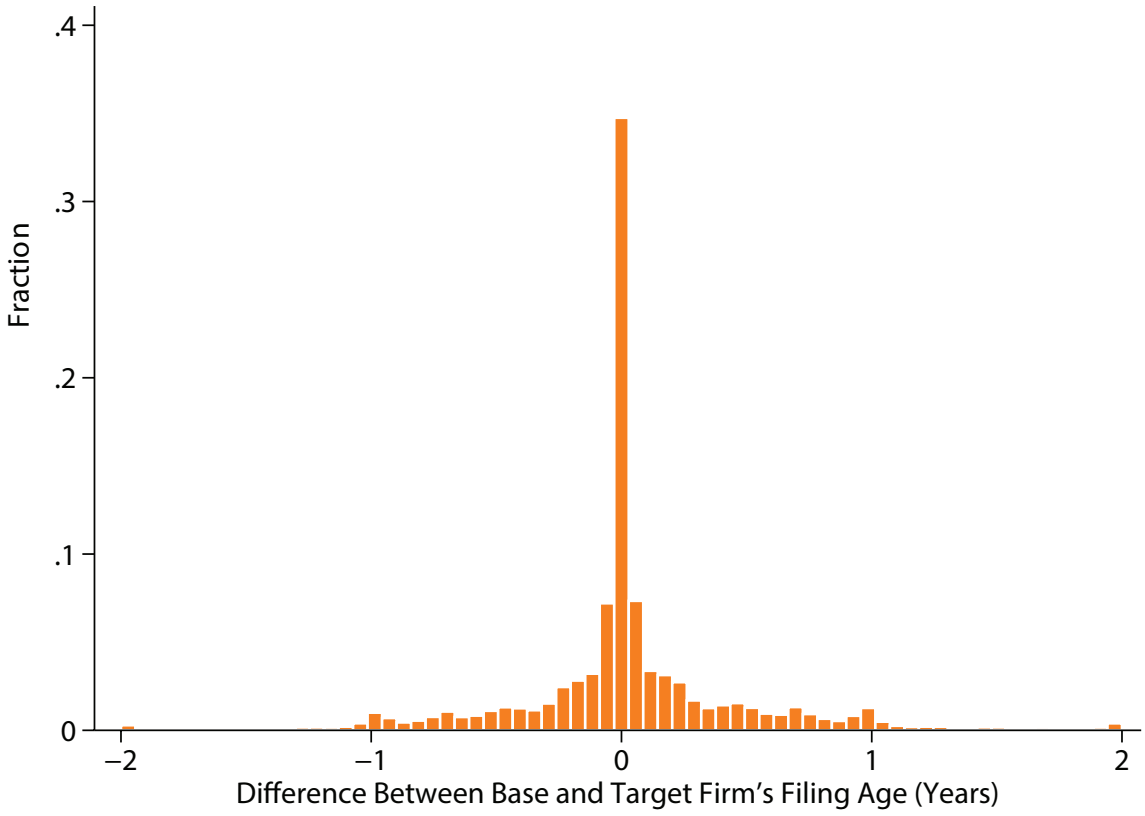


Figure 6. Distribution of Difference in Filing Age Between Base and Peer Firm This figure is a histogram of the difference in the filing age measured in years between the base and peer firm accessed by EDGAR users in the sample period 2008-2011.

Table 1.
Traffic Statistics

This table provides statistics on the sample of SEC EDGAR search traffic occurring between January 2008 and December 2011. Panel A reports our filtering process and the number of observations remaining after each filtering step. Step 1) reports the total number of filing downloads and the total number of daily unique visitors. In Step 2) we restrict searches for firms in the S&P1500 index as of January 1, 2008. In Step 3) we restrict traffic to users who search at least two unique firms (CIK-based) in order to apply our co-search *Annual Search Fraction* algorithm. In Step 4) to reduce the influence of bulk downloaders, we restrict searches to users who do not download more than 50 unique firms in a given day. In Step 5) we keep only the traffic page views for 10-K, 10-Q, 8-K, 14-A, and S-1 reports. Finally, in Step 6, we remove consecutive searches for base firm i by the same user in (a) and double counting of searches for the same pair of firms i and j by the same user more than once on a given day in (b). We link our search traffic from the SEC filings through the Wharton Research Data Services Central Index Key (CIK) to GVKEY master link file, which we then use to link to PERMNO via the CRSP/Compustat Merged Database. Panel B reports the number of base firms with SEC traffic coverage (defined as having at least 10 peer firms based on same-day searches). Also reported are the median numbers of total peer firms available by calendar year for the entire S&P1500 base firm sample.

Panel A. Data Filtering Steps

Filter Rule	# of Daily Visitor Page Views Obs	# of Daily Unique Visitors Obs
1) Raw Sample	3.53 billion	39.9 million
2) S&P1500 Firm Searches	810.7 million	16.2 million
3) Keep if Unique CIK>1	746.4 million	5.1 million
4) Keep if Unique CIK≤ 50	224.7 million	4.9 million
5) Keep 10-K, 10-Q, 8-K, 14-A, S-1 Searches	136.5 million	4.5 million
	# i then j search linkages	# Daily Unique Visitors Obs
6a) Removing consecutive searches for i by same user on a given day	12.2 million	3.2 million
6b) Remove double-counted users searching for same pair of firms i and j	10.4 million	3.2 million

Panel B. Coverage of S&P1500 Universe Conditional on 10 Peer Firms

Year	S&P500 Firms	S&P1500 Firms	Median Number of Peer Firms
2009	484	1460	469
2010	485	1461	493
2011	485	1462	520

Table 2.
Correspondence with Standard Industry Classifications

This table provides summary statistics on the degree of correspondence between our peer firm measure and other industry classification schemes. In Panel A, we sort firms by their degree of closeness in terms of search traffic and identify the top 10 SBPs for each base firm. For each peer, we report the average raw *Annual Search Fraction*, as well as the proportion that falls into each industry classification scheme. We define and compute *Annual Search Fraction* as

$$f_{ij}^{t} = \frac{\Sigma_{d=1}^{365}(\text{Daily Searches for } j \text{ after } i)_d}{\Sigma_{d=1}^{365}(\text{Daily Searches for } i \text{ and then any firm } j \neq i)_d}$$

between base firm i and peer firm j in calendar year t, or the unique number of daily unique visitor searches for firm j's information after searching for firm i's information. The entire sample comprises of search traffic on EDGAR between the years 2008 and 2011 and is comprised of a fixed S&P1500 sample as of January 1, 2008. The GICS, SIC, and NAICS classifications are all as of 2011. Panel B reports the degree of correspondence between our peer identification scheme and other industry groupings for each GICS2 industry sector. To construct this panel, we group each base firm by its GICS2 and report the fraction of its top 10 SBP firms that have the same industry classification (for GICS2, GICS6, SIC2, and NAICS3).

Panel A. Correspondence by Degree of Closeness of the Top 10 SBP Firms

Peer Firm Rank	Search Fraction	Same GICS2	Same GICS6	Same SIC2	Same NAICS3
1	0.07	0.83	0.70	0.68	0.65
2	0.04	0.77	0.62	0.62	0.60
3	0.03	0.74	0.56	0.56	0.54
4	0.02	0.70	0.52	0.52	0.49
5	0.02	0.67	0.47	0.48	0.45
6	0.02	0.65	0.45	0.46	0.43
7	0.01	0.64	0.43	0.44	0.42
8	0.01	0.62	0.40	0.42	0.40
9	0.01	0.61	0.39	0.40	0.38
10	0.01	0.59	0.37	0.38	0.36
Total	0.02	0.68	0.49	0.50	0.47

Panel B. Correspondence by GICS2 Industry for the Top 10 SBP Firms

GICS2 Groupings	Same GICS2	Same GICS6	Same SIC2	Same NAICS3
Energy	0.78	0.73	0.59	0.54
Materials	0.51	0.44	0.36	0.35
Industrials	0.54	0.34	0.29	0.26
Consumer Discretionary	0.66	0.49	0.42	0.42
Consumer Staples	0.62	0.48	0.47	0.43
Health Care	0.68	0.51	0.49	0.40
Financials	0.83	0.65	0.73	0.74
Information Technology	0.68	0.40	0.44	0.41
Telecommunication Services	0.67	0.48	0.72	0.67
Utilities	0.78	0.44	0.79	0.78

Table 3.
Comparison of R^2 Values: Monthly Returns

This table compares the average R^2 values from monthly cross-sectional regressions of the form

$$R_{i,t} = \alpha_t + \beta_t R_{p_i,t} + \epsilon_{i,t}$$

using CRSP returns data from January 2009 to December 2011. Columns (1)∼(3) report average R^2s from monthly cross-sectional regressions, regressing base firm (i) returns in a given month (t) on the concurrent returns of a portfolio (p_i) of 10 peers. Column (1) considers an equal-weighted portfolio of 10 randomly selected firms from the base firm's GICS6 industry and belonging to the S&P1500; Column (2) considers an equal-weighted portfolio (SBP EW) of the top 10 SBP firms, ranked by the prior calendar year's *Annual Search Fraction* f_{ij}, defined as the percentage of daily unique users searching for firm j after searching for firm i on the same day; Column (3) considers a portfolio (SBP TW) consisting of the top 10 SBP firms, with each peer firm weighted by the prior calendar year's *Annual Search Fraction* (relative to the top 10 peer firms). SBPs and portfolio weights are generated based on prior calendar year's EDGAR search traffic (e.g., the regressions in 2009 are generated with weights from calendar year 2008). Columns (4) and (5) test for the significance of the differences in average R^2s between the two SBP portfolio formulations and the GICS6 peer portfolios. Column (6) tests for the significance of the differences in average R^2s between the the SBP EW and SBP TW portfolios.

The results are reported for the sample of S&P1500 and S&P500 base firms. To facilitate comparisons, all the regressions are conducted using the same underlying set of firms. For example, when a base firm does not have at least 10 GICS peers but does have 10 peers based on the search traffic (and vice versa), we exclude the firm from the regression. The variable N in parentheses represents the average cross-sectional sample size for each monthly regression and standard errors are reported in square brackets. Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

	GICS6 (1)	SBP EW (2)	**SBP TW (3)**	(2)-(1) (4)	**(3)-(1) (5)**	(3)-(2) (6)
SP1500 Base Firms	**0.069*** **	0.114***	**0.125*** **	0.044***	**0.055*** **	0.011***
(N=1240)	**[0.007]**	[0.008]	**[0.010]**	[0.003]	**[0.005]**	[0.003]
SP500 Base Firms	**0.116*** **	0.189***	**0.214*** **	0.073***	**0.098*** **	0.024***
(N=407)	**[0.011]**	[0.013]	**[0.013]**	[0.007]	**[0.007]**	[0.005]
Number of Months	**36**	36	**36**	36	**36**	36

Table 4.
Comparison of R^2 Values: Valuation Multiples, Financial Ratios, and Other Financial Information

This table compares the average R^2 from several monthly cross-sectional regressions of the form

$$Var_{i,t} = \alpha_t + \beta_t Var_{p_i,t} + \epsilon_{i,t}$$

using most recently observable quarterly financial statement data from Compustat and market capitalization data from CRSP on March, June, September, and December of each year from 2009 to 2011. Columns (1)∼(3) report average R^2s from quarterly cross-sectional regressions, regressing base firm (i) Var in a given month (t) on the concurrent Var of a portfolio (p_i) of 10 peers. Each row considers a different Var, as defined in Table A2. Column (1) considers an equal-weighted portfolio of 10 randomly selected firms from the base firm's GICS6 industry; Column (2) considers an equal-weighted portfolio (SBP EW) of the top 10 SBP firms, ranked by the prior calendar year's *Annual Search Fraction* f_{ij}, defined as the percentage of unique users searching for firm j after searching for firm i on the same day; Column (3) considers a portfolio (SBP TW) consisting of the top 10 SBP firms, with each peer firm weighted by the prior calendar year's *Annual Search Fraction* (relative to the top 10 SBP firms). SBP firms and portfolio weights are generated based on prior calendar year's EDGAR search traffic (e.g., the regressions in 2009 are generated with weights from calendar year 2008). Columns (4) and (5) test for the significance of the differences in average R^2s between the two SBP portfolio formulations and the GICS6 peer portfolios. Column (6) tests for the significance of the differences in average R^2s between the SBP EW and SBP TW portfolios.

Regressions are performed for the sample of S&P1500 base firms. To facilitate comparisons, all the regressions are conducted using the same underlying set of base firms. For example, when a base firm does not have at least 10 GICS peers but does have 10 peers based on the search traffic (and vice versa), we exclude the firm from the regression. In addition, for regressions involving *pe* we also drop observations with negative net income before extraordinary items and for regressions involving *rnoa* we drop observations when values are missing for current assets, current liabilities, or property, plant, and equipment. The variable N in parentheses represents the average cross-sectional sample size for each quarterly regression and standard errors are reported in square brackets. Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

Table 4.
(Continued)

	GICS6 (1)	SBP EW (2)	SBP TW (3)	(2)-(1) (4)	(3)-(1) (5)	(3)-(2) (6)
Valuation Multiples						
pb (N=793)	0.028***	0.121***	0.124***	0.093***	0.096***	0.003
	[0.003]	[0.008]	[0.011]	[0.009]	[0.012]	[0.005]
evs (N=795)	0.234***	0.403***	0.439***	0.169***	0.205***	0.036***
	[0.006]	[0.009]	[0.008]	[0.008]	[0.006]	[0.003]
pe (N=676)	0.022***	0.025***	0.035***	0.003	0.013*	0.009***
	[0.005]	[0.005]	[0.007]	[0.005]	[0.006]	[0.003]
Financial Statement Ratios						
rnoa (N=783)	0.192***	0.238***	0.278***	0.046***	0.086***	0.041***
	[0.012]	[0.012]	[0.012]	[0.009]	[0.009]	[0.005]
roe (N=790)	0.021***	0.056***	0.067***	0.036***	0.046***	0.010***
	[0.006]	[0.007]	[0.008]	[0.005]	[0.005]	[0.003]
at (N=792)	0.434***	0.596***	0.617***	0.161***	0.183***	0.021***
	[0.009]	[0.008]	[0.008]	[0.005]	[0.006]	[0.001]
pm (N=790)	0.150***	0.261***	0.298***	0.111***	0.147***	0.036***
	[0.009]	[0.015]	[0.016]	[0.013]	[0.013]	[0.003]
lev (N=797)	0.029***	0.065***	0.066***	0.037***	0.037***	0.000
	[0.003]	[0.008]	[0.005]	[0.007]	[0.005]	[0.004]
Other Financial Information						
ltgrowth (N=658)	0.125***	0.171***	0.180***	0.046***	0.055***	0.009
	[0.007]	[0.010]	[0.015]	[0.008]	[0.012]	[0.007]
salesgrowth (N=761)	0.128***	0.159***	0.175***	0.030**	0.047***	0.016***
	[0.012]	[0.014]	[0.017]	[0.010]	[0.013]	[0.004]
rdpersales (N=792)	0.629***	0.655***	0.686***	0.025***	0.056***	0.031***
	[0.008]	[0.007]	[0.008]	[0.004]	[0.004]	[0.003]
Number of Quarters	12	12	12	12	12	12

Table 5.
Explaining Search-Based Peer (SBP) Choices

The sample consists of the top 10 SBPs to each base firm in our sample as well as all other firms from the same GICS2 sector. The dependent variable is an indicator for being a top 10 SBP to a base firm in our sample. Explanatory variables are ranked absolute % difference in firm fundamentals between the base firm and the peer firm, where 1 indicates $|\frac{Var_{peer}}{Var_{base}} - 1| \leq 25\%$, 2 indicates $25\% < |\frac{Var_{peer}}{Var_{base}} - 1| \leq 50\%$, and 3 indicates $50\% < |\frac{Var_{peer}}{Var_{base}} - 1|$. "Different GICS4" is an indicator variable equaling 1 when the base firm and the peer firm belong to different 4-digit GICS industry groups. Year fixed effects are included throughout. Odd [even] columns of this table reports probit coefficients [marginal effects ("*MFX*")]. Marginal effects are evaluated at 1 for all explanatory variables with the exception of "Different GICS4" and the year fixed effects, which are evaluated at 0. Standard errors are clustered at the base-firm level. Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

	Fundamental Sample		*Analyst Sample*	
	(1)	(2)	(3)	(4)
Var	Top 10	Top 10 *MFX*	Top 10	Top 10 *MFX*
size	-0.1408***	-0.0454***	-0.0680***	-0.0254***
	(0.007)	(0.003)	(0.010)	(0.004)
pb	-0.0193***	-0.0062***	-0.0072	-0.0027
	(0.004)	(0.001)	(0.006)	(0.002)
pe	-0.1722***	-0.0556***	-0.1433***	-0.0535***
	(0.004)	(0.002)	(0.005)	(0.002)
rnoa	-0.1068***	-0.0345***	-0.0755***	-0.0282***
	(0.005)	(0.002)	(0.006)	(0.002)
roe	-0.0324***	-0.0105***	-0.0103*	-0.0038
	(0.005)	(0.002)	(0.006)	(0.002)
at	-0.1065***	-0.0344***	-0.1024***	-0.0383***
	(0.005)	(0.002)	(0.007)	(0.003)
evs	-0.0733***	-0.0236***	-0.0705***	-0.0263***
	(0.005)	(0.002)	(0.007)	(0.003)
lev	-0.0837***	-0.0270***	-0.0799***	-0.0299***
	(0.005)	(0.002)	(0.007)	(0.003)
salesgrowth	-0.0719***	-0.0232***	-0.0600***	-0.0224***
	(0.004)	(0.001)	(0.006)	(0.002)
rdpersales	-0.1194***	-0.0385***	-0.1093***	-0.0409***
	(0.007)	(0.003)	(0.010)	(0.004)
coverage			-0.0973***	-0.0364***
			(0.010)	(0.004)
eps spread			-0.0136**	-0.0051**
			(0.005)	(0.002)
ltgrowth			-0.0784***	-0.0293***
			(0.007)	(0.003)
ltgrowth spread			-0.0508***	-0.0190***
			(0.005)	(0.002)
Different GICS4	-0.2020***	-0.0607***	-0.3031***	-0.1057***
	(0.015)	(0.004)	(0.019)	(0.007)
Observations	2,264,948	2,264,948	544,451	544,451
Pseudo R^2	0.0614	0.0614	0.0630	0.0630

Table 6.
Comparison of R^2 by Base Firm's Industry and Size

This table compares the average R^2 values from monthly cross-sectional regressions of the form

$$R_{i,t} = \alpha_t + \beta_t R_{p_i,t} + \epsilon_{i,t}$$

using CRSP returns data from January 2009 to December 2011. In both Panels, Columns (1)∼(3) report average R^2s from monthly cross-sectional regressions, regressing base firm (i) returns in a given month (t) on the concurrent returns of a portfolio (p_i) of 10 peers. Column (1) considers an equal-weighted portfolio of 10 randomly selected firms from the base firm's GICS6 industry; Column (2) considers an equal-weighted portfolio of the top 10 SBP firms, ranked by the prior calendar year's *Annual Search Fraction* f_{ij}, defined as the percentage of unique users searching for firm j after searching for firm i on the same day; Column (3) considers a portfolio consisting of the top 10 SBP firms, with each peer firm weighted by the prior calendar year's *Annual Search Fraction* (relative to the top 10 SBP firms). SBPs and portfolio weights are generated based on prior calendar year's EDGAR search traffic (e.g., the regressions in 2009 are generated with weights from calendar year 2008). Columns (4) and (5) test for the significance of the differences in average R^2s between the two SBP portfolio formulations and the GICS6 peer portfolios. Column (6) tests for the significance of the differences in average R^2s between the equal-weighted and traffic-weighted SBP portfolios.

The four rows in each Panel of the table report results for subsets of base firms. Panel A reports results by the two-digit GICS classification of base firms in the S&P1500 universe (we exclude industries with fewer than 15 firms). Panel B reports results by market cap decile of base firms in the S&P1500 universe. Decile breakpoints are measured at the end of December in the prior calendar year of the S&P1500 sample. To facilitate comparisons, all the regressions are conducted using the same underlying set of firms. For example, when a base firm does not have at least 10 GICS peers but does have 10 peers based on the search traffic (and vice versa), we exclude the firm from the regression. The variable N in parentheses represents the average cross-sectional sample size for each monthly regression and standard errors are reported in square brackets. Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

Table 6.
(Continued)

Panel A. R^2 by Base Firm's Industry

	GICS6 (1)	SBP EW (2)	SBP TW (3)	(2)-(1) (4)	(3)-(1) (5)	(3)-(2) (6)
Energy (N=83)	0.007*	0.083***	0.100***	0.076***	0.093***	0.017*
	[0.004]	[0.015]	[0.019]	[0.015]	[0.019]	[0.009]
Materials (N=73)	0.024***	0.059***	0.079***	0.035***	0.055***	0.021**
	[0.007]	[0.013]	[0.015]	[0.009]	[0.012]	[0.009]
Industrials (N=149)	0.010***	0.047***	0.057***	0.037***	0.047***	0.010
	[0.003]	[0.007]	[0.011]	[0.006]	[0.009]	[0.008]
Consumer	0.025***	0.076***	0.104***	0.051***	0.079***	0.028***
Discretionray (N=214)	[0.005]	[0.009]	[0.012]	[0.006]	[0.010]	[0.006]
Consumer Staples	0.009*	0.028**	0.049***	0.019*	0.040***	0.021**
(N=49)	[0.005]	[0.012]	[0.014]	[0.010]	[0.012]	[0.008]
Health Care (N=147)	0.010***	0.051***	0.074***	0.040***	0.063***	0.023***
	[0.003]	[0.009]	[0.012]	[0.008]	[0.011]	[0.008]
Financials (N=223)	0.042***	0.101***	0.111***	0.059***	0.069***	0.010*
	[0.009]	[0.013]	[0.014]	[0.008]	[0.010]	[0.005]
Information	0.010***	0.020***	0.027***	0.010**	0.017***	0.007**
Technology (N=235)	[0.003]	[0.004]	[0.005]	[0.004]	[0.005]	[0.003]
Utilities (N=63)	0.022***	0.051***	0.055***	0.029***	0.033***	0.004
	[0.008]	[0.011]	[0.012]	[0.008]	[0.009]	[0.008]
Number of Months	36	36	36	36	36	36

Panel B. R^2 by Base Firm's Size Decile

	GICS6 (1)	SBP EW (2)	SBP TW (3)	(2)-(1) (4)	(3)-(1) (5)	(3)-(2) (6)
1 (N=124)	0.019***	0.037***	0.050***	0.018***	0.030***	0.013
	[0.004]	[0.008]	[0.010]	[0.006]	[0.009]	[0.008]
2 (N=115)	0.038***	0.065***	0.084***	0.027***	0.046***	0.019***
	[0.006]	[0.008]	[0.012]	[0.006]	[0.010]	[0.007]
3 (N=118)	0.057***	0.093***	0.104***	0.036***	0.047***	0.011*
	[0.008]	[0.011]	[0.013]	[0.007]	[0.009]	[0.006]
4 (N=128)	0.075***	0.121***	0.121***	0.046***	0.046***	-0.000
	[0.009]	[0.012]	[0.013]	[0.008]	[0.010]	[0.006]
5 (N=118)	0.093***	0.139***	0.150***	0.046***	0.057***	0.011
	[0.012]	[0.016]	[0.016]	[0.011]	[0.013]	[0.009]
6 (N=128)	0.108***	0.147***	0.164***	0.039***	0.056***	0.018*
	[0.010]	[0.014]	[0.014]	[0.009]	[0.010]	[0.009]
7 (N=126)	0.103***	0.181***	0.189***	0.078***	0.086***	0.008
	[0.011]	[0.016]	[0.015]	[0.009]	[0.010]	[0.008]
8 (N=119)	0.113***	0.177***	0.193***	0.063***	0.080***	0.016**
	[0.013]	[0.016]	[0.019]	[0.010]	[0.012]	[0.008]
9 (N=135)	0.145***	0.222***	0.247***	0.077***	0.101***	0.025**
	[0.015]	[0.016]	[0.017]	[0.009]	[0.011]	[0.009]
10 (N=126)	0.128***	0.192***	0.223***	0.064***	0.095***	0.031***
	[0.013]	[0.017]	[0.017]	[0.011]	[0.011]	[0.009]
Number of Months	36	36	36	36	36	36

Table 7.
Comparison of R^2 by Robustness Test: Alternative Co-Search Algorithm and Size Matching

This table compares the average R^2 values from monthly cross-sectional regressions of the form

$$R_{i,t} = \alpha_t + \beta_t R_{p_i,t} + \epsilon_{i,t}.$$

Unless otherwise specified, each robustness specification is a specific variant of the baseline case of 10 peer firms using the sample of S&P1500 base firms with S&P1500 peer firms with an unique CIK cutoff of 50 firms. The number of available base firms for each specification is reported in parentheses. The first row references the baseline result mirrored in Table 3. For the second specification, we redefine the time window for measuring *Annual Search Fraction*s from a yearly aggregation of daily co-searches to session co-searches. Each session is defined as a contiguous block of search activity separated by a minimum of one hour of inactivity by the same IP on the same day. In the third specification, we use an equal-weighted portfolio of 5 randomly selected peer firms from the base firm's GICS6 industry and market cap quartile. Specification 4 considers an equal-weighted portfolio of 5 firms whose market cap are the closest to the base firm's and who belong to the same GICS6 as the base firm. In the 5th specification, we reformulate our *Annual Search Fraction* f_{ij} to include the percentage of unique users searching for firm j before or after searching for firm i on the same day. In the 6th specification, we take into account the popularity (or baseline probability) of a peer firm j being searched in any search sequence. Specifically we recalculate our *Annual Search Fraction* as

$$\hat{f_{ij}} = \frac{f_{ij} \equiv \frac{i \text{ then } j}{i}}{\frac{!i \text{ then } j}{!i}}$$

Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

	GICS6 (1)	SBP EW (2)	SBP TW (3)	(2)-(1) (4)	(3)-(1) (5)	(3)-(2) (6)
1. Baseline (N=1240)	0.069***	0.114***	0.125***	0.044***	0.055***	0.011***
	[0.007]	[0.008]	[0.010]	[0.003]	[0.005]	[0.003]
2. Size Quartile Matched GICS6 (5 Peers) (N=1077)	0.059***	0.104***	0.113***	0.045***	0.054***	0.009**
	[0.007]	[0.008]	[0.009]	[0.005]	[0.005]	[0.003]
3. Nearst Size Matched GICS6 (5 Peers) (N=1344)	0.066***	0.097***	0.104***	0.032***	0.038***	0.006**
	[0.007]	[0.007]	[0.008]	[0.006]	[0.006]	[0.003]
4. Session Level (N=1240)	0.069***	0.112***	0.124***	0.043***	0.054***	0.011***
	[0.007]	[0.008]	[0.009]	[0.003]	[0.005]	[0.003]
5. Upstream and Downstream Traffic (N=1240)	0.069***	0.116***	0.129***	0.046***	0.060***	0.013***
	[0.007]	[0.008]	[0.010]	[0.004]	[0.005]	[0.003]
6. Baseline Probability Correction (N=1240)	0.069***	0.110***	0.119***	0.041***	0.050***	0.009***
	[0.007]	[0.008]	[0.009]	[0.003]	[0.005]	[0.003]
Number of Months	36	36	36	36	36	36

Table 8.
Comparison of R^2 by Robustness Test: Alternative Data Sampling Criterion

This table compares the average R^2 values from monthly cross-sectional regressions of the form

$$R_{i,t} = \alpha_t + \beta_t R_{p_i,t} + \epsilon_{i,t}.$$

Unless otherwise specified, each robustness specification is a specific variant of the baseline case of 10 peer firms using the sample of S&P1500 base firms with S&P1500 peer firms with an unique CIK cutoff of 50 firms. The number of available base firms for each specification is reported in parentheses. The first row references the baseline result mirrored in Table 3. In specification 2, we use the traffic sample comprised of searches by users who search for more than 50 unique firms(CIK-based) on a given day to demonstrate the relative lack of intelligence of bulk downloaders. For specifications 3 and 4, we divide our baseline sample into two partitions, depending on whether the search was for .txt (10% of data) or .html (90% of data) filings on EDGAR. For specifications 5 and 6, we divide our baseline sample into two partitions, depending on the filing age of the base firm. The filing age is defined as the difference between the date of the click and the filing date of the file accessed. The partition is based on whether the filing age is more (60% of data) or less than 1 month old (40% of data). For specification 7, using the entire baseline sample, we weight base firm's searches by filing age. The weighting scheme is such that filings with zero age (click on same day as filing) have a maximum weight of 1, while filings with age greater than 2 year (11% of sample) have a weight of 0 and filings in between are weighted linearly. Significance levels are indicated by *, **, *** for 10%, 5%, and 1%, respectively.

	GICS6 (1)	SBP EW (2)	SBP TW (3)	(2)-(1) (4)	(3)-(1) (5)	(3)-(2) (6)
1. Baseline (N=1240)	0.069***	0.114***	0.125***	0.044***	0.055***	0.011***
	[0.007]	[0.008]	[0.010]	[0.003]	[0.005]	[0.003]
2. Robot Sample	0.069***	0.010***	0.008***	-0.059***	-0.061***	-0.002**
(N=1240)	[0.007]	[0.002]	[0.001]	[0.006]	[0.006]	[0.001]
3. HTML Only Sample	0.069***	0.115***	0.126***	0.045***	0.057***	0.011***
(N=1240)	[0.007]	[0.008]	[0.009]	[0.003]	[0.005]	[0.003]
4. TXT Only Sample	0.070***	0.031***	0.031***	-0.039***	-0.038***	0.000
(N=1179)	[0.007]	[0.003]	[0.004]	[0.005]	[0.005]	[0.002]
5. Filing Age $\leq$ 6 Month	0.069***	0.111***	0.125***	0.042***	0.055***	0.014***
(N=1239)	[0.007]	[0.008]	[0.009]	[0.003]	[0.005]	[0.003]
6. Filing Age $>$ 6 Month	0.069***	0.098***	0.108***	0.029***	0.038***	0.009***
(N=1238)	[0.007]	[0.008]	[0.008]	[0.003]	[0.004]	[0.002]
7. Filing Age Time	0.069***	0.116***	0.127***	0.047***	0.058***	0.011***
Weighted (N=1240)	[0.007]	[0.008]	[0.010]	[0.004]	[0.005]	[0.003]
Number of Months	36	36	36	36	36	36