Aggregate attention*

Paul J. Irvine
Neeley School of Business
Texas Christian University

Siyi Shen
CUHK-Shenzhen

Tao Shu
CUHK-Shenzhen

April, 2022

Abstract

We study some of the aggregate properties of investor attention to the stock market. We introduce a framework that constructs a null hypothesis of what rational aggregate attention should look like. This framework states that aggregate attention should be proportional to the aggregate wealth invested in each stock. We find that the distribution of attention at first appears rational. However, much of this attention is directed away from the high market cap names that should attract the most attention. Attention is notably more volatile from month to month than market cap, making attention unpredictable from a market maker perspective This pattern generates pricing errors that result in the most popular stocks performing poorly in the upcoming month. This poor performance appears to be a short-term reversal of pricing errors generated by the unpredictable liquidity demands volatile attention can generate.

*Preliminary, all errors are our own. Thanks to Allen Berger, David Rakowski, Joey Engelberg, Kevin Pisciotta, and seminar participants at the University of South Carolina, the University of Nebraska, Iowa State University, and the University of Kansas. Corresponding author: p.irvine@tcu.edu.

Aggregate Attention

Abstract

We study some of the aggregate properties of investor attention to the stock market. We introduce a framework that constructs a null hypothesis of what rational aggregate attention should look like. This framework states that aggregate attention should be proportional to the aggregate wealth invested in each stock. We find that the distribution of attention at first appears rational. However, much of this attention is directed away from the high market cap names that should attract the most attention. Attention is notably more volatile from month to month than market cap, making attention unpredictable from a market maker perspective This pattern generates pricing errors that result in the most popular stocks performing poorly in the upcoming month. This poor performance appears to be a short-term reversal of pricing errors generated by the unpredictable liquidity demands volatile attention can generate.

1 Introduction

Attention is a scarce resource. How do investors allocate their attention across stocks? Do they allocate attention rationally? What constitutes a measure of rational attention? This paper intends to provide evidence on these questions. We use a large data set of investor posts from the website Stocktwits, to evaluate the distribution of attention across stocks. In addition, we evaluate how attention responds to stock returns, news articles, and other factors proposed in the literature to proxy for attention. The goal of this paper is to provide a broad set of facts on aggregate investor attention, that can hopefully guide the burgeoning literature on attention and its effects in financial economics.

What should be the rational level of investor attention? It is almost tautological nowadays to assume that an individuals level of attention is limited, sometimes this bound is imposed by cognitive capability, in other settings by the restrictions of the effort required, or the limited time involved, sometimes by all of these factors. An early attempt to incorporate these limitations to explain stock returns is Merton (1987) who notes, quite correctly, that investors seem limited in the set of stocks they 'know about', and thus pay attention to. Casual observation leads to the obvious conclusion that different individual investors pay attention to different sets of stocks. So how do these distinct attention sets aggregate in the stock market? We propose a simple framework that assumes investors should aggregate attention in proportion to their wealth invested in different stocks. Fortunately, we know how wealth aggregates across stocks since we can easily calculate market capitalization weights. Therefore, we propose a null hypothesis that attention across investors should aggregate in proportion to a stock's market cap weight.

We examine the aggregate properties of investor attention using data from the Stocktwits web site. Stocktwits is a site where investor can post short notes on individual stocks. These notes are aggregated into a thread that is continually updated as new investors post their notes or reactions to the market action, news reports, dividend news, or other posts. Stocktwits has grown considerably since its introduction in 2008, and currently has the advantage of covering a large and broad set of investors and stocks. In using this data, we do assume that posting about a stock implies the investor is paying attention to the stock. This does not seem like an extreme assumption since to post an investor must identify the stock by ticker symbol, so clearly the attention is specific to that stock, and this attention is recorded when they take enough interest in the stock to write a post about it. In doing so, they are very likely follow the thread containing other posts on the same stock, thus applying their limited supply of attention to the posted stock.

Existing papers using network data concentrate on the effects of abnormal attention on sentiment or stock returns. For example, Da, Engelberg and Gao (2011) use Google search intensity as a proxy for attention and show that interest in the ticker is related to IPO initial returns. Ben-Rephael, Da, and Israelsen (2017) use the Bloomberg news network to examine the effects of abnormal institutional attention on the speed of return responses to news events. Because of restrictions on the way Google search intensity and Bloomberg news are disseminated, they are problematical for measuring aggregate attention. Social networks are studied by Giannini, Irvine and Shu (2018) who relate social media sentiment to overpricing by physically distant investors. Rakowski, Shirley and Stark (2020) use Twitter outages to infer the effects of social media activity on volume and returns. As in Giannini et al. (2019), Rakowski et al. (2020) find that social media activity can be particularly revealing around earnings announcements. A number of more obscure attention and return studies have been performed in international settings, usually with Twitter since it has the most accessible API. These international studies produce mixed results; usually they report a relation between attention and returns, but it is not always a relation that is consistent with other findings (Yoshinaga and Rocco, 2020). Stocktwits data has proven to be particularly useful in Cookson and Neissner's (2020) study of disagreement and investor type, in Giannini et al.'s (2019) study of investor disagreement around earnings announcements, as a data source to define political leanings in Cookson, Engelberg, and Mullins (2020), and in Cai, Yung, and Zhu's (2019) study of sentiment and post-earnings announcement drift.

In his treatise on attention and effort, Kahneman (1973) discusses why individuals appear to selectivity attend some stimuli, in preference of others This selective attention is set in a framework where attention requires effort, a resource that is limited, and therefore must be distributed selectively. Most of the literature on attention in psychology has focused on what stimuli attract individual attention. In economics, a natural focus point is how limited attention affects consumer choice (De Clippel, Eliaz, and Rozen, 2014). However, Kahneman's (1973) ideas of a capacity on attention are used in accounting and finance papers like Hirshleifer and Teoh (2003), Peng and Xiong (2006), DellaVigna and Pollet (2009), and Hirshleifer, Kim and Teoh (2009). In this paper, we focus on aggregate attention first examining how the collective selective attention by individuals aggregates to a distribution of attention across stocks.

Our first empirical tests examine the concentration of investor attention and compares the level of concentration in attention to the concentration of market value and volume. We represent concentration as a power law, where the exponent of the power law can be estimated empirically, and yields a compact way to describe concentration across a large set of stocks. Using a sample of 100 stocks, the initial indication is that attention is rational in that investors allocate their attention in a similar way to the distribution of market capitalization is allocated across stocks. However, when we expand the set of stock examined to 200, 500, or 1,000 stocks, we find that investor attention tends to be more and more concentrated than market capitalization. In fact, attention and trading volume tend to reflect a similar pattern, in that volume is more concentrated than market cap as well. This change in attention concentration as we expand the sample indicates that investor attention is rational relative to market cap, at least for the top 100 stocks, but investors are relatively inattentive as we go down the market cap scale. This finding provides behavioral support for the 'neglected firm' effect, first proposed by Arbel and Strebel (1982).

Some of the characteristics of aggregate attention can be important for models such as Hendershott, Menkveld, Praz, and Seasholes (2021) who extend the slow moving capital ideas of Bogousslavsky (2016) and Duffie (2010) to explain mispricing and mean reversion in prices at different time scales. Their model relies on two groups of inattentive investors whose lack of attention drives episodic liquidity demands that can distort prices. Their model has little behavioral support, it is a proposal that appears to fit several market regularities. We provide behavioral support to this theory in several ways, including the finding that aggregate attention is quite volatile on a month to month basis. About 60 percent of the stocks that attract the most attention in a particular month are also attracting the most attention in the following month. This percentage implies that attention-driven liquidity demands can rotate significantly and could be the driving force for liquidity shocks that impact efficient pricing. At this point it should be noted that the top volume stocks are more stable month to month than the top attention stocks, and volume is the key driver of the market maker problem in Hendershott et al. (2021). However, since attention in Hendershott, et al. (2021) is empirically represented by retail order flow, our evidence indicates that order flow is an imperfect proxy for attention, since only between 35 and 53 percent of firms are in the same high volume-high attention groups across two adjacent months.[1] Although attention tracks volume more closely than attention tracks market cap, there are still significant differences, and attention is not always perfectly reflected in trading volume.

Other models of attention, though limited in focus, have proven particularly powerful in finance applications. Kacperczyk, Van Niewerburgh and Veldkamp (2016) present a model where fund managers optimally focus attention on either systematic or aggregate factors depending on the business cycle, with systematic factors attracting more attention in recessions, and idiosyncratic factors attracting more attention in booms. They claim that such a pattern of attention allocation over the business cycles produces the same pattern of time varying skill as observed in the data. Chinco (2021) attempts to infer the ex-ante likelihood of bubbles, using a model where returns above a certain threshold stimulate speculative attention to particular stocks. Attention and returns are tied together in Chinco (2021) through

[1] We examine the top attention and volume groups using sets of between 50 and 1,000 stocks. As this threshold rises, there are more stocks in common between attention and volume.

social interactions that become more persuasive when past returns reach a threshold level. The implicit tie into attention in Chinco (2021) is that most investor attention to particular stocks is latent, until attention is stimulated by the arguments of their personal network. This attention to particular stocks reaches a level that overwhelms the remaining rational investors who would normally arbitrage price back to fundamental levels. We directly test whether such a threshold level of attention exists for individual stocks, and find a consistent increment in attention when returns reach a monthly threshold of +/- 20 percent.

The set of stocks that investors pay attention to can change, usually if a new stock comes into the opportunity set of a particular investor. Precisely when a new stock comes into an investors opportunity set is difficult to measure, so Barber and Odean (2008) instead test this idea by examining attention-grabbing events, such as returns, volume and news. We use the Barber and Odean (2008) idea of attention grabbing events, to motivate an investigation of these events on investor attention. In the Chinco (2021) model, past returns are the a bubble triggering mechanism, but one could imagine news events being a trigger as well, perhaps by driving returns above the Chinco (2021) threshold. In this paper we hope to provide some evidence on how *much* returns and news stimulate attention. In this way, we hope to provide some parameters for the next generation of models that explore the effects of attention on stock prices. We find that both returns and news have a strong effect on investor attention, but the effect of abnormal volume is modest.

To conclude, we examine the effect of abnormal attention on future returns. We find that future month $t + 1$ returns are negatively related to attention, particularly for the highest attention portfolio. When we examine this portfolio, we find that current returns are particularly high, suggesting that the future returns are likely a reversal from prices that were driven away from fundamentals due to the demand generated by the high level of investor attention. We view these results as consistent with the model of Hendershott et al. (2021). The unpredictable nature of attention that we show in the data makes the market makers inventory problem particularly difficult. When the episodic liquidity demands in

Hendershott et al. (2021) are unpredictable, as they are in our attention data, the large demand arising from high attention can temporarily distort prices and create pricing errors. The fact that our month $t + 1$ returns appear to be reversals from prior-month attention driven price effects, suggests that the uncertain nature of investor attention is an important factor in the development of the pricing errors found in Hendershott et al. (2021).

2 A Framework for aggregate attention

2.1 Aggregate Attention

Most of the attention-based literature in economics uses an attention-augmented decision utility of the form $Max_a\ U(a, x, m)$, where a is a particular action to take, x is a signal of the true value, which can be multidimensional such as when a purchase has several quality dimensions, and m is an attention parameter. In this literature, m is usually parameterized [0, 1], from no attention to full attention. But this parametrization does not really suit our focus on aggregate attention. Economists usually confront the attention problem as one of inattention to a signal, wherein if an actor is paying full attention to the signal, the variance of the signal would approach zero. Therefore, where there are no limitations on attention, the agent's optimal action would be the fully informed action where $m = 1$ (Gabaix and Laibson 2006, Chetty, Looney and Kroft 2009, and Gabiax 2014). A degree of inattention to a signal yields a suboptimal attention parameter, $m < 1$, and the inattentive actor would make a suboptimal decision. Gabaix (2018) states the case that many of the problems found in behavioral economics can be framed in this inattention framework, including inattention to taxes, nominal price illusion, hyperbolic discounting, and most interesting to a finance audience, overreaction and underreaction.

A common feature of this literature is an inattentive ($m = 0$) default. One application of this framework is Greenwood and Shleifer (2014), where overreaction and underreaction are caused by an investor having to pay attention to a large number of AR(1) processes, say stock prices, or interest rates, and the inability to identify the true process in each case can lead investors to anchor on the average autocorrelation, used as the inattentive default. Investors thus, incorrectly evaluate the autocorrelation of a specific price process, and underreaction and overreaction follow directly.

How much attention do investors pay to stocks? How much attention *should* they pay to stocks? To make observations on the cross-sectional patterns of aggregate attention meaningful, we need some benchmark for rational investor behavior. When one sets out to develop such a model, the researcher is presented with a large number factors that likely influence how much attention an investor pays to the stock market in general, and to particular stocks. Some of these factors could be (i) their aggregate wealth in the market, (ii) the ability of particular stocks to generate positive or negative alphas, or (iii) their opportunity cost of paying attention to the market. Using this limited set of proposed factors, we produce a general attention function as:

$$a_{i,j} = f(W_{i,}, \alpha_j, c_i), \tag{1}$$

where $a_{i,j}$ is the attention of investor i in stock j, α_j, is stock $j's$ potential outperformance or underperformance, W_i is the investor's total wealth in the stock market, and c_i is investor $i's$ opportunity cost of paying attention to the market. Some of these variables are notoriously hard to measure, but we can make significant progress towards an aggregate attention benchmark if we assume only that one of the partials, $f'_j(w_{ij}) > 0$, where w_{ij} enters Equation (1) from the well-known definition of an investor's portfolio wealth:

$$W_i = \sum_{j=1}^{N} v_{i,j} \tag{2}$$

where, for N stocks, $v_{i,j}$ is the value of investor $i's$ wealth in the stock j, and $w_{i,j}$ is then equal to $v_{i,j}$ normalized by portfolio wealth, W_i. Given our partial derivative assumption, investors will pay proportionally more attention to stocks that represent a higher proportion of their invested wealth.

We do not know what each individual investor's holdings, $w_{i,j}$ are in a particular asset, so we make no predictions about individual attention, but we can make a reasonable conjecture about stocks in aggregate. Fortunately, we have an easy way to calculate benchmark for the proportion of aggregate wealth in a security in the market capitalization weight. Since aggregate wealth in the market is the sum of all individual positions, aggregate wealth in a stock is equal to the sum of all individual investments, or w_j after normalizing by total wealth. Given our partial derivative assumption, $f_j'(w_{ij}) > 0$, the greater the aggregate wealth of all investors in stock j, the more attention stock j should receive.

We do not know investors' attention capacity, their wealth in the market, or their opportunity costs of attention, so we can't say anything about the total amount of attention they spend contemplating their portfolios. Several well-off investors that I know personally invest in index mutual funds, primarily through retirement accounts and pay very little attention to the fluctuations of the market, but these index investors will be dependent on investors that pay attention to keep prices relatively efficient and their index strategy a reasonable one. We can use the market cap weight benchmark to focus on the relation between the proportion of attention allocated to a particular stock and its market cap weight. Using this idea as our null hypothesis: Attention to a particular stock should be proportional to its relative importance to all investors, the latter measured by its market cap weight. If Apple is ten times the market cap of Boeing, then our null hypothesis is that investors should pay ten times as much attention to Apple relative to Boeing. Formally, in aggregate investors should pay attention to the stocks in the market in proportion to their representative market capitalization. Under this null, attention should be proportional to:

$$a_j :: w_j \tag{3}$$

where w_j is the market cap weight of a particular stock j, and a_j is the relative amount of aggregate attention received by stock j.

Clearly, investors could pay too much attention to stock j, whether they are attracted by news, past returns, or abnormal volume as in Barber and Odean (2008), or some particular attention-grabbing feature of the firm's operations.[2] But if they do, then the attention they pay to some other stock k will be deficient. Hence the parameter a_j can be greater or less than w_j, the market cap weight. Most investors will only pay attention to a subset of stocks (Merton, 1987), so a_j will clearly be less than w_j for the stocks they are not aware of. Since most, if not all, investors are likely to hold different portfolios, all investors are likely to pay too much attention to certain stocks, and little or no attention to other stocks. But how do these different piecemeal sets of attention allocation aggregate across all investors? Answering this question is the basis of the first set of empirical tests in the paper, and to our knowledge the first attempt to identify investors' aggregate attention function.

To examine aggregate attention, we need some compact measure to indicate whether the aggregate set of attention across all investors is rational, at least as under the assumption that aggregate attention proportional to market cap weight is rational. To do this, we rely on a power law function, $Y = kX^{-\zeta}$, where the exponent, ζ is referred to as the power law exponent. The power law relation can tell us whether the aggregate level of attention coincides with the aggregate distribution of market capitalization. The particulars on power law estimation follow immediately below. When we examine these power law distributions, we see there are many cases where the power law coefficients of attention are quite similar to the power law coefficients of market cap. However, when we dig into the data, we find that the overlap of stocks within the top n market cap stocks and the top n attention stocks is much less than 100 percent. We will refer to these differences as *Distraction*, defined as:

$$Distraction_j = w_j - a_j. \tag{4}$$

Because we know little about the investors attention-augmented utility function, we cannot say much about how much effort should an investor spend on allocating attention to stocks. Although some would argue 'none', Grossman and Stiglitz (1980) show that 'none'

[2] By operations, we envision firm's who began a presence on this internet in the 1997-2000 internet bubble, or electric vehicles and Bitcoin today.

is not an equilibrium for all investors. As fewer investors pay attention to stock prices, the potential benefits of attention increase for those whose opportunity cost of attention is lower, or whose cost of acquiring information is smaller. Observationally, many investors pay close attention to stock prices, and others prefer indexing with little attention paid to individual stocks. We are interested in how these different levels of investor attention aggregate across all stocks. To put some economic content into our empirical analysis, we need to think about what the levels of aggregate attention should look like across stocks. We use aggregate wealth as our benchmark. Since aggregate wealth in a particular stock relative to other stocks is represented by w_j, the *Distraction* measure serves as our null. Under this null hypothesis, aggregate attention should be proportional to investors aggregate wealth in a particular stock. Empirically, calculating the weights in a_j and w_j, is feasible using market cap weights as the null for the empirical attention weights, a_j, but it is more convenient to initially examine the attention rank of a stock against the market cap weight rank. Using ranks maintains the ordinal ranking of market cap weight percentages, but has the advantage of being easily translatable into a power law, a convenient function for assessing concentration across a large set of stocks. We construct *Distraction* as market cap rank minus attention rank so that under the null: $Distraction = 0$, where the attention rank of a stock is precisely equal to its market cap weight. A positive level of *Distraction* represents an overweighting of attention under the null, and a negative level of *Distraction* represents an underweighting of attention.

2.2 Power Laws

Power laws are simple exponential models that have recently adapted to economic situations. For example, Gabaix (2011) estimates the power law of firm sales to GDP to argue that idiosyncratic shocks to important firms can have widespread effects in the economy. Goldstein et al. (2009) find that the degree of broker concentration for institutional order flow is easily expressed as a power law. Blakrishnan, Miller and Shanker (2008) examine the distribution of daily stock volume using a power law. Saglam, Moallemi, and Sotiropoulos

(2019) apply a fractional exponent power law process to calculate price impact costs of institutional trades. Gabaix (2009, 2016) discusses many examples of power laws applied to economics and finance including, city size, number of firm employees, income, wealth, and CEO compensation.

A power law is a simple mathematical relation between two variables:

$$Y = kX^{-\zeta} \tag{5}$$

Where k is a constant, that is often separated and empirically less interesting. After taking the log of the equation. ζ is referred to as the power law coefficient, and is negative by construction since rank declines as total counts, be it population, employees, wealth, or attention, falls. However, the power law coefficient is often discussed in absolute terms, such as a higher (absolute) Gini coefficient denotes a greater degree of wealth inequality. The power law coefficient can then be estimated as a linear equation usually by using the rank of a particular firm as the Y variable, and the raw number be it firm size, wealth, or in this case, the number of Stocktwits posts in a particular period. Taking logs and using our attention variable as the independent variable produces the linear equation:

$$\ln(Rank_{i,t}) = k - \zeta \ln(Posts_{i,t}) + \varepsilon \tag{6}$$

where $Rank_{i,t}$ is the rank of a particular stock i in terms of posts in month t, and $Posts_{i,t}$ is the number of posts about the stock in that month. The power law coefficient, ζ, is the coefficient of interest, it describes how concentrated is the distribution of attention across stocks. As we move down in rank from most popular ($Rank = 1$) to the second most popular stock ($Rank = 2$), the raw number of posts will drop off at a speed determined by ζ. In an analogous manner, we can examine the power laws of market capitalization and trading volume. A higher ζ means a higher degree of inequality in the distribution, therefore if $\zeta(attention) > \ \zeta(market\ cap)$, we can conclude that investor's attention is too concentrated in just a few stocks relative to what a rational distribution of attention should be given the aggregate level of investment in the stock market. Our tests are not limited in scale, so we can test how the power law coefficient, and the relation between attention and market capitalization varies across different numbers of ranked stocks. Specifically, over the Top 50, 100, 200, 500 and 1,000 stocks for each variable. Typically, the regressions have extremely high R^2 values, indicating that the simple power law model does a good job of explaining relative concentration across a large group of stocks.

2.2.1 Volume

Motivated by papers that link attention to trading volume, we also examine the power law coefficient for monthly trading volume. Barber and Odean (2008) show the link between attention grabbing events and order imbalance. Hendershott, et al. (2021) derive a model where investors are inattentive in such a way that makes their liquidity demands difficult for liquidity providers to predict. The authors contend that these liquidity shocks are associated with significant pricing errors. As a first step, we will look at how well the power law for attention tracks the power law for volume. It will also be important to see how predictable from period to period is the set of high attention and high volume stocks. Even if the power law coefficients between attention and volume are similar, what stocks investors are paying attention to is key. If the set of stocks that investors pay attention to varies considerably from month to month, liquidity providers will have a difficult time predicting where liquidity shocks occur, and pricing errors can result. In this way, we are testing whether investor behavior is consistent with the liquidity provider problem in Hendershott, et al. (2021).

3 Data

The Stocktwits data set for investor attention contains over 76 million posts covering 75 months from January, 2011 through March, 2017. The data consists of a series of posts identified by a '$SYMBOL' hashtag as pertaining to a particular stock. Since 2008, Stocktwits number of posts has been growing rapidly. However, data before 2011 is sparser and does not cover as large a universe of stocks every month. The data is obtained directly from Stocktwits API, but at present the data set cannot be extended to more recent months because Stocktwits has restricted research access to their API. All posts in the sample have a single hashtagged stock as the subject. Occasionally, investors also post about indices, currencies, or commodities, but we restrict the sample to single stock posts with share codes less than 30. The bulk of these posts cover share codes 10 and 11.

Other similar data sets have been accessed and are reasonable substitutes for Stocktwits. However, due to the volume of data, existing studies are sometimes limited in scope or scale. For example, Bordino et al., (2012) study the relation between daily volume and Google search queries, but only cover the NASDAQ 100 stocks over a single year. Four years of Twitter activity is used by Rakowski, Shirley, and Stark (2020) to examine return and volume predictability. Stocktwits has been used as a data source by Giannini, Irvine, and Shu (2018), Giannini et al. (2019), Cookson and Neissner (2020), Cai, Yung, and Zhu (2019), Cookson, Engelberg, and Mullins (2020).

It seems unlikely that we could present a convincing analysis of aggregate attention without controlling for the presence of news. Therefore, we obtain a news data sample comes from Ravenpack. We only include articles with relevance score equal to 100. The relevance score is a Ravenpack provided confidence score to indicate how certain their algorithm is that an article is really about a specific stock. Requiring a relevance score of 100 is a standard filter (Gao, Parsons and Shen, 2018). For each stock we calculate the number of articles each month from three sources: the Dow Jones Newswire, PR Newswire, and Web edition, a heading that includes major publishers, government and regulatory agencies, and local and regional newspapers. We also record a news sentiment score. Although also a Ravenpack proprietary algorithm that is unfortunately opaque, the news sentiment measure has been used effectively by Hendershott, Livdan, and Schurhoff (2015). The sentiment score is bounded between -1 and 1 based on Ravenpack's Event Sentiment Score.

Data on returns and volume comes from the CRSP monthly database. Stock financial information and earnings report date data are taken from the CRSP-Compustat merged database. Finally, data on analyst coverage is collected from the IBES database.

3.1 Sample

Panel A of Table 1 presents aggregate statistics on the Stocktwits post sample. Total posts and number of stocks are yearly totals, while average posts per stock and the maximum number of posts per stock are monthly averages. The sample is growing rapidly from around 55,000 post per month in 2011 to 2.6 million posts per month in 2017. This growth in activity is reflected in the average number of posts per month, which rises from 32.9 in 2011 to 768.2 in 2017. The stock universe covered is fairly broad throughout the sample period, with a minimum number of 3,655 stocks in 2011, and a maximum of 4,731 in 2015.

Panel B of Table 1 presents similar summary statistics for the news article sample. The total number of recorded articles has also been growing, but more modestly, from 56,000 articles per month in 2011 to 126,000 articles per month in 2017. The average number of articles per stock does not show the same growth as the Stocktwits sample as it is fairly stable in the range of 33 to 39 articles per month. With the total number of stocks covered in rough alignment to the pattern of the Stocktwits sample, the growth in the number of articles reflects a more balanced pattern, with more stocks receiving some attention, so that the overall distribution is less skewed over time.

To get an idea of the specifics of the post and news data, Table 2 presents the 40 individual stocks that were most often the Top 20 attention gathering stocks in a particular month. Frequency is the number of times the stock was in the Top 20 most mentioned stocks, 75 being the maximum. Stocktwits rank, Market Cap rank and Volume rank, are the average ranks for posts, market cap, and shares traded when the stocks made the Top 20. Apple (AAPL) is the most mentioned stock, being in the Top 20 every month, and gathers the most posts, with an average Stocktwits rank of only 1.32. Amazon and Google (Alphabet) were also in the Top 20 every month. Facebook (FB) is in the Top 20, 59 times, especially impressive since they were not public for the entire sample period. Typically, stocks that are often in the Top 20 category have similar ranks for either size or volume, but not always. Certain smaller capitalization stocks, like PLUG (Plug Power), ZNGA (Zynga) or GPRO (GoPro) sometimes capture an outsize level of attention. Large financials like BAC (Bank of America), GS (Goldman Sachs), and JMP (JP Morgan) are also prominent. Surprisingly, stocks that are perennial losers like BBRY (Blackberry) or JCP (JC Penney) also make the list, showing that investors have a consistent interest in these underperforming stocks (Odean, 1999), perhaps foreshadowing the Gamestop episode. Most volume ranks are within a reasonable range of the Stocktwits rank considering that there are often upwards of 4,000 stocks in the sample. Several high priced stocks, such as LNKD (LinkedIn) and PCLN (Priceline) likely are under ranked on our shares traded metric relative to what their ranks would be under a dollar volume metric. The table also reflects a heavy sprinkling of technology stocks.

Similar data is presented in Table 3 for the stocks that were most often in the Top 20 news articles in a month. This list is a bit more predictable than the posts list in Table 2. Most of the stocks are large cap stocks. AAPL again leads this list and is the most covered stock by news reports. In this list several large financials, automakers, and Dow Jones stocks, such as GM (General Motors), T (AT&T), DB (DeutscheBank), and MS (Morgan Stanley) are heavily reported on, but have more moderate attention ranks. The only stocks that are relatively low on all ranks other than news, yet still are heavily reported, are the financials DB, and RJF (Raymond James Financial). FCAU was the symbol of Fiat Chrysler, reflecting the fact that the automobile industry tends to have more news articles than Stocktwits interest. Another financial RY (Royal Bank) makes this news list, but has the lowest Stocktwits rank and the second lowest volume rank. The news appears slanted, at least relative to investor attention, towards Dow stocks, financials, and automakers.

4 Results

4.1 The power law of attention

We examine the concentration of attention across stocks by estimating the power law coefficient using Equation (6). First, we sort stocks every month by the number of posts, so that the stock with the highest number of posts has $Rank = 1$, and continue until all stocks with post activity in a month have been ranked. We then take logs and estimate Equation (6) for 5 different sets of stocks from 50 stocks to 1,000 stocks.[3] Figure 1 plots the distribution of the power law coefficient, ζ, over 75 months of estimation. For clarity, Figure 1 presents pairwise distributions of 100, 200, 500, and 1,000 stocks. A more negative ζ coefficient indicates a distribution that has higher concentration towards the top end of the relevant set of stocks.

Panel A of Figure 1 reports the power law coefficient distributions of attention and size (market capitalization). We first look at the pairwise distributions of the Top 100 stocks and find a notable pattern. The two distributions overlap considerably, and the means of the two different distributions are quite close (-1.72 and -1.65). What this means is that, for the Top 100 stocks, investors are allocating their attention in much the same way as the market allocates market cap weights. There is no ex ante reason why these two patterns should overlap so much. In fact, anecdotally many suggest that investors are much too highly concentrated in attention-grabbing stocks, but this anecdotal impression is incorrect. For the Top 100 stocks, investors appear to allocate their attention rationally, at least under the null of Section 2.

The results change a bit as we add more stocks to the power law estimation. For 200 stocks, the mean of the attention power law distributions has crept away from the power law coefficient of market cap. Although both averages have fallen, there is a larger gap between the two (-1.61 and -1.41). This finding indicates that investors are paying too much attention

[3] We add 0.5 to the rank before taking logs as suggested by Gabaix (2016).

to the top stocks, relative to the null. As we increase the number of stocks analyzed to 500 and then 1,000, the distributions appear to separate more and more, and the distribution of market cap power law coefficients is more compact than the diffuse distribution of attention power law coefficients. As we analyze more stocks, investor pay relatively too much attention to the most important stocks, relative to their market cap weights. If these results are a proxy for how humans pay attention across a large set, the results show behavioral support for the neglected firm effect (Arbel and Strebel, 1982). We do not appear to pay enough attention to the lower market cap weight stocks.

For comparison purposes, Panel B presents the power law distributions of attention and trading volume. For all sets of stocks the distributions overlap considerably, and the means are reasonably close together. However, when we examine 1,000 stocks, the distributions begin to separate with volume even more concentrated than attention. Nevertheless, we conclude that, at least in aggregate, attention and volume tend to have the same concentration. Whether they are concentrated on the same set of stocks is a question we will address below.

Finally, Panel C compares the distributions of size and volume. The pattern is similar to that of Panel A, but somewhat more pronounced. In general, volume is more concentrated than market cap, and by the time we examine the 1,000 stock distributions, the distributions are completely separate. As a rule, volume is more concentrated in the most active names than market capitalization, and tends to track attention much more closely than market cap.

This tracking pattern is easily seen in Figure 2, which presents the power law coefficient results across time for sets of 100, 500, and 1,000 stocks. In the 100 stock group the three power law coefficient distributions tend to track each other. The overall concentration for market cap stays pretty consistent, and while attention and volume coefficients track market cap quite well for the first half of the sample period, there appears to be an increase in concentration of both attention and volume in the second half of the sample period. For the 500 and 1,000 stock groups, the distribution of market cap coefficients is consistently lower than volume and attention, which tend to track each other quite closely. Notably, the attention coefficients are more volatile indicating that the level of attention concentration across stocks is less predictable. The increase in concentration for attention and volume halfway through the sample period that seems to be a feature of the Top 100 graph, is less apparent in the Top 500 and Top 1,000 graphs, although a slight increase in concentration for these measures could be inferred.

4.1.1 Inside the power law distributions

The evidence we have seen so far indicates that investors allocate their attention in a reasonable approximation of rational allocation in proportion to market cap weights. Table 4 presents the summary statistics of the power law regressions. We see that volume in general is the most concentrated, regardless of how we define a set of stocks. This means they have the most negative ζ coefficients. But this is not a universal rule since market cap is actually the most concentrated in the set of 50 stocks. All estimated coefficients decline as we add more stocks to the test set. This finding indicates that both attention and volume tend to migrate towards a distribution that tracks the market capitalization distribution, but as we saw in Figures 1 and 2, attention and volume tend to remain a little more concentrated at the top than the market capitalization distribution. Market cap is the only variable that approaches, and actually reaches the Zipf's law coefficient of 1.0. A Zipf's law coefficient of 1.0 is found in such diverse data as word frequency and city size, and implies that the nth largest stock has a market cap that is $\frac{1}{n}$ the market cap of the largest stock.

However, when we dig deeper into the data we find patterns that suggest the relatively neat and rational results from the power law distributions conceal a good deal of month to month volatility. Panel A of Table 5 reports month-to-month own correlations for our three measures. $Frequency$ represents the average number of stocks in a particular sample in month t, that are also in the same sample in month $t+1$. $Percent$ reports the $Frequency$ as a percentage for easy comparison across stock groups. The data shows that the distribution of market cap is very stable; about 97% of the stocks that are in a Top group in one month, are also in the group the next month. The *Rank correlation* is a Spearmann correlation of the overall ranks between months t, and $t+1$, and for market cap is very close to 1.0. For market cap, these very high correlations indicate that, even when a stock falls out of the Top 50 or Top 100, it is replaced by a similarly ranked stock, since the rank of the replacement stock cannot be that far out of the Top group for the rank correlation to remain so high. We directly examine this contention in Panel B that shows the average month t rank of the Top n stocks in month $t+1$. This Panel shows that for size, time t rank of the overall rank of the $t+1$ Top n stocks is very close to the minimum. For example, if the Top 50 stocks remained exactly the same from month t to month $t+1$, the mean Replacement rank would equal 25.5. The actual mean Replacement rank, and the relative rank expressed as a percentage, shows that the replacement stocks for market cap, were already quite highly ranked in month t. In other words, it would be easy to predict the stocks likely to move into a Top n market cap group by examining the stocks that just missed the cutoff for that group in the previous month.

Attention shows a more volatile pattern. The *Frequency* of stocks in a particular group from month to month ranges from 61 to 65 percent. Even when we extend the sample to 1,000 stocks there is considerable turnover in the stocks that attract the most attention. The rank correlation for attention is much lower than the market cap sample, indicates that much of this turnover comes from stocks well outside the boundary of a particular group. We prove this contention for attention in Panel B that shows that the stocks replacing the Top n stocks in month $t+1$ are difficult to predict from their month t attention rank. For example the Top 50 stocks in month $t+1$ had a mean attention rank in month t of 101.86, considerably above the minimum of 25.5.

Volume settles in the middle of the range between size and attention. About 80 to 87 percent of the high-volume stocks are represented in the same category from month to month. The rank correlation and the replacement rank statistics also suggest that the stocks that replace the month t stocks in month $t+1$ come from outside the month t distribution, but not too far outside it. The risk to market stability and efficient pricing comes when a smaller stock attracts attention that also generates volume. This volatility generates unpredictable liquidity demands, which can affect stock prices, (Hendershott, et al. 2021). Alternatively, since the stocks that attract the most volume have a more stable distribution than the stocks that attract the most attention, not all attention-grabbing stocks generate outsize volume demands. When there is no associated liquidity demand, the unusual amount of attention is benign with respect to the markets. Just how frequent each type of attention-grabbing event will be, is assessed in Panel C.

Panel C present the cross-sectional correlations in a particular month t. These cross-sectional distributions answer questions such as how many of the Top 50 market cap stocks are also in the Top 50 attention stocks? The correlations are done in a pairwise fashion, as in Figure 1. There is not a good deal of overlap between the size and attention groups. Together with the results in Figure 1, these numbers indicate that attention, which appears rational from an average power law coefficient perspective, is actually not rational because it includes so few of the Top market cap stocks. In this Panel, the size of the sample makes a considerable difference; only 36 percent of Top 50 market cap stocks are also in the Top 50 attention stocks, and the rank correlation between size and attention is extremely low at 0.248.

The attention-volume comparison also shows only modest cross-correlation, between 36% and 62% of the stocks in the high attention group are also in the high volume group. Roughly one-third to two-thirds of the attention shocks also generate a volume shock. Combined with the relatively low cross-correlations between volume and size these results imply that when an attention shock does generate a volume shock, it will be extremely difficult to predict what stocks will be hit with liquidity shocks. Thus, the focus on inattention in Hendershott et al. (2021) is important when we look at actual investor attention data. It is very hard for the market maker to predict what stocks will receive the largest liquidity demands from month to month. Thus, unpredictable attention presents a difficult inventory management problem for the market maker. As attention, and thus potential liquidity demands, cannot be well predicted, the result is temporary price pressure that moves prices away from fundamentals. The rank correlations for all the comparisons in this Panel are quite low; all below 0.40. This finding suggests that any attention-driven liquidity demands are not very likely to come from stocks that are just outside the set of usual suspects, rather they will come from much lower ranked stocks that are neither large, nor trade heavily on a regular basis. The conclusion on the difficulty liquidity providers have in predicting where attention-related liquidity shocks will occur, and preparing with offsetting inventory, is apparent.[4]

4.1.2 Sample robustness

We contend that the Stocktwits sample is a valid data set for measuring aggregate attention since it covers many stocks and contains many different investor types.[5] However, despite the professionals on the site, it could be an be criticized as being too representative of retail investors. We examine this possible objection in two ways. First, in unreported tests, we separate market cap into institutional ownership, and retail ownership. If the Stocktwits sample tilts to retail, then the cross-correlations between attention and size reported in Table 5 Panel C, should be stronger for the stocks that are heavily-retailed owned, and weaker for the stocks that have high institutional ownership. In unreported tests, we calculate these cross-correlations separately for both measures of market cap, and find no significant differences. The institutional-retail ownership percentage of a stock does not appear to have a significant affect on our results.

We discuss in the conclusion some of the difficulties in using either Twitter or Google search volume to capture aggregate attention. However, thanks to the generosity of Loughran and McDonald (2017), we obtain their EDGAR search log data from 2007-2015, and use the search log data as an additional measure of investor attention. We investigate this dataset in detail in the Appendix, repeating a number of our tests using EDGAR search logs as a

[4] Table 7 examines how strongly abnormal attention is correlated with abnormal volume.

[5] Stocktwits posters can self-designate their experience ranging from novice to expert, as well as designate their most common investing style Cookson and Neisner (2020) investigate these different investing styles.

measure of attention. In brief, the EDGAR search data exhibits higher own correlation from month to month than the Stocktwits sample. EDGAR search logs also have much higher power law coefficients for many of our sets of Top stocks than Stocktwits or market cap power law coefficients. Yet, the cross-correlations between EDGAR search as a measure of attention, market cap, and volume are quite similar to those reported in Table 5 Panel C for the Stocktwits data. All of our main conclusions about aggregate attention are verified using the EDGAR search measure of attention. Attention is too concentrated relative to market cap weights, particular for more than the Top 100 stocks, indicating that EDGAR search also exhibits the neglected firm effect. EDGAR search attention is also more volatile month to month than either market cap or volume, consistent with the results for Stocktwits data. Finally, the same cross-sectional correlation patterns exist in EDGAR search that exist in the Stocktwits data, and these cross-correlations are not affected by the institutional or retail ownership level in a stock.

4.2 Returns and attention

We know that extreme returns captures investor attention, but what does the returns-attention function look like? This is an important question for modeling cascades and information bubbles. For example, Chinco's (2021) model relies on a return threshold to ignite the investor confidence necessary to begin a stock bubble; but he has to pick an arbitrary past return that convinces ordinary investors that the speculators could be correct. Modeling the returns-attention function will help us determine if there is some level of past return that ignites investor attention. Motivated by the attention and trading study of Barber and Odean (2008), we are particularly interested on the effect of news and abnormal volume, as well as returns, on attention. But a number of other explanatory variables suggest themselves as well.

4.2.1 A nonparametric exploration of returns and attention

We begin by examining how returns relate to different measures of attention including the absolute number of posts in a month, the log of the raw number of posts in a month, the Stocktwits activity rank of a firm, and *Distraction*, the difference between the market cap rank of a stock and the Stocktwits rank. Because we do not wish to specify the relation as linear, we run a non-parametric local regression (Cleveland, 1979) of the form:

$$Attention_t = g(Return_t) + e_t. \tag{7}$$

In this specification, g is a local linear or cubic regression function and e is a random error. For every observation $Return_{i.t}$, where i is a stock-month return, and t designates the particular month, the function $g(Return_t)$ is estimated using observations near $Return_{i.t}$ to form a local approximation. In a local regression, weighted least squares is used to fit functions of the predictors at the center of each local neighborhood near $Return_{i.t}$.[6] The result should be an approximation of the returns-attention relation without specifying a particular functional form on the data.

With over 180,000 stock-month observations, we chose to restrict the size of the output by showing four representative local regression plots using the period February, 2017 to March, 2017. Several other subperiods were examined in other years and other months, and the pattern is generally similar. One reason to restrict the plots to a shorter time frame is that the number of Stocktwits posts increases throughout the sample, and thus showing a graph of posts or log posts over the entire sample would be misleading.[7]

Panel A of Figure 3 presents the local regression plots of returns against the number of posts and the log of posts. The striking feature of these plots are that contemporaneous returns strongly influence posting activity. The first plot of the raw number of posts is the

[6] Data in each neighborhood is weighted by a decreasing function of its distance from the center of the neighborhood.

[7] Returns are winsorized at 1% and 99% (approximately +/- 40%) for clarity in these graphs.

more interesting. Returns between approximately -15% and +10% per month produce no particular effect on posting activity, as the total number of posts fluctuates slightly below its mean. However, there is a marked increase in slope, at a returns trigger point at about -25% and +15% where the amount of attention increases, and continues steeply upward as absolute returns continue to increase. The plot of the log of the number of posts is much smoother, showing the same general pattern but with less of a returns trigger point. The mean number of posts during these months is 1,185, a figure than can triple when returns reach more extreme levels. Returns not only capture attention, they capture a lot of attention relative to normal posting activity. We also see a fairly symmetric increase for both positive and negative returns, a finding consistent with Odean (1999) and Barber and Odean (2008) who propose that many contrarian investors focus on negative returns, just as optimistic investors focus on positive returns.

The plots in Panel B focus on rank, the first presenting Stocktwits rank, which gives some idea of how much a stock can move up in rank as a function of returns. The mean rank in the first plot is 1520, and stocks with returns around zero tend to be ranked lower than average. Rank increases rapidly and symmetrically as returns increase and decrease. The plot indicates that is possible for a typical firm to increase their attention rank by over 500 points if they have a month of particularly high or low returns. The final plot shows how returns are related to *Distraction*, the difference between Stocktwits rank and the market cap rank. In this graph the mean level of *Distraction* is greater than 0 (the null) at 485. Because some quite small firms market cap firms in the CRSP universe nevertheless receive attention, the mean is drawn away from 0, although the median is lower. The *Distraction* graph does show return effects that are large in magnitude. Small stocks with either high positive or high negative returns attract attention far out of proportion to their market cap weight.

4.3 News and attention

Panel C of Figure 3 presents two graphs outlining the relation between news and attention. The first graph plots the relation between the number of news articles and Stocktwits rank, the second graph plots the relation between the number of news articles and *Distraction*. These plots are less striking than our returns and attention plots, except for a kink below the mean number of articles (51 in this period), the number of news articles rises sharply until about Stocktwits rank 500, and them more gradually increases towards a rank of 1. The second graph shows that *Distraction* is strongly positive for just a few news articles, but as the number of news articles increases, the plot steeply reaching a minimum at about 100 news articles per month, and then rises gently in the number of articles. This kinked patterns are likely related to small firms that usually receive little or no news coverage in a month. When they do receive news coverage their attention rank increases sharply, but quickly dissipates as larger, higher ranked firms normally receive a steady volume of news articles, and the bulk of the graph reflects this relation.

4.4 Predicting attention

4.4.1 Explanatory variables

In this Section we estimate several panel data regressions for alternative measures of attention including the log of the number of posts, Stocktwits rank, and *Distraction*. The primary explanatory variables of interest are returns, news, and volume. For *Returns*, we split the returns sample into positive and negative to see whether the point estimates of the positive and negative slope differ, since the plots in Figure 3 look very symmetric, a closer examination is warranted. For *News* we use the total number of news articles in a month from our Ravenpack database (Tables 1 and 3). Volume is measured as abnormal volume, as in Barber and Odean (2008) and Da et al. (2011), we follow the latter example and calculate abnormal volume ($AbnVol$) relative to the average of the last three trading months prior to the observation month.

We also include a number of regressors that prior literature or investigation lead us to believe might influence the level of attention. First, we include *Advertise*, the advertising to sales ratio of the stock. We include this ratio because Grullon, Kanatas, and Weston (2004) find that advertising is positively related to the size of the shareholder base. In addition, Lou (2014) claims that firms use advertising specifically to attract retail investor attention. Lou's claims seem to contradict the findings in Da et al. (2011) who report that the advertising to sales ratio is often negatively related to individual investor trading reported through SEC Rule 11ac1-5 (Dash 5 reports). However, since Lou's (2014) claim seems tenable, we include the ratio here as a predictor of investor attention. We also include return threshold variables, that are dummy variables set to 1 if the contemporaneous stock return is greater than 20%, or less than -20%. Chinco (2021) motivates us to include these threshold coefficients, since he hypothesizes that surpassing certain return thresholds can trigger a jump in investor interest in a stock. We also include the book-to-market ratio (BtM) since Giannini et al. (2018) report that Stocktwits coverage tilts away from value firms. We include a dummy variable, *Announce*, that takes a value of 1 if earnings are announced in a particular stock-month. Following Da et al. (2011) and Giannini et al. (2018) we also include *Size*, the log of the market value of equity for a firm, the log of 1+ analyst coverage, *Coverage*, and idiosyncratic volatility, $Ivol$. We also include the news *Sentiment* variable from Ravenpack. Finally, we include dummy variables for *Tech*,the technology industry and *Pharm*, the drug industry, since observation leads us to believe that these type of stocks are investor favorites.[8]

Summary data on these variables is presented in Table 6. Firm size averages about $8.0B, news sentiment and abnormal volume are centered around 0, which they should be from construction. Idiosyncratic volatility is the standard deviation of the residuals from the market model, these reported daily averages gross up to 38 percent annually for the mean value and about 28 percent annually for the median value. Book-to-market averages 0.44, negative book values occur in almost 30 percent of the sample, and these observations are

[8] The dummy variable *Tech* is set to 1 when the stock's SIC codes first 3 digits are 737. Likewise, the *Pharm* dummy variable uses the 3-digit SIC code 283.

set equal to 0. The advertising-to-sales ratio averages 1.3%, with almost 75 percent of all firms reporting no expenditures on advertising. About 7.3 percent of the sample firms are *Tech* firms, and 8.9 percent are Pharmaceuticals.

4.4.2 Determinants of attention

Table 7 presents the results of Panel data estimation of attention levels. In Columns (1) and (2), the dependent variable is the log of the number of posts about the stock in a month. In Column (3), the dependent variable is Stocktwits rank, so that a negative coefficient represents a variable that moves the stock closer to the Top rank (= 1). Column (4) estimates the determinants of *Distraction*, where a positive *Distraction* represents a stock that captures proportionally more attention than its market cap rank would warrant. For reasons outlined below, *Distraction* is a skewed variable that tends to have a positive mean, so for analysis we standardize *Distraction* by transforming it into a standard normal variable. The standardized variable is almost perfectly correlated with raw *Distraction*, but has the advantage of having easy to interpret effects on future returns in Table 8.

The first two regressions measure total post activity, given the change in the number of posts over the sample period, these regressions use monthly fixed effects, and 'Huber-White' standard errors (Rogers, 1993). Larger stocks capture more attention, consistent with our null hypothesis that aggregate attention should follow market cap weight. Last months returns (*Lag Returns*) are positively related to attention, indicating a certain amount of trend following in attention. A biological explanation of trend following is provided by Anderson et al. (2016) who show that attention to activities that have produced rewards in the past is associated with a dopamine release caused by the positive feedback the attention to that stock has rewarded the individual with in the past. Their findings indicate that attention this month could be chemically rewarding if attention last month produced positive rewards, and is an novel explanation for trend following in an efficient market.[9] In Column

[9] The other nine authors on this paper are omitted from the text and references. Interested readers can refer to: http://dx.doi.org/10.1016/j.cub.2015.12.062

(1) positive returns do not have a significant effect on attention, though negative returns do. To explain the lack of reaction in positive returns, it is revealing to look at the return threshold coefficients in Column (2). Both positive and negative return threshold coefficients are significant, indicating that returns greater than twenty percent have a marked effect on the level of posts, in particular for positive returns. The impact of a threshold value of returns on attention is consistent with Chinco's (2021) assumption that there is a threshold in returns that tends to excite the interest of speculative investors.

News and news sentiment are both positively related to the number of posts, but abnormal volume is not. This latter result is surprising given Barber and Odean's (2008) conclusion that abnormal volume is a stronger influence than either news or returns on the trading order imbalance. Part of the explanation of this difference is given by Barber and Odean (2008) who note that their order imbalance activity measure is almost tautologically related to abnormal volume. Table 7 also includes more possible variables as influences on attention, which might capture some of the abnormal volume effect. Analyst coverage is strongly associated with attention, despite its high correlation with size.[10] Idiosyncratic volatility is associated with positive attention; investors pay attention to volatile stocks. Book-to-market is negatively associated with attention, confirming the finding in Giannini et al. (2018) that growth stocks tend to capture more attention. The coefficient on advertising is positive and significant indicating some support for the conjectures of Grullon et al. (2004) and Lou (2014). The earnings announcement month, not surprisingly, also tends to capture attention associated with this important information release. Both technology and pharmaceutical firms are positively associated with greater attention.

Column (3) uses Stocktwits rank as the dependent variable. Stocktwits rank is calculated each month, as is *Distraction*, so the growth in sample size that necessitated monthly fixed effects is not an issue in these specifications. However, Table 2 reveals that certain firms are highly ranked relative to their market cap on a regular basis. To capture any firm-

[10] The regression results are similar if we drop either of these two highly correlated variables from the specification.

specific attributes that are not controlled for with the set of included regressors, we include stock fixed effects in the specifications in Columns (3) and (4). The discussion of these results will focus on coefficient estimates that are different for Stocktwits rank than those of the log of total posts specifications. These differences begin with the fact that the returns threshold effect is even more clearly delineated in this specification. Both return variables are insignificant, but both threshold dummy variables are significant. This finding indicates that firms receive relatively more attention only when their returns reach a certain threshold. The abnormal volume coefficient is negative and significant in this specification, indicating a rank closer to 1, but the advertising coefficient has no significant effect on Stocktwits rank. Technology firms are associated with a rank closer to 1, but pharmaceutical firms are not.

Finally, Column (4) examines *Distraction*, the market cap rank less the Stocktwits rank. In this specification, the coefficient on market cap is negative and significant, which is a little odd given the dependent variable already controls for market cap rank. The explanation likely arises from the asymmetry arising from the fact that there are some small stocks that can, at times, capture a good deal of investor attention, whereas a large cap stock like Apple, cannot go up from the number 1 position in the Stocktwits rank, but could occasionally drop to number 2 or 3. Last month's returns are significantly associated with *Distraction*, and the returns results reflect the same threshold level of attention gathering that is represented in Column (3). News and Sentiment are significantly positively related to *Distraction*, as is abnormal volume. Generally, the variables that affect Stocktwits rank affect *Distraction*, not too surprising given that Stocktwits rank is used in the calculation of *Distraction*, and we already know from Table 5 that market capitalization rank is quite stable.

4.5 Distraction and future returns

We conclude our examination of aggregate attention with an exploration of whether *Distraction* affects future returns in month $t + 1$. Our motivation for this analysis comes from Ibbotson and Idzorek (2014), and Ibbotson, Idzorek, Kaplan and Xiong (2018) who hypothesize that many risk premiums can be related to the popularity of a stock, or specifically, the premiums are related to unpopularity. The economic motivation behind this hypothesis is that popularity drives demand and thus prices and future returns. In one sense, the theory is similar to Merton (1987) where investors only trade in the set of stocks of which they are aware. Ibbotson et al. (2018) use brand recognition, comparative advantage, and firm reputation as measures of popularity. Although *Distraction* may not capture comparative advantage, brand recognition and reputation should be strongly related to Stocktwits attention, making *Distraction* an excellent measure of popularity. Thus, the Stocktwits rank directly measures a stock's popularity with investors, and the variable is benchmarked against the null of market cap rank, so it has both a popular and an unpopular dimension.

In Table 8, we test the relation between *Distraction* and returns in a regression framework (Panel A), and a portfolio sort (Panel B). The regression uses month $t+1$ return as the dependent variable, and includes three different measures of *Distraction*. In Columns (1) and (4) we use the standardized *Distraction* itself, Columns (2) and (5) examine positive Distractions only, and Columns (3) and (6) examine negative Distractions only. We split the variable into separate analysis because the Ibbotson et al. (2014, 2018) theory states specifically that unpopular stocks will tend to have a risk premium, so we test the overall effects as well as the effects of unpopular stocks (*Distraction* $(-)$), and popular stocks (*Distraction* $(+)$) separately. We control for a stock's market cap and book-to-market ratio, since these are characteristics well known to be related to returns. Noting the trend following in the specification of Table 7, it is also important to control for the momentum effect, so we include a number of lagged returns including last month's return ($Ret1$), as well as other past months returns for months two to three ($Ret2-3$), four to six ($Ret4-6$), and seven through twelve ($Ret7-12$). To test for a popularity effect, we run monthly regressions over the 75 month sample period, and present Fama-MacBeth average coefficients and standard errors in Table 8.

Column (1) reports a significantly negative relation between *Distraction* and future returns. More popular stocks, relative to market cap weight, tend to do worse in month $t+1$. This is confirmed in the specifications in Columns (2) and (3). Positive *Distraction*, the relatively popular stocks earn lower returns, while negative *Distraction* stocks earn insignificant negative returns. These results, using a direct proxy for popularity, contradict the Ibbotson et al. (2014, 2018) popularity theory, wherein unpopular stocks should have a risk premium. In defense of the popularity theory, we only have 75 months of data, and only examine month $t+1$ returns. Ibbotson et al. (2014, 2018) also posit that unpopularity could account for the return effects associated with certain well-known factors like size and book-to-market, so the posited popularity effect could be associated with the inclusion of these variables. But exclusion of these variables, does not change the returns associated with *Distraction*, so we are left with a puzzle. Why do popular stocks earn a negative return premium, especially in the presence of past returns to control for momentum effects?

We have a number of extreme returns in the sample, monthly returns ranging from -93 percent to over +400 percent. To control for the possibility that these extreme outliers affect our conclusions about the popularity hypothesis, we winsorize the data at 1% and 99% (<-35% or >44%) and rerun our regressions. These results are presented in Columns (4)-(6). Removal of the extreme outliers from the returns distribution, increases the t-statistics on past returns, but does not remove the negative premium associated with popular *Distraction* stocks. Stocks that receive aggregate attention greater than their market cap weight percentage, earn lower future returns in the upcoming month.

Panel B of Table 8, presents a portfolio sort of *Distraction* and month $t+1$ returns. To construct this Table, we sort *Distraction* into deciles and examine the next month's return in each decile. The portfolio sort reveals that the popularity effect is not linear in the value of *Distraction*. Instead, there is almost no difference in monthly return across the five lowest deciles, thereafter, average monthly return drops slightly in deciles 6, 7, 8, and 9, but drops precipitously in decile 10. The negative relation between *Distraction* and returns is concentrated in the most popular stocks.

This simple portfolio sort reveals that month t returns in decile 10 are considerably higher than those in any other decile, and this fact provides an explanation for the low future returns in this decile. It is likely that high attention firms in this decile, (average standardized *Distraction* is 1.96 in this decile), and the unpredictable nature of attention, cause the type of difficult to predict liquidity shocks outlined in the Hendershott et al. (2021) model. These liquidity shocks generate the temporary mispricing found in that paper, a mispricing effect that mean reverts in the following month.

We expand on our conjecture of attention-driven reversals in Panel C that double-sorts Carhart (1997) 4-factor alphas first by the correlation between volume rank and attention rank, and then on *Distraction*. Correlation is measured as the correlation between the monthly volume rank of stock and its contemporaneous-month attention rank estimated over a past 12-month rolling window. The past correlation of attention and volume is created as an instrument for tendency of attention to generate volume in month t. Panel C reveals that the higher *Distraction* (Popularity of a stock) and the more positive the past correlation between attention and volume, the lower are month $t + 1$ alphas. For example, stocks in the most popular quintile that are also in the quintile with the highest correlation between past attention and volume have month $t + 1$ alphas of -1.62 percent. In row 1 where attention and volume are negatively correlated, the amount of attention a stock receives has an insignificant affect on prices. The difference between the highest *Distraction* stocks (Popular) and the lowest *Distraction* stocks (Neglected) produce alphas of only -0.28 percent in month $t + 1$. Conversely, in row 5, the most positive correlation quintile, the difference between the highest *Distraction* stocks and the lowest *Distraction* stocks have month $t + 1$ alphas of -1.80 percent. The difference in spreads across correlation quintiles supports our earlier speculation that attention that does not generate a volume shock has a benign effect on stock prices, but attention that does generate a volume shock significantly distorts month t stock prices, a price distortion that mean reverts in month $t + 1$, producing significant negative returns in that month.

There is one more noteworthy finding in Panel B of Table 8. Except for portfolio 1, month t returns are higher, sometimes considerably higher, than month $t + 1$ returns. Since there are 74 months of overlapping returns in the sample, and only one month of independent future returns, this finding would be near impossible in a complete data set. However, the Stocktwits data is a censored data set because if there is no posting activity in a month, the stock will not be in the data set. Although we present data in Tables 2 and 3 on the most often mentioned stocks, there must be some fraction of stocks that only receive a mention when their returns are noteworthy. Contemporary returns can only be consistently higher than future returns if returns play a significant factor in whether some infrequently mentioned stocks receive *any* attention. The final lesson on Panel B is that we should be careful not to underestimate the effect of high returns on investor attention.

5 Conclusion

We present an analysis of a large data set on aggregate investor attention. The paper builds a framework for understanding attention that states, in aggregate investors should allocate their attention across stocks consistent with their wealth across stocks, so that aggregate, attention should be proportional to market cap weight. We first examine whether this is true by estimating the power laws for attention, market cap, and trading volume across different sets of stocks. We find that aggregate attention looks rational, since it closely aligns with market cap weight for the most active sets of stocks. However, as we expand the universe of stocks, investors tend to possess a degree of neglect for smaller capitalization stocks relative to their market cap weight.

When we investigate the components of different sets of stocks, we find that this apparent rationality is ephemeral, as investor attention only covers between 36 and 53 percent of the individual stocks they should be paying attention to under the null. Further, the attention portfolios exhibit significantly more month-to-month turnover than the market cap portfolios. This makes investor attention, and the large liquidity demands that attention can generate, unpredictable for liquidity providers, who are unable to predict inventory demand when attention is so volatile. We interpret these findings as supporting the pricing error results of Hendershott et al. (2021) who predict that inattentive investors cause unpredictable liquidity shocks that can generate these pricing errors. High-levels of investor attention have an unpredictable component that has pricing implications. We find support for this conjecture in the data since high attention portfolios predict negative future returns, but have high current returns. These return patterns suggest that unpredictable levels of attention generate liquidity shocks that market makers are not prepared for, and pricing errors occur in high attention months. These pricing errors are subsequently corrected in the following month.

This paper presents the first large scale collection of facts on investor attention in the stock market. We hope that these facts will guide the burgeoning research literature on attention. In actuality, we present a number of facts that are consistent with assumptions already theoretically proposed in the literature, including the relation of analyst coverage to attention (Atiglan et al. 2020), and a return threshold affect of attention (Chinco, 2021). On the other hand, proposed effects from variables such as abnormal volume (Barber and Odean, 2008) and advertising (Lou, 2014), have only modest affects on investor attention.

One of the potentially important findings in the paper is that aggregate attention reflects a behavioral tendency to concentrate too much on large cap stocks, generating the neglected firm effect. Small market cap weight firms, get less attention than they should under the null hypothesis. It would be important to confirm this effect using other measures of attention. One confounding issue is that some people may post on Stocktwits not only to share information with the crowd, but also to receive recognition of their opinions or efforts. Lightly followed stocks could discourage posters since feedback would be relatively rare. On the other hand the recent stream in GameStop (GME) during the wallstreetbets episode scrolled too fast for human comprehension of all the posts, and specific feedback was rare. One way to tell if there is a reward/recognition effect clustering attention in highly-followed stocks, would be to look at similar attention measures and see if they also exhibit similar concentration patterns. Unfortunately, Twitter, while accessible, would likely have an even worse recognition problem. Google Search Volume has potential, since it is unlikely that investors search for ticker symbols as a way of garnering recognition. Unfortunately, Google Trends provides a measure of popularity that is calculated relative to a stock's own historical search volume. Specifically, a percentage popularity relative to a stock's all time high number of searches over the relevant time period. This service unfortunately provides no raw data. However, it might be possible to get a crude measure of popularity by running comparative searches relative to a numeraire stock, such as Apple. The limitation with this method is that the relative popularity measures are imprecise (2 digit comparison), and most smaller stocks relative to Apple may not produced meaningfully differentiated statistics. We resolve these difficulties by using aggregate Edgar search log data as an alternative measure of aggregate attention, and outline the results in the Appendix.

References

[1] Anderson et al. (2016) The role of dopamine in value-based attention orienting, *Critical Biology,* 26, (4), 550-555.

[2] Arbel, A. and P. Strebel, 1982. The neglected and small firm effects, *The Financial Review,* 17(4), 201-218.

[3] Atiglan, Y., T. Bali, K. O. Demirtas, and A. D. Gunaydin, 2020. Left-tail momentum, underreaction to bad news, costly arbitrage, and equity returns, *Journal of Financial Economics,* 135 (3), 725-753.

[4] Ben-Rephael, A., Z. Da, and R. Israelsen, (2017). It depends where you search: Institutional investor attention and underreaction to news, *Review of Financial Studies,* 30 (9), 3009-3047.

[5] Balakrishnan, P.V., J. Miller, and S. Shankar, 2008. Power law and evolutionary trends in stock markets, *Economics Letters,* 98, 194-200.

[6] Barber, B., and T. Odean, 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies*, 21 (2), 785-818.

[7] Bogousslavsky, V., 2016. Infrequent rebalancing, return autocorrelation and seasonality, *Journal of Finance,* 71 (6), 2967-3006.

[8] Bordino, B., Battison, S., Caldarelli, G., Cristelli, M, Ukkonned, A., and I. Weber. 2012. Web search queries can predict stock market volumes, *PLoS ONE*, 7 (7).

[9] Cai, Q., K. Yung and Z. Zhu, 2019. Sentiment, attention and earnings momentum, Working paper, Old Dominion University.

[10] Carhart, M. 1997. On persistence in mutual fund performance, *Journal of Finance,* 52 (1), 57-82.

[11] Chetty, R, A. Looney, and K. Kroft, 2009. Salience and taxation: Theory and evidence, NBER Working paper No. 13330.

[12] Chinco, A., 2021. The ex ante likelihood of bubbles, forthcoming, Management Science.

[13] Cleveland W., 1979. Robust locally-weighted regression and smoothing scatterplots, *Journal of the American Statistical Association,* 74, 829-836.

[14] Cookson, T., and M. Neissner, 2020. Why don't we agree? Evidence from a social network of investors, *Journal of Finance,* 75 (1), 173-228.

[15] Cookson T., J. Engelberg, and W. Mullins, (2020). Does partisanship shape investor beliefs? Evidence from the COVID 19 pandemic, *Review of Asset Pricing Studies,* 10 (4), 863-893.

[16] DellaVigna, S., and J. Pollet. 2009. Investor inattention and Friday earnings announcement, *Journal of Finance,* 64 (2), 709-749.

[17] Da, Z., J. Engelberg and P. Gao, 2011. In search of attention, *Journal of Finance,* 56 (5), 1461-1499.

[18] De Clippel, G., K. Eliaz, and K. Rozen, 2014. Competing for consumer inattention, *Journal of Political Economy,* 122 (6), 1203-1234.

[19] Duffie, D. 2010. Presidential address: Asset price dynamics with slow-moving capital, *Journal of Finance,* 65, 1237-1267.

[20] Gabaix, X., 2009. Power laws in economics and finance, *Annual Review of Financial Economics,* 1, 255-93.

[21] Gabaix, X., 2014. A sparsity-based model of bounded rationality, *Quarterly Journal of Economics,* 129 (4), 1661-1710.

[22] Gabaix, X., 2016. Power laws in economics: An introduction, *Journal of Economic Perspectives,* 30(1), 195-206.

[23] Gabaix, X., 2019. Behavioral inattention, in Handbook of Behavioral Economics, North Holland, Amsterdam, Netherlands.

[24] Gabaix, X., and D. Laibson, 2006. Shrouded attributes, consumer myopia, and information suppression in competitive markets, *Quarterly Journal of Economics,* 121 (2), 505-540.

[25] Giannini, R. P. Irvine and T. Shu, 2018, Nonlocal disadvantage: An examination of social media sentiment, *Review of Asset Pricing Studies*, 8 (2), 293-336.

[26] ____________________________, 2019, The convergence and divergence of investors' opinions around earnings news: Evidence from a social network, *Journal of Financial Markets,* 42 (1), 94-120.

[27] Goldstein, M., P. Irvine, E. Kandel, and Z. Wiener, 2009. Brokerage commissions and institutional trading patterns, *Review of Financial Studies,* 22 (12), 5175-5212.

[28] Greenwood, R., and A. Shleifer, 2014. Expectations of returns and expected returns, *Review of Financial Studies,* 27 (3), 714-746.

[29] Grossman, S., and J. Stiglitz, 1980. On the impossibility of informationally efficient markets, *American Economic Review,* 70 (3), 393-408.

[30] Grullon, G., G. Kanatas, J. Weston, 2004. Advertising, breadth of ownership, and liquidity, *Review of Financial Studies,* 17 (2), 439-461.

[31] Hendershott, T., D. Livdan and N. Schurhoff, 2015. Are institutions informed about news? *Journal of Financial Economics,* 117, 249-287.

[32] Hendershott, T., A. Menkveld, R. Praz, and M. Seasholes, 2021. Asset price dynamics with limited attention, forthcoming, *Review of Financial Studies.*

[33] Hirshleifer, D. and S. H. Teoh, 2003. Limited attention, information disclosure, and financial reporting, *Journal of Accounting and Economics,* 36 (1-3), 337-386.

[34] Hirshleifer, D., S.Lim and S.H. Teoh, 2009. Driven to distraction, extraneous events and reaction to earnings news, *Journal of Finance,* 64 (5), 2289-2325.

[35] Ibbotson, R., and M. Izdorek, 2014. Dimensions of popularity, *Journal of Portfolio Management,* 40 (5), 68-74.

[36] Ibbotson, R., and M. Izdorek, P. Kaplan and J. Xiong. 2018. Popularity: A bridge between classical and behavioral finance, CFA Research Institute, Charlottesville, VA.

[37] Kacperczyk, M., S. Van Niewerburgh, and L. Veldkamp, 2016. A rational theory of mutual funds' attention allocation, *Econometrica,* 84 (2), 571-626.

[38] Kahneman, D., 1973. Attention and effort, Prentice Hall, Englewood Cliffs, NJ.

[39] Lee, C., P. Ma, and C. Wang, 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches, *Journal of Financial Economics,* 116, 410-431.

[40] Lou, D., 2014. Attracting attention through advertising, *Review of Financial Studies,* 27 (6), 1797-1829.

[41] Loughran, T., and B. McDonald (2017). The use of EDGAR filings by investors, *Journal of Behavioral Finance,* 18 (2), 231-248.

[42] Merton. R., 1987. A simple model of capital market equilibrium with incomplete information, *Journal of Finance,* 42 (3), 483-510.

[43] Odean, T., 1999. Do investors trade too much, *American Economic Review,* 89 (5), 1279-1298.

[44] Rakowski, D., S. Shirley, and J. Stark. 2020. Twitter activity, investor attention, and the diffusion of information, *Financial Management*, 50 (1), 1-44.

[45] Rogers, W., 1993. Regression standard errors in clustered samples, *Stata Technical Bulletin,* 13, 19-23.

[46] Peng, L., and W. Xiong, 2006. Investor attention, overconfidence, and category learning, *Journal of Financial Economics,* 80 (3), 563-602.

[47] Saglam, M., C. Moallemi, and M. Sotiropoulos, 2019. Short-term trading skill: An analysis of investor heterogeneity and execution quality, *Journal of Financial Markets,* 42 (1), 1-28.

[48] Yoshinago, C., and F. Rocco, 2020. Investor attention: Can Google search volumes predict stock returns, *Brazilian Business Review,* 17 (5).

Appendix

Stocktwits is one measure of investor attention. The Stocktwits sample is broad, has a reasonably long time series, and our conclusions do not appear to be influenced by the degree of institutional or retail ownership in particular stocks. However, the level of certainty regarding the contentions in the paper would clearly be strengthened if we alternative data sources that confirmed the observations in the paper. The conclusion briefly discusses the drawbacks of two alternatives: Twitter and Google Search Volume, and outlines the difficulties in applying these data sets to the aggregate attention problem. However, one alternative is provided by Loughran and McDonald (2017) who use the Freedom of Information Act to obtain SEC Edgar search logs from 2003-2015, and make the data available to other researchers.[11] In this Appendix we examine several of our conclusions on aggregate attention using the Edgar search logs as a measure of aggregate investor attention.

The Edgar search log

The processing and cleaning of the original 3,319 daily Edgar search log files obtained from the SEC is outlined in Loughran and McDonald (2017). They describe several filters that limit the amount of data actually analyzed. The most relevant filters for our analysis include source-based data errors for particular months in 2005 and 2006. Due to data integrity issues, we eliminate this period entirely and focus on two periods after the data issues are resolved; the 108 month period from January, 2007 to December, 2015, when the data set ends, and the 60 months where the Edgar search log data overlap with our Stocktwits data from January, 2011 to December, 2015. Another critical issue is the exclusion of automatic robotic requests, which is done by Loughran and McDonald (2017) using a filter that sets a maximum on daily requests to 50 requests per day from a particular IP address (Lee, Ma, and Wang, 2015). The robot filter is clearly imperfect, and alternative robot definitions can produce vastly different results, as noted by Loughran and McDonald (2017). We follow their suggestion

[11] This data is described and available at: https://sraf.nd.edu/data/edgar-server-log/

and use the *htm* file type, which is a type directly viewable in a web browser, as the best proxy for individual search of company filings. Despite the great care taken by Loughran and McDonald (2017) the reader should be aware that the resulting Edgar search log is not likely to be free of automatic robotic requests. The authors only conclude that the count of *htm* file type views is most likely to be associated with individual users.

Edgar search data

Summary statistics for the Edgar search data is presented in Appendix Table A1. The data shows the extensive size of the market for Edgar searches, beginning with over 6.4 million searches in 2007, and rising to over 28.0 million searches in 2015. This general rise is searches is reflected in the average per stock which rises from 123.8 per year in 2007, to 655.7 annually in 2015. The use of Edgar as a research tool appears to be quite broad, covering most of the NYSE-Nasdaq universe of publicly listed stocks. The number of stocks search in 2007 is 4,471, and this number gradually declines and then rises again at the end of the sample, reflecting the total number of public firms.

The concentration of Edgar search data

To estimate the properties of the Edgar search log data as a measure of attention, we first estimate the power law coefficients for the same five levels of stocks we estimate for Stocktwits attention, market capitalization, and trading volume. Figure A1 compares the monthly power law coefficients for Edgar search data against market cap for the 60 month overlapping period of January, 2011 to December, 2015. We also present power law coefficient statistics for this period, and the 2007-2015 period in Panels A and B of Table A2.

Panel A reveals that the power law coefficients for the Edgar data match market cap concentration closely for the most popular 100 stocks in the 2011-2015 period. However, as we extend the sample to include more stocks, the power law coefficients for the Edgar search data stay relatively high, never falling below -1.6, considerably above the concentration of market capitalization. This indicates that the Edgar search data is much more concentrated in the largest securities than is market cap, trading volume or Stocktwits attention as soon as we include more than the top 100 stocks in the test sample. Overall, the Edgar search sample shows even more of a tendency for the neglected firm affect than does the Stocktwits sample. The reason for this is not clear, as stated above, the Edgar data likely still contains a number of automated searches, but how this feature of the data affects the power law coefficients of this data set is only subject to speculation without greater information regarding the individuals using the data set. It is possible that automated searches are concentrated in the largest stocks, but this is a speculative contention.

Power law coefficient estimates for the full 2007-2015 period are presented in Panel B of Table A2. These estimates are higher than those for the 2011-2015 period, indicating that Edgar searches were even more concentrated in the top stocks in the 2007-2010 period. As use of the service increased over time, the distribution of Edgar searches tended to broaden out and cover more stocks, but inclusion of this period does not alter the conclusions resulting from the estimates in Panel A.

Correlation

Panels A and B of Appendix Table A3 present own correlation statistics for the Edgar data. These results use the 2011-2015 overlap period for comparison with the data in the paper. These own correlations are qualitatively similar to the result in Table 5 of the paper for the Stocktwits attention data. The most searched stocks in Edgar vary from month to month more than either market cap or trading volume, consistent with our results in Table 5. Qualitatively, Edgar search is somewhat more consistent than Stocktwits attention, indicating that the month to month searches in Edgar tend to focus on the same set of stocks more than do Stocktwits users. This may reflect that Stocktwits as a measure of attention is more volatile, or it could reflect the contamination of regular repeated searches in Edgar that are automatically undertaken.

When we examine the stocks that replace the top stocks from month to month in Panel B, we see that Edgar search replacement stocks are more difficult to predict than those of either market cap or volume. For example, the average month $t-1$ replacement rank of the 50 largest market cap stocks is 26.7, similar to our full sample estimates. This number can be compared to the minimum rank of 25.5 if the set of stocks did not change from month $t-1$ to month t. For trading volume the replacement rank is 42.0, but for Edgar Search the mean replacement rank is 72.0. This number is lower than the 101.9 replacement rank for the Stocktwits sample, but still reflects the same conclusion that the stocks that attract the most attention in a particular month are more difficult to predict from past data than either market cap, which is quite predictable, or trading volume, which is moderately predictable.

We examine the cross correlations between Edgar search, market cap and trading volume in Panel C of Table A3. Here we find results that are quite comparable to those in Panel C of Table 5. The overall pattern of cross-correlations is similar using either Edgar search data or Stocktwits as the measure of attention. For some sets of stocks Stocktwits attention is more highly correlated with market cap and trading volume, for other cutoffs Edgar search is more highly correlated. Overall, both measures of attention show a moderate level of positive correlation with market cap and trading volume, producing similar results.

Finally, in Panel D of Table A3 we present the market cap-Edgar search cross correlations for market capitalization measured in two different ways. Institutional market cap is the total market capitalization of the company multiplied by the percentage institutional ownership. Retail market cap is the remainder, or total market cap multiplied by 1 minus the percentage institutional ownership. We estimate the panel due to the concern that robotic searches are likely still present in the data set. If we assume that robotic searches are more likely to be generated by institutional rather than retail investors, we could see differences in the Edgar search-market cap cross-correlations for institutional verses retail ownership. However, there is little difference across the two samples, indicating no particular bias towards institutional or retail stocks in the Edgar search sample. This reflects the same conclusion as when the same exercise is performed for the Stocktwits attention measure.

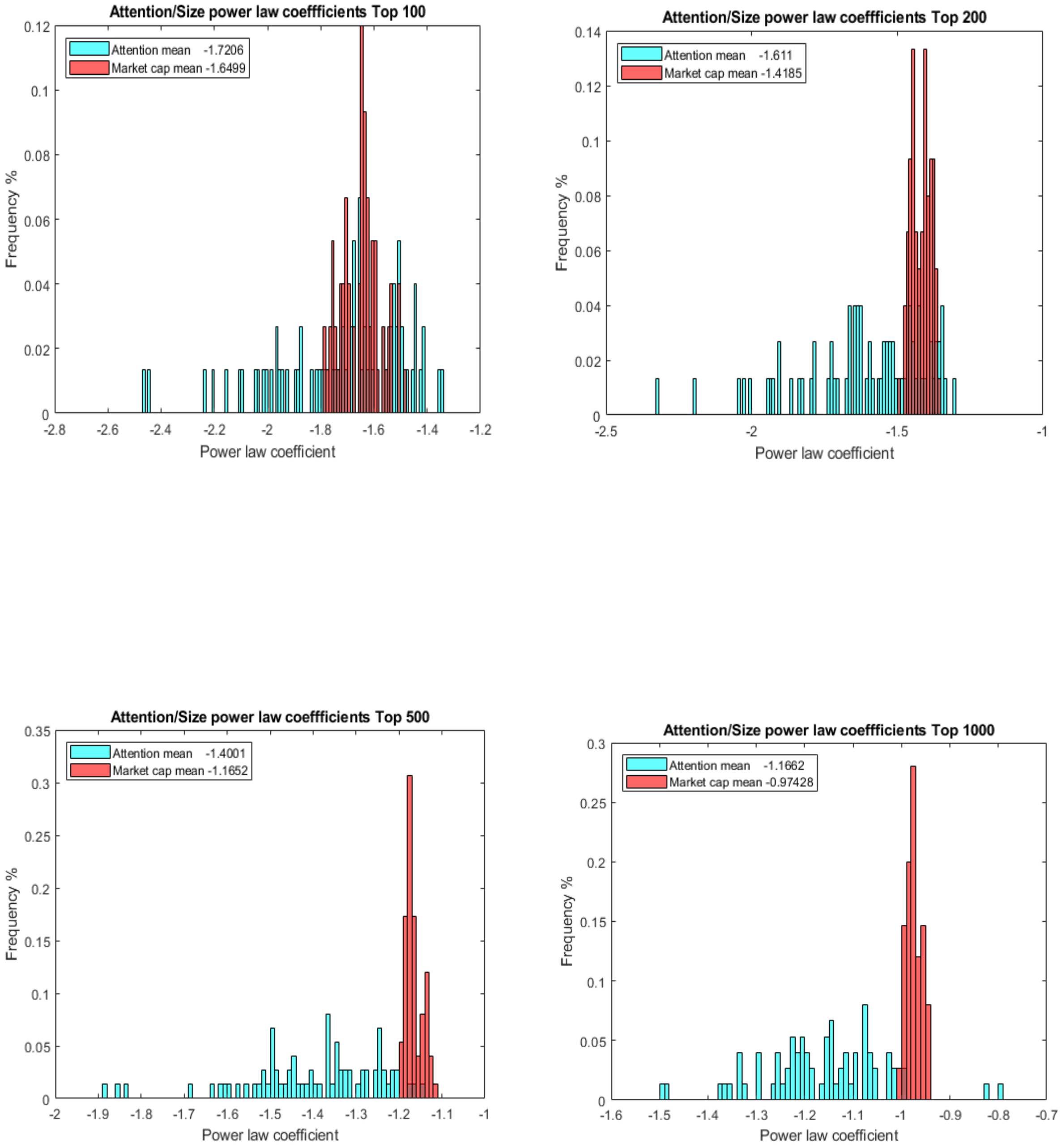
News concentration

Finally, we conclude with an examination of the power law coefficients for the Ravenpack data set. Aside for completeness from our perspective, such estimates aid researchers in understanding the focus of the Ravenpack data when undertaking examinations of news data effects on the stock market. We report the power law coefficients for the Ravenpack sample in Panel C of Table A2 and present visual representations of monthly estimates in Panel B of Figure A1. We find news to have similar concentration as market cap for the set of 50 or 100 stocks. Yet for larger sets of stocks, though the power law coefficients for market cap, Stocktwits attention, and trading volume all tend to fall towards one, the power law coefficients for news articles decrease much more slowly, only falling to -1.53 for the set of 1,000 stocks. Unlike the coefficients for the Edgar search data, they are monotonically decreasing as we add more stocks to the power law regressions, but they are falling at a much lower rate than the coefficients of market cap. This indicates that, perhaps not surprisingly, the news is also biased towards the largest stocks. Researchers should keep this fact in mind when developing insights using the Ravenpack news data set.

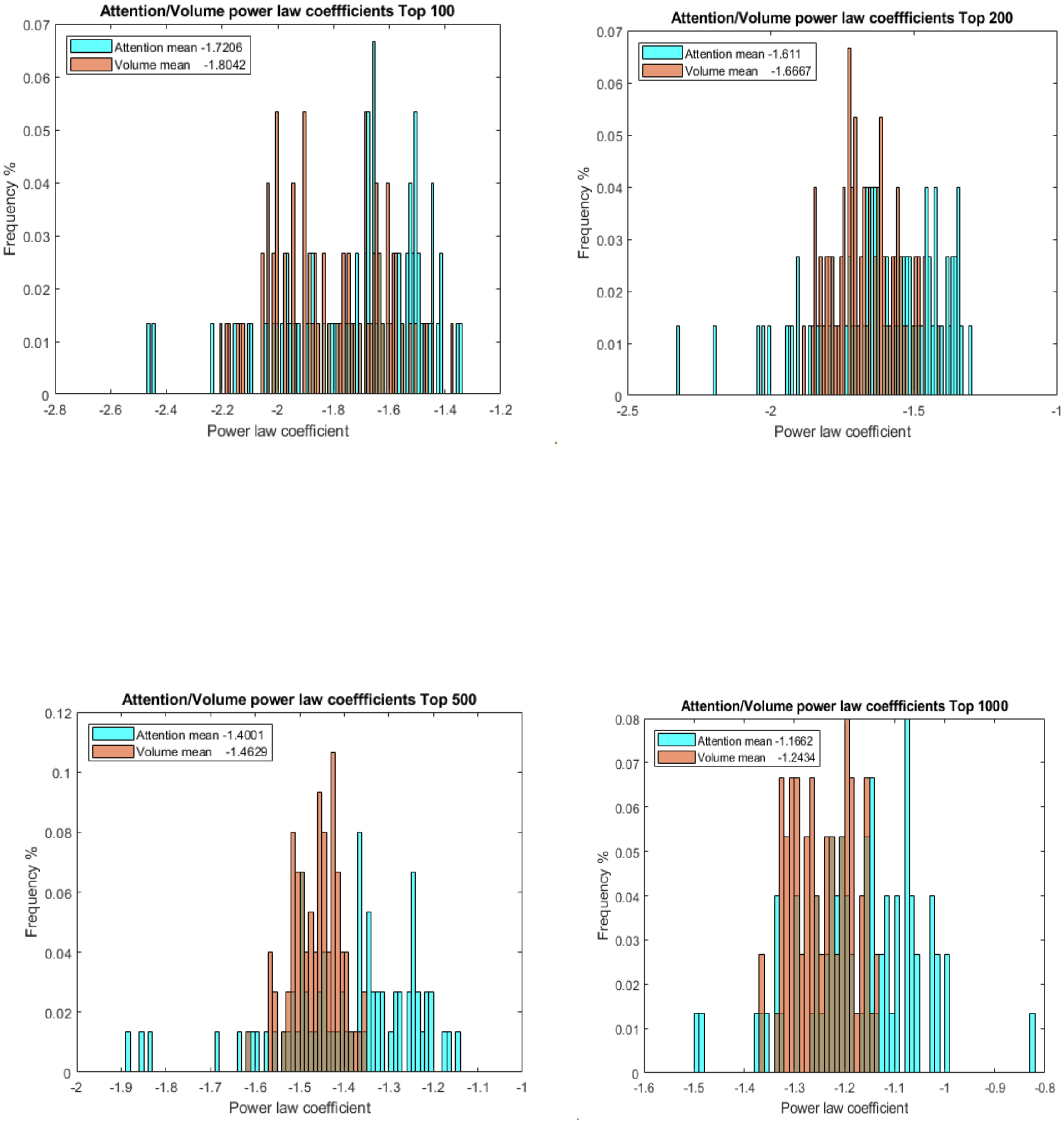
Figure 1 – Power law coefficients as estimates of concentration

These figures present pairwise histograms of power law coefficient distributions. Linear power law equations are estimated over 75 months for Stocktwits attention, market capitalization, and trading volume. The figures show the distribution of coefficients and the sample average.

Panel A. – Attention versus Size.



Panel B. – Attention versus Volume



Panel C. Size versus Volume

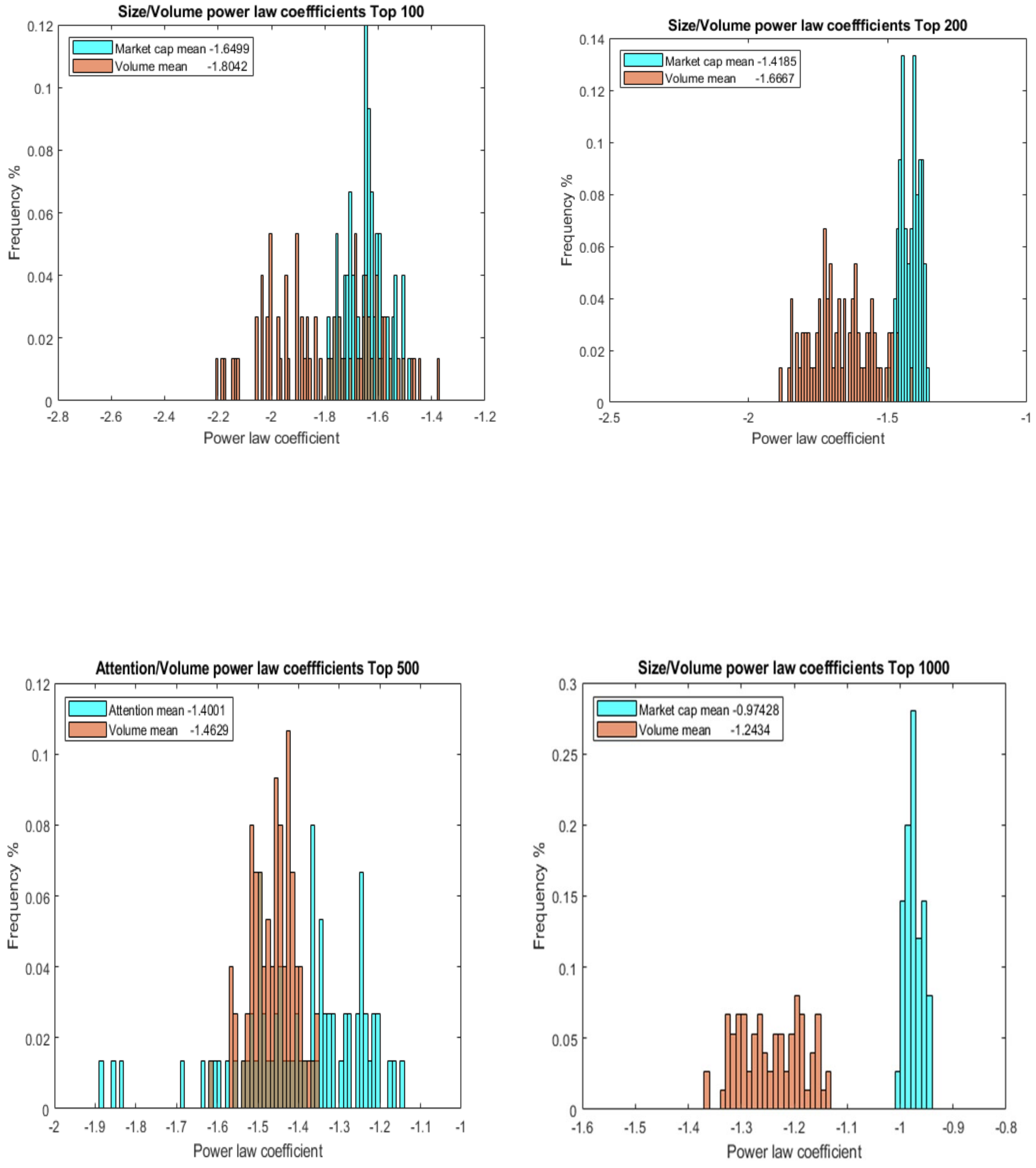
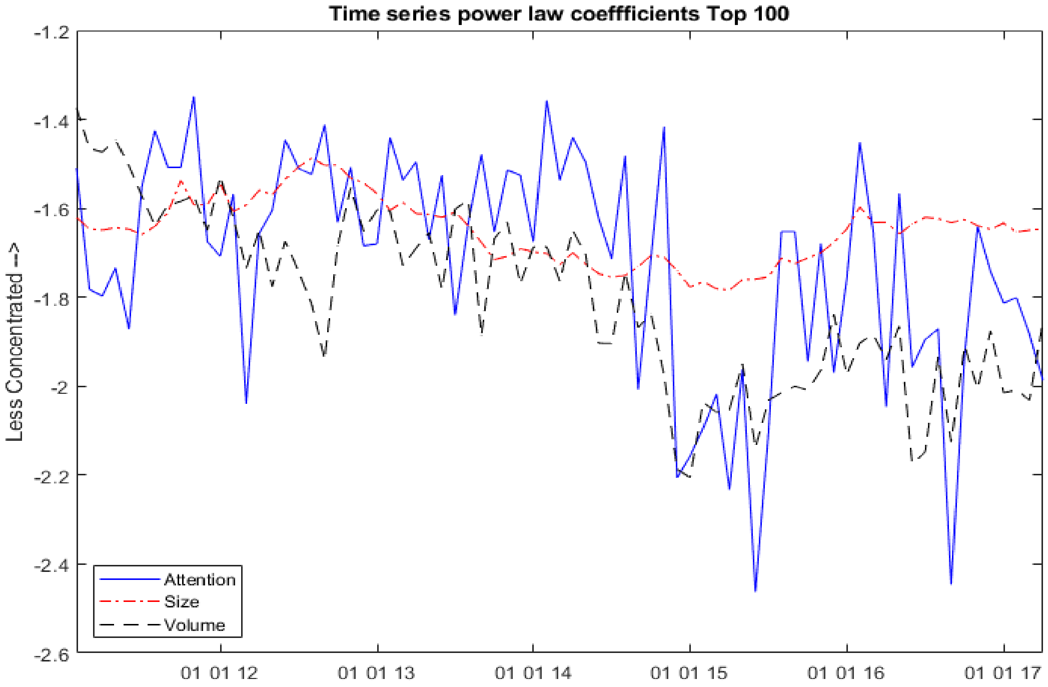
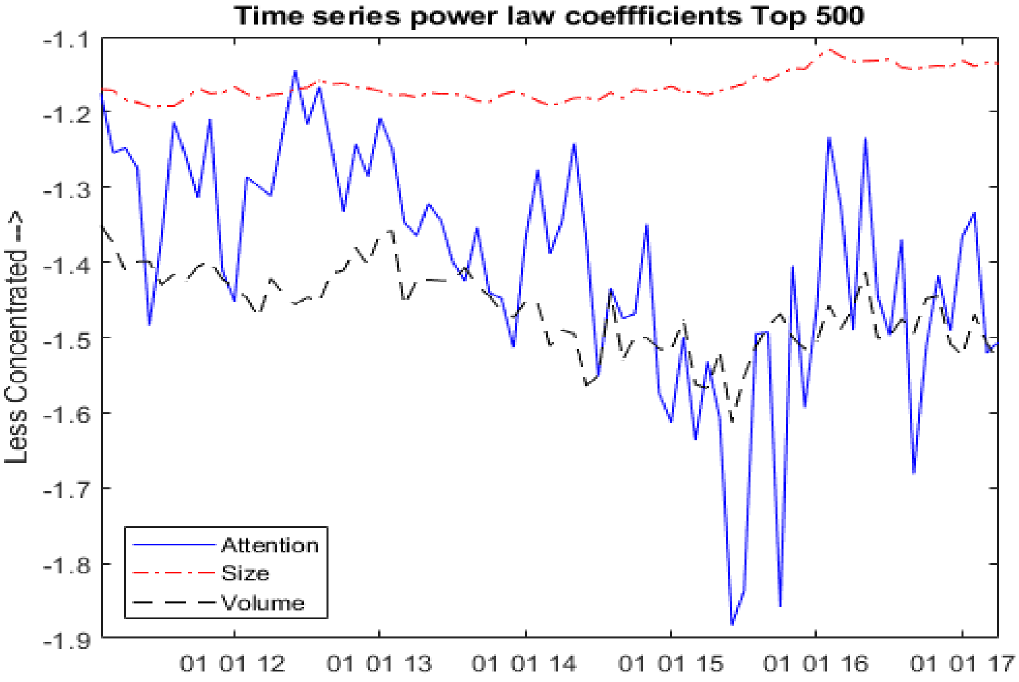


Figure 2. The time series evolution of power laws

These figures present the power law coefficients of Stocktwits posts, market capitalization and trading volume over time. The coefficients are the same as those presented in Figure 1, but plotted by calendar month, from January 2011, through March, 2017.





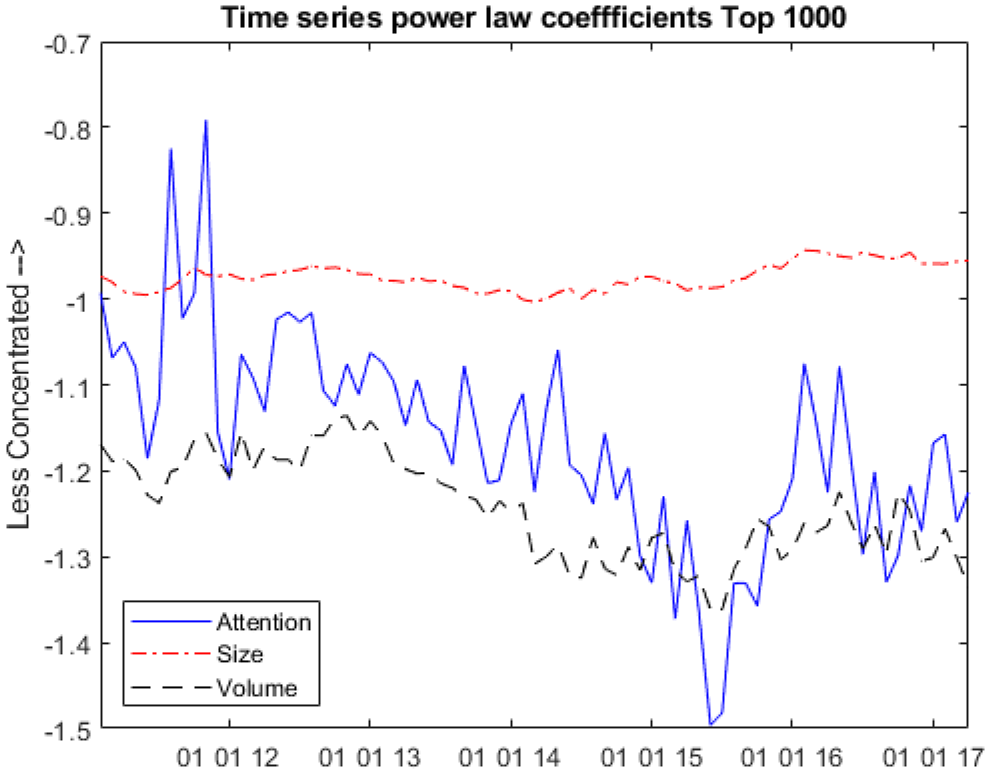
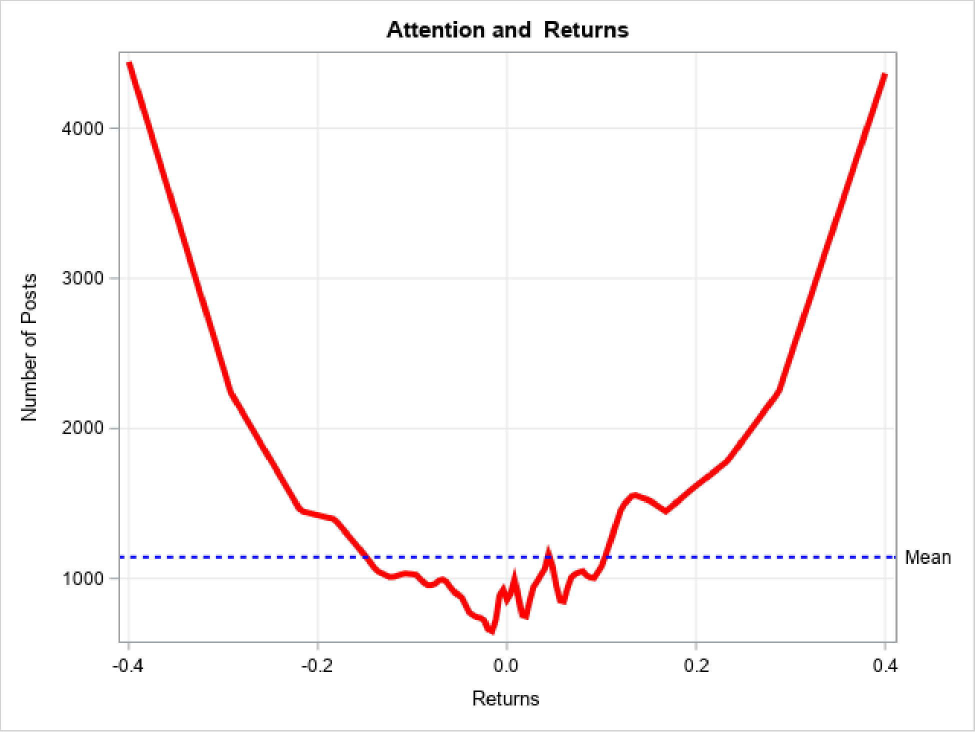
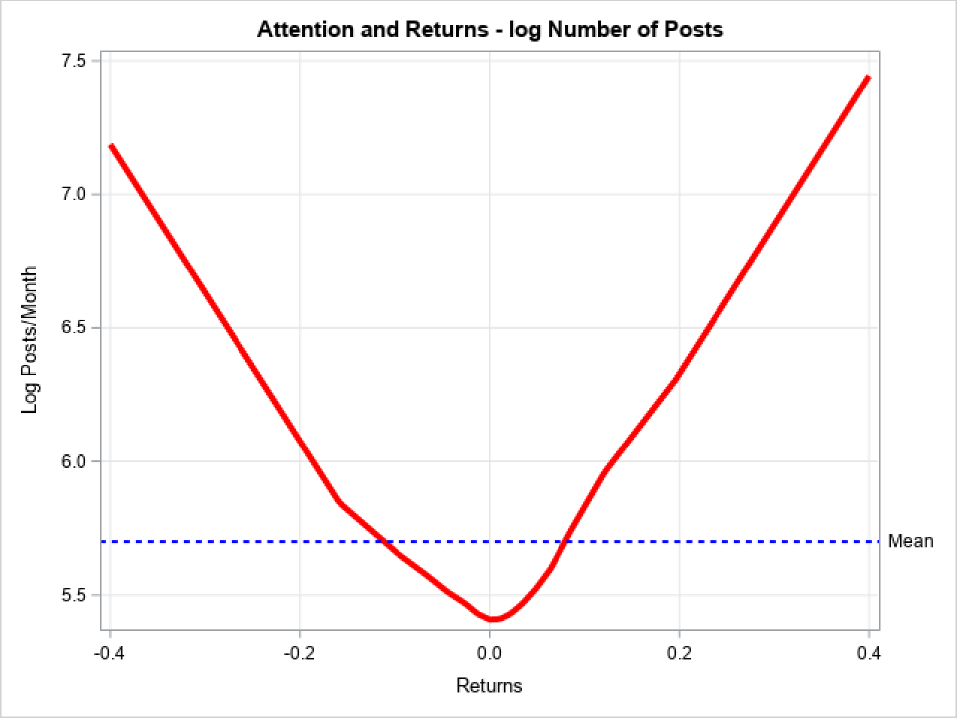

Time series power law coeffficients Top 1000
Less Concentrated -->
-0.7
-0.8
-0.9
-1
-1.1
-1.2
-1.3
-1.4
-1.5
01 01 12
01 01 13
01 01 14
01 01 15
01 01 16
01 01 17
Attention
Size
Volume

Figure 3. Attention and Returns

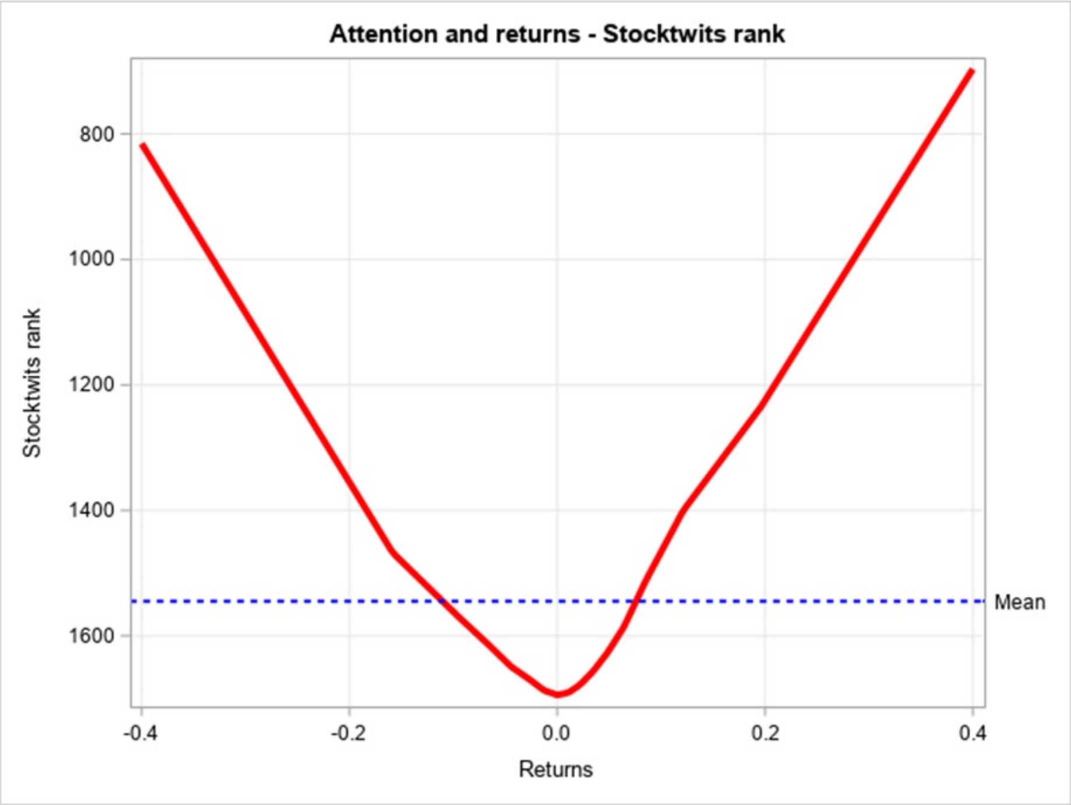
These figures present the fitted values from non-parametric regressions of attention and returns. Returns are winsorized at 1% and 99% (approximately +- 40% /month) for clarity. This representative data comes from the period February, 2017 to March, 2017.

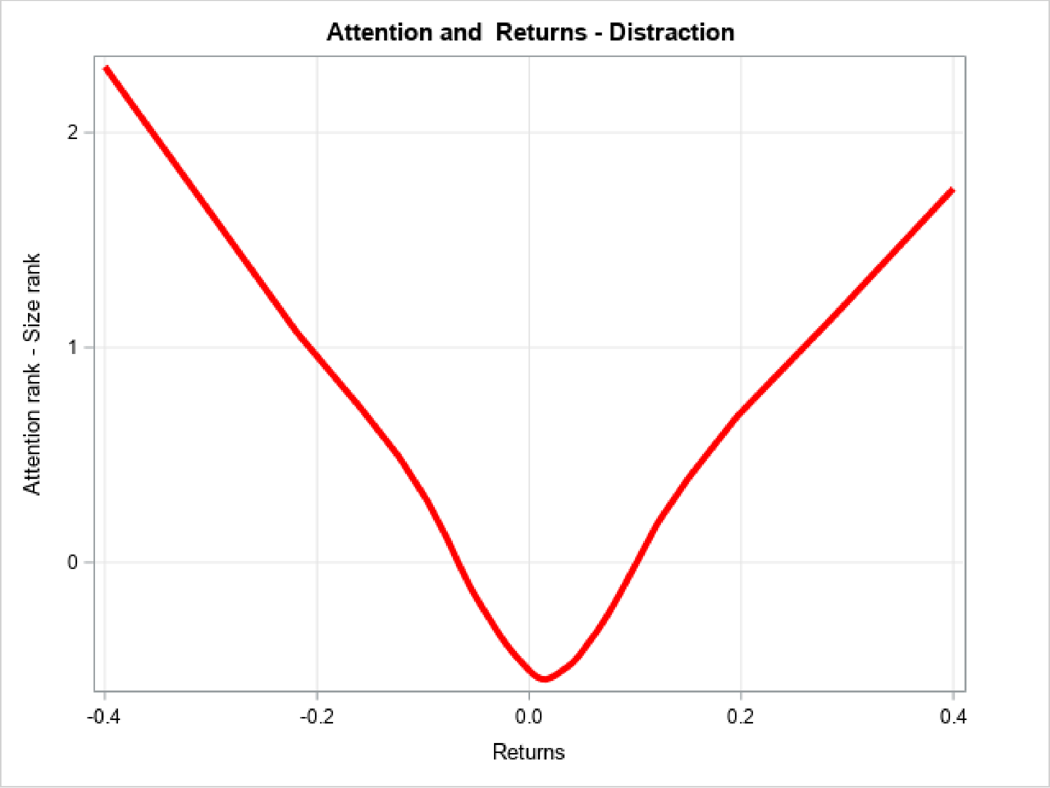
Panel A. Number of posts and returns



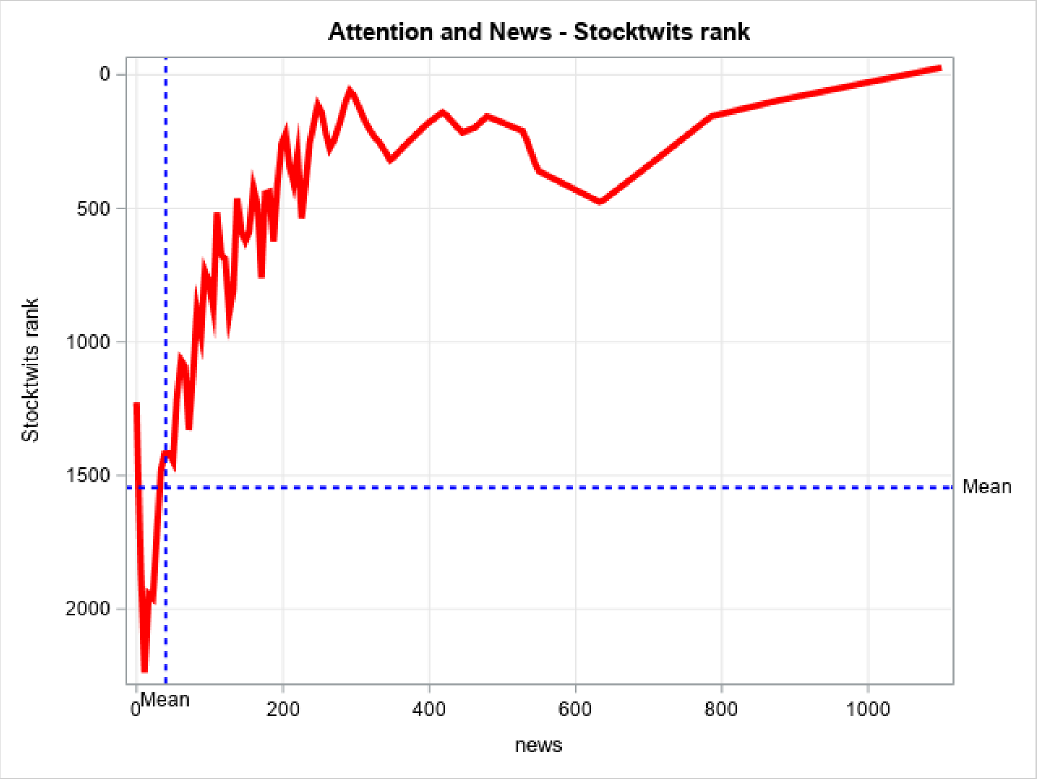


Panel B: Returns and attention rank





Panel C: News and Stocktwits rank



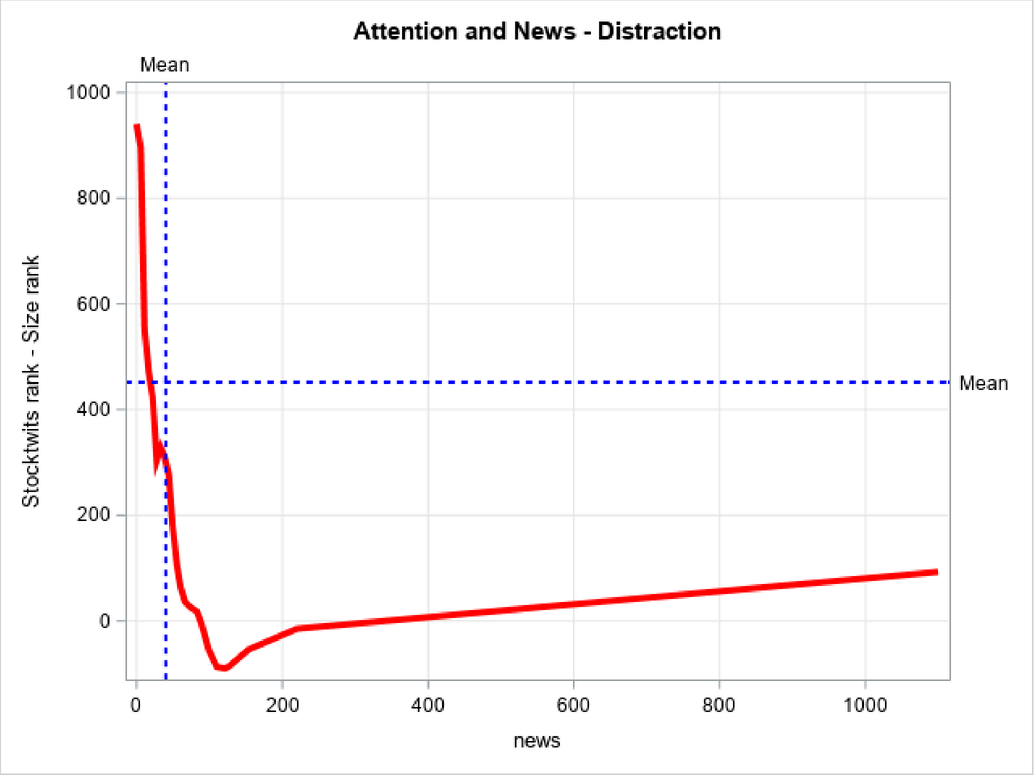


Table 1

This table presents summary information on the Stocktwits data set and the Ravenpack news data set. *Total Posts* is the number of single-stock posts in the year. *Total Articles* is the total number of Dow Jones, PR Newswire and Web Edition articles per year on the sample stocks. *Average/Stock* and *Max* are monthly averages and maximums within the relevant year. Number of stocks is the total number of different stocks covered by the relevant data each year. All data is from January, 2011 – March, 2017.

Panel A: Stocktwits summary data

Year	Total Posts	Average/ Stock	Max	Number Stocks
2011	661,280	32.9	4,002	3,655
2012	1,559,803	58.1	10,123	3,976
2013	4,257,018	151.8	43,283	4,018
2014	14,795,233	460.0	54,302	4,290
2015	21,150,513	531.7	51,146	4,731
2016	26,594,848	747.6	74,429	4,243
2017	7,836,715	768.2	54,926	4,277

Panel B: Ravenpack summary data

Year	Total Articles	Average/ Stock	Max	Number Stocks
2011	668,835	33.7	3,506	3,604
2012	1,012,395	38.2	3,792	3,889
2013	1,067,027	38.7	3,145	3,927
2014	1,091,906	34.4	3,769	4,239
2015	1,331,775	34.4	3,406	4,541
2016	1,384,619	39.2	2,663	4,174
2017	379,844	38.5	2,299	3,981

Table 2. Rank analysis of highly mentioned stocks

This table presents rank information on often-mentioned stocks. *Frequency* is the number times the stock was among the 20 most highly mentioned stocks in a month. *Stocktwits* is the average rank over all months conditional on the stock being in the Top 20. *Market Cap* is the average monthly rank of the stock's market capitalization. *Volume* is the average monthly rank of the stock's volume. All data is from January, 2011 – March, 2017.

TICKER	Top 20	Stocktwits rank	Size rank	Volume rank
AAPL	75	1.32	1.17	22.09
AMZN	75	5.32	18.56	263.41
GOOG	75	3.11	10.88	517.33
MSFT	73	9.60	2.89	7.04
NFLX	71	7.39	270.32	222.76
FB	59	3.22	46.88	10.32
BAC	56	10.04	24.0	1.05
TSLA	47	7.72	197.85	173.81
JPM	37	12.3	12.86	17.19
TWTR	37	6.46	269.05	19.51
GS	36	12.31	48.89	231.14
INTC	32	14.81	22.84	7.91
C	31	12.68	23.52	9.52
LNKD	31	11.65	507.19	430.16
PCLN	27	12.96	90.15	1053.2
GILD	24	9.71	29.42	50.04
CMG	23	14.17	333.91	1119.8
IBM	22	15.09	14.59	202.27
GPRO	20	10.6	1317.4	106.3
YHOO	18	15.44	144.06	24.78
BBRY	17	10.12	698.24	23.65
HPQ	17	12.47	79.76	24.76
RIMM	16	10.75	363.63	26.25
WMT	16	14.13	8.5	82.94
DIS	15	14.33	22.6	104.13
MS	15	13.4	101.73	24.33
T	15	15.8	10.27	17.6
DDD	13	11.31	733.69	249.38
GMCR	13	14.08	380	218.85
F	12	15.3	78.33	6.58
SCTY	12	13.08	809.83	240.33
PLUG	11	9.82	2365.4	42.18
CSCO	10	14.7	27.8	6.1
VRX	10	11.5	418.4	19
YELP	10	12.1	986	186.6
GE	9	14.67	6.7	5.33
LULU	9	13.44	514.89	398.0
ZNGA	9	9.78	1199.9	23.89
FIT	8	13.13	1529.0	112.63
JCP	8	17.2	1082.9	15.13

Table 3. Rank analysis of top news mention stocks

This table presents summary rank information on the stocks that were most often mentioned in the news. *Top 20 News* is the number times the stock was among the 20 most highly mentioned stocks in a month. *News rank* is the average of the *News rank* among all firms. *News, Stocktwits, Size* and *Volume* rank is the average rank over all months conditional on the stock being in the top 20. Data is from January, 2011 – March, 2017.

TICKER	Top 20 News	News rank	Stocktwits rank	Size rank	Volume rank
AAPL	75	1.45	1.32	1.17	22.0
MSFT	75	3.69	10.1	2.88	7.11
GM	72	8.50	91.2	89.4	47.4
JPM	63	8.76	22.1	11.6	22.5
GS	58	12.5	29.3	53.0	278.0
FB	57	5.68	3.28	44.6	10.5
T	57	9.58	45.1	11.9	16.4
BAC	56	10.3	12.9	23.4	1.09
IBM	56	12.1	39.4	15.0	224.2
AMZN	54	12.0	4.98	17.5	257.0
F	51	12.9	47.6	78.8	7.00
GE	51	12.5	53.2	5.86	6.69
VZ	45	10.3	45.3	19.2	38.8
BA	40	12.8	79.3	48.3	215.4
GOOG	39	2.05	2.51	10.7	424.3
WMT	37	12.8	34.4	8.86	87.7
YHOO	36	10.9	38.3	156.0	30.3
DB	35	11.2	600.3	140.0	477.8
WFC	34	15.2	50.9	10.0	21.0
TWTR	31	8.74	7.10	277.5	18.9
C	30	12.9	18.5	23.3	11.7
HPQ	29	13.3	59.0	82.7	31.4
CSCO	28	12.8	37.0	27.9	6.68
CMCSA	26	13.6	175.3	35.9	42.5
MS	26	11.7	135.7	83.3	41.4
RIMM	19	9.95	73.6	414.5	23.3
TSLA	18	10.5	9.06	180.9	183.1
NFLX	17	13.2	6.06	309.9	203.0
INTC	16	13.9	18.4	23.06	8.38
BBRY	15	13.0	16.7	766.0	27.2
DIS	12	14.7	30.2	25.1	87.7
EBAY	12	16.4	65.2	75.2	57.2
FCAU	12	11.5	537.1	367.2	140.2
ORCL	12	13.8	43.7	16.4	17.4
QCOM	11	15.0	70.5	40.7	52.9
RJF	10	10.8	1515	576.3	1122
PFE	9	12.0	42.1	15.4	7.11
RY	9	14.5	901.2	47.5	1107
MCD	8	14.7	60.2	41.8	153.3
VRX	8	9.00	13.1	462.9	21.1

Table 4. Power Law Statistics

This table presents average power law coefficients across five sets of stocks, Stocks are ranked by their rank with respect to the variable in question. Sample covers the time period from Jan, 2011-March, 2017.

Panel A. Attention

	Attention				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-1.788	-1.721	-1.611	-1.400	-1.166
Standard Deviation	0.270	0.249	0.216	0.155	0.124
Mean R-Square %	97.1	98.1	98.5	98.2	96.5
N=	75	75	75	75	75

Panel B. Market Capitalization

	Size				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-2.021	-1.650	-1.419	-1.165	-0.974
Standard Deviation	0.183	0.074	0.034	0.020	0.016
Mean R-Square	93.2	94.3	95.7	96.3	96.1
N=	75	75	75	75	75

Panel C. Trading Volume

	Volume				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-1.937	-1.804	-1.667	-1.463	-1.243
Standard Deviation	0.308	0.202	0.111	0.054	0.060
Mean R-Square	96.9	97.9	98.3	98.3	97.2
N=	75	75	75	75	75

Table 5

This table presents own and pairwise correlations for the five different sets of stocks examined. In Panel A *Frequency* is the average number of stocks in month t+1 that were in the sample in month t. *Rank correlation* is the correlation between the rank of stocks in month t, and the rank of stocks in month t+1. Panel B presents the average rank of the stocks in month *t-1* that become Top Rank stocks in month *t*. Panel C presents *Frequency* and *Rank correlation* statistics for pairwise comparisons.

Panel A. Average time-series correlations

	Size			**Attention**			**Volume**		
Stocks	Frequency	Percent	Rank Correlation	Frequency	Percent	Rank Correlation	Frequency	Percent	Rank Correlation
50	48.21	96.42%	0.989	30.77	61.54%	0.670	40.19	80.38%	0.790
100	97.34	97.34%	0.983	61.28	61.28%	0.607	81.92	81.92%	0.813
200	194.45	97.23%	0.991	120.95	60.48%	0.595	165.83	82.92%	0.831
500	488.57	97.71%	0.993	308.23	61.65%	0.576	425.67	85.13%	0.841
1,000	977.42	97.74%	0.994	650.43	65.04%	0.559	878.43	87.84%	0.863

Panel B: Replacement Stocks

	Size		**Attention**		**Volume**	
Stocks	Replacement Rank	Percent Minimum	Replacement Rank	Percent Minimum	Replacement Rank	Percent Minimum
50	26.70	104.7%	101.86	399.5%	41.96	164.5%
100	51.77	102.5%	167.35	331.4%	83.50	165.3%
200	102.61	102.1%	287.75	286.3%	157.58	156.8%
500	254.20	101.5%	539.05	215.2%	330.54	132.0%
1,000	506.13	101.1%	805.15	160.9%	585.06	116.9%

Panel C: Average cross-sectional correlations

	Size-Attention			**Attention-Volume**			**Size-Volume**		
Stocks	Frequency	Percent	Rank Correlation	Frequency	Percent	Rank Correlation	Frequency	Percent	Rank Correlation
50	17.99	35.98%	0.248	18.37	36.74%	0.252	16.92	33.84%	0.081
100	38.01	38.01%	0.320	38.73	38.73%	0.339	35.96	35.96%	0.343
200	75.89	37.95%	0.351	86.45	43.23%	0.328	81.23	40.62%	0.333
500	228.51	45.70%	0.372	262.08	52.42%	0.333	276.41	55.28%	0.312
1,000	531.23	53.12%	0.353	625.16	62.52%	0.368	650.61	65.06%	0.377

Table 6. Regression summary statistics

Market Cap is the value of equity in $Millions. *News* is the number of RavenPack news articles in a month. *Sentiment* is the average Ravenpack sentiment for the news articles in that month. *AbnVol* is abnormal volume measured as month *t* trading volume (in 1,000s) relative to the average over the previous 3 months. *Coverage* is the number of analysts that cover the firm, *Ivol* is the idiosyncratic volatility in month *t*. *BtM* is the book-to-market ratio, negative book values are set to 0. *Advertise* is the percentage of firm sales spent on advertising. *Tech* is a dummy variable set to 1 for SIC codes whose first three digits are 737. *Pharm* is a dummy variable set to 1 for SIC codes whose first three digits are 283.

	Mean	**Std Dev.**	**Median**	**Min**	**Max**
Market Cap ($MM)	8,033	2,634	1,338	0.85	75,071
News	36.6	86.9	19.0	0	3,792
Sentiment	0.04	0.10	0.0012	-0.86	0.88
AbnVol (000s)	2.95	425.3	-0.87	-63,096	56,009
Coverage	9.3	7.24	7	0	56
Ivol	0.025	0.026	0.018	0.004	1.42
BtM	0.44	0.76	0.31	0	80.8
Returns %	1.07	12.99	0.94	-93.53	1598.44
Advertise	0.013	0.185	0	0	18.75
Tech	0.073	-	-	-	-
Pharm	0.089	-	-	-	-

Table 7. Determinants of Attention

The dependent variables in these regressions in the log of posts in a month, the Stocktwits post rank in a month, and a standardized measure of *Distraction,* the Stocktwits posts rank less the market capitalization rank. All variables are defined in the Table 6 column header except for *Ret > 20%* and *Ret < 20%*. These variables are two threshold dummy variables set to 1 if the monthly return is greater than the 20% threshold value. T-statistics are presented in parentheses below the coefficient estimate.

Variable	(1) Log Posts	(2) Log Posts	(3) Stocktwits Rank	(4) Distraction (std.)
Ln (Mkt Cap)	0.255	0.261	-97.9	-0.382
	(20.67)	(21.60)	(-11.77)	(-51.78)
Lag Return	0.469	0.47	-295	0.212
	(8.72)	(8.91)	(-26.00)	(24.54)
Ret (+)	0.236	-0.373	-5.06	0.004
	(1.54)	(-2.85)	(-0.09)	(0.09)
Ret (-)	-1.05	-0.83	-56.5	0.041
	(-6.62)	(-4.43)	(-1.33)	(1.41)
Ret > 20%		0.504	-185	0.142
		(10.73)	(-10.02)	(9.58)
Ret < -20%		0.117	-82.5	0.076
		(2.82)	(-7.61)	(8.60)
News	0.003	0.003	-0.436	0.0004
	(26.76)	(26.94)	(-5.08)	(4.77)
Sentiment	0.230	0.221	-127	0.072
	(5.52)	(5.36)	(-7.59)	(5.44)
AbnVol	0.00001	0.00001	-0.024	0.0001
	(1.05)	(0.77)	(-2.50)	(2.54)
Coverage	0.45	0.45	-70.1	0.028
	(29.98)	(30.47)	(-6.13)	(4.68)
Ivol	18.95	18.95	-5646	4.57
	(19.09)	(19.87)	(-19.33)	(18.87)
BtM	-0.063	-0.062	-27.9	0.017
	(-8.15)	(-8.09)	(-3.89)	(2.6)
Advertise	0.053	0.053	-2.97	-0.008
	(2.61)	(2.65)	(-0.28)	(-0.09)
Announce	0.169	0.165	-125	0.105
	(5.7)	(5.53)	(-27.74)	(27.54)
Tech	0.062	0.062	-280	0.182
	(3.44)	(3.42)	(-3.37)	(7.39)
Pharm	0.507	0.493	116	-0.098
	(16.71)	(16.61)	(3.37)	(-3.69)
Month FE	Y	Y	N	N
Stock FE	N	N	Y	Y
Winsorized	N	N	N	N
R-square	19.9	20.2	23.4	60.6
N	187,870	187,870	187,870	187,870

Table 8. Distraction and future returns

Panel A presents average coefficients and cross-sectional t-statistics from Fama-MacBeth regressions of returns in month *t+1* against *Distraction*, and stock characteristics. *Distraction* is the standardized difference between the Stocktwits rank sand the Market Cap, so that a positive Distraction indicates a popular stock relative to its market cap weight. *Distraction* (+) contains only positive *Distraction* values, and *Deviation* (-) is analogous. *Ret1* is the stock return in month t-1, and *Ret2-3* is the stock return in months t-2 to t-3. *Ret4-6*, and *Ret7-12* are past returns over the specified months. Columns (1)-(3) present unfiltered returns. Columns (4)-(6) winsorize returns at the 1% and 99% levels. T-statistics are presented in parentheses below the coefficient estimate. Panel B sorts returns into *Distraction* deciles. Panel C presents Carhart (1997) 4-factor alphas double-sorted into quintiles first by the Correlation between *Volume* rank and *Attention* rank, and then on *Distraction (Popularity-Neglected)*. *Correlation* is measured as the correlation between the monthly volume rank of stock and its contemporaneous-month attention rank estimated over a past 12-month rolling window. Robust Newey and West (1987) *t*-statistics are reported in parentheses.

Panel A. Fama-MacBeth regression coefficients

Variable	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0130	0.0167	-0.0168	-0.0099	0.0011	0.0506
	(1.14)	(1.26)	(-1.11)	(-0.89)	(0.08)	**(-4.04)**
Distraction	-0.0044			-0.0056		
(std.)	**(-3.16)**			**(-4.48)**		
Distraction (+)		-0.0074			-0.0106	
		(-3.25)			**(-5.40)**	
Distraction (-)			-0.002			-0.0009
			(-1.22)			(-0.71)
Ln (Mkt Cap)	-0.0005	-0.0005	0.0009	0,0005	0.0002	0.0024
	(-1.07)	(0.83)	(1.55)	(-1.14)	(0.33)	(5.09)
BtM	0.0064	0.0064	0.0066	0.0061	0.0061	0.0063
	(6.11)	**(6.10)**	**(6.18)**	**(-7.17)**	**(7.14)**	**(7.29)**
Ret1	0.0063	0.006	0.0059	0.0074	0.007	0.0071
	(0.87)	(0.83)	(0.80)	(1.18)	(1.12)	(1.11)
Ret2-3	0.002	0.0018	0.0018	0.0053	0.0049	0.0052
	(0.32)	(0.31)	(0.29)	(1.04)	(0.06)	(1.01)
Ret4-6	0.0067	0.0062	0.0074	0.0062	0.0055	0.0069
	(1.54)	(1.40)	(1.70)	(1.62)	(1.42)	(1.81)
Ret7-12	0.0064	0.0059	0.0067	0.0079	0.0072	0.0083
	(1.77)	(1.65)	(1.84)	**(2.43)**	**(2.23)**	**(2.53)**
N	75	75	75	75	75	75

Panel B: Portfolio sorts

Distraction	1	2	3	4	5	6	7	8	9	10
Returns *t+1*	0.99	0.96	0.93	0.92	0.97	0.77	0.85	0.44	0.40	-0.44
Returns *t*	0.98	1.12	1.33	1.44	1.64	1.60	1.51	1.48	1.17	2.64

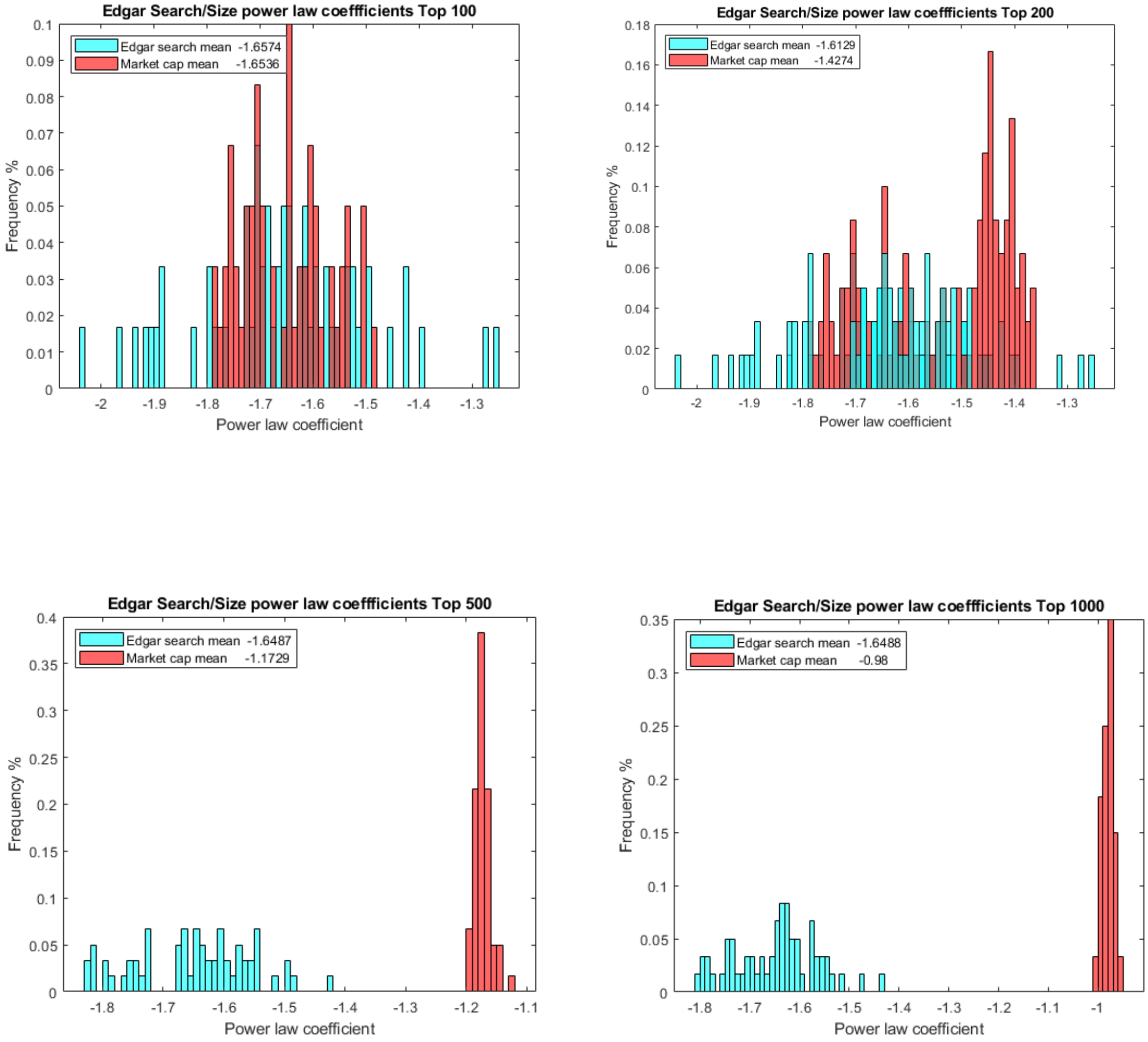
Panel C: Double sorted Returns

Deviation	1 Neglected	2	3	4	5 Popular	Popular -Neglected
1 Negative Correlation	0.22	0.08	0.07	0.12	-0.07	-0.28
	(2.76)	(0.74)	(0.58)	(1.07)	(-0.15)	(-0.62)
2	0.07	0.20	0.06	-0.35	-0.55	-0.62
	(0.66)	(2.16)	(0.52)	(-2.62)	(-1.17)	(-1.37)
3	-0.08	-0.21	0.14	-0.14	-0.88	-0.80
	(-0.63)	(-2.13)	(0.95)	(-0.48)	(-2.03)	(-1.80)
4	0.17	-0.18	-0.28	-0.35	-1.07	-1.25
	(1.18)	(-1.82)	(-2.15)	(-1.33)	(-2.39)	(-2.59)
5 Positive Correlation	0.18	-0.06	-0.25	-0.22	-1.62	-1.80
	(1.33)	(-0.44)	(-1.52)	(-0.55)	(-2.09)	(-2.07)

Figure A1 – Power law coefficients: Edgar search and Ravenpack News

These figures present pairwise histograms of power law coefficient distributions. Linear power law equations are estimated over 60 months from 2011-2015 for the EDGAR search data in Panel A, and for the 75 months from January, 2011 to March, 2017 for the Ravenpack news data in Panel B. The figures show the distribution of coefficients and the sample average.

Panel A. – Edgar Search Attention versus Size.



Panel B. – News versus Size

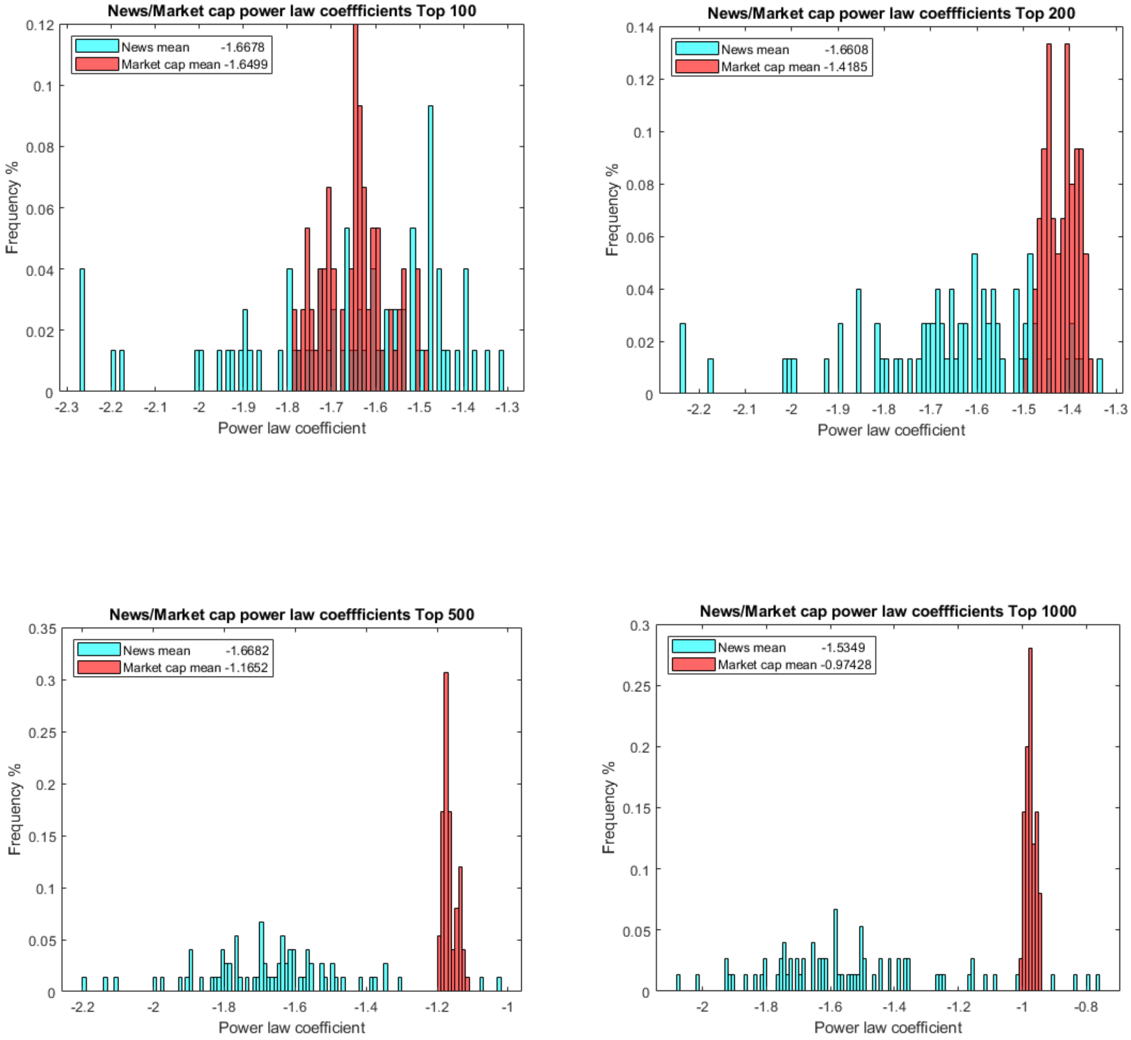


Table A1

This table presents summary information on the Edgar search data set from January 2007 to December 2015. *Total Searches* is the number of single-stock searches. *Average/Stock* and *Max* are monthly averages and maximums within the relevant year. Number of stocks is the total number of different stocks covered by the relevant data each year.

Year	Total Searches	Average/ Stock	Max	Number Stocks
2007	7,191,086	127.3	20,408	4,471
2008	9,556,532	173.6	18,951	4,200
2009	14,980,007	288.7	44,904	3,978
2010	17,308,764	339.4	45,084	3,827
2011	20,164,200	407.6	71,065	3,693
2012	19,863,850	403.1	317,983	3,601
2013	24,984,371	509.7	165,360	3,594
2014	31,499,368	613.9	82,630	3,696
2015	32,662,749	642.9	140,494	3,659

Table A2

This table presents average power law coefficients across five sets of stocks, Stocks are ranked by their rank with respect to the variable in question.

Panel A. Edgar 2011-2015

	Edgar				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-1.790	-1.657	-1.613	-1.649	-1.649
Standard Deviation	0.278	0.156	0.113	0.096	0.083
Mean R-Square %	97.0	97.7	98.6	99.3	99.6
N=	60	60	60	60	60

Panel B. Edgar 2007-2015

	Edgar				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-1.920	-1.768	-1.707	-1.721	-1.706
Standard Deviation	0.305	0.207	0.165	0.129	0.102
Mean R-Square	96.9	97.6	98.5	99.3	99.6
N=	108	108	108	108	108

Panel C. News

	News				
Number of Stocks	50	100	200	500	1,000
Mean Coefficient	-1.774	-1.668	-1.661	-1.668	-1.534
Standard Deviation	0.295	0.225	0.190	0.204	0.279
Mean R-Square	97.0	98.1	98.9	99.2	97.7
N=	75	75	75	75	75

Table A3

This table presents own and pairwise correlations for five different sets of stocks. In Panel A *Frequency* is the average number of stocks in month *t+1* that were in the sample in month t. *Rank correlation* is the correlation between the rank of stocks in month t, and the rank of stocks in month *t+1*. Panel B presents the average rank of the stocks in month *t-1* that become Top Rank stocks in month *t+1*. Panel C presents *Frequency* and *Rank correlation* statistics for pairwise comparisons at time *t*. Panel D presents the cross-correlation between EDGAR search and market capitalization when size is measured as institutional ownership only, or retail ownership only.

Panel A. Average time-series correlations

	Size			**Edgar Search Attention**			**Volume**		
			Rank			Rank			Rank
Stocks	Frequency	Percent	Correlation	Frequency	Percent	Correlation	Frequency	Percent	Correlation
50	48.21	96.42%	0.989	40.47	80.94%	0.784	40.19	80.38%	0.790
100	97.34	97.34%	0.983	76.69	76.69%	0.834	81.92	81.92%	0.813
200	194.45	97.23%	0.991	148.98	74.49%	0.806	165.83	82.92%	0.831
500	488.57	97.71%	0.993	380.86	76.17%	0.757	425.67	85.13%	0.841
1,000	977.42	97.74%	0.994	784.86	78.49%	0.750	878.43	87.84%	0.863

Panel B: Replacement Stocks

	Size		**Edgar Search Attention**		**Volume**	
	Replacement	Percent	Replacement	Percent	Replacement	Percent
Stocks	Rank	Minimum	Rank	Minimum	Rank	Minimum
50	26.70	104.7%	72.03	282.5%	41.96	164.5%
100	51.77	102.5%	113.87	225.5%	83.50	165.3%
200	102.61	102.1%	199.37	198.4%	157.58	156.8%
500	254.20	101.5%	387.19	154.6%	330.54	132.0%
1,000	506.13	101.1%	668.03	133.5%	585.06	116.9%

Panel C: Average cross-sectional correlations –2011 to 2015 only

	Size-Edgar Search			**Edgar Search-Volume**			**Size-Volume**		
			Rank			Rank			Rank
Stocks	Frequency	Percent	Correlation	Frequency	Percent	Correlation	Frequency	Percent	Correlation
50	25.07	50.14%	0.345	20.83	41.66%	0.175	17.53	35.06%	0.048
100	47.62	47.62%	0.469	40.02	40.02%	0.409	37.22	37.22%	0.318
200	86.37	43.19%	0.529	87.71	43.86%	0.375	82.62	41.31%	0.328
500	246.42	49.28%	0.478	253.15	50.63%	0.404	280.13	56.03%	0.323
1,000	531.05	53.11%	0.473	527.70	52.77%	0.456	655.35	65.54%	0.387

Panel D: Institutional and retail ownership.

Institutional	**Size-Edgar Search**			**Retail**	**Size-Edgar Search**		
			Rank				Rank
Stocks	Frequency	Percent	Correlation	Stocks	Frequency	Percent	Correlation
50	23.83	47.66%	0.305	50	26.80	53.60%	0.381
100	46.28	46.28%	0.405	100	46.33	46.33%	0.548
200	84.82	42.41%	0.518	200	86.58	43.29%	0.598
500	230.68	46.14%	0.580	500	226.70	45.34%	0.487
1000	486.6	48.66%	0.501	1000	467.83	46.78%	0.501