Catching Fire: the Diffusion of Retail Attention on Twitter*

Nitesh Chawla, Zhi Da, Jian Xu, and Mao Ye
August 2015

Abstract

Increasingly, investor attention is triggered by communication on social networks. We track the diffusion of retail attention to specific financial news in real time by monitoring how the news is retweeted on Twitter. Using a unique de-identified brokerage account dataset from TD Ameritrade (provided as part of an academic sharing effort), we find the diffusion of retail attention to be highly correlated with intraday retail trading patterns, especially among investors with large stock holdings. The resulting retail attention leads to lower bid–ask spreads and positive price pressure on the news day that is completely reverted the next day. The amount of retail attention the news generates on Twitter can be predicted using characteristics of the users, accounts and tweets. The fact that predicted retail attention generates similar results alleviates concerns about reverse causality and endogeneity.

Key Words: Retail attention, Diffusion, Price Pressure, Social Media

* Chawla: University of Notre Dame, Department of Computer Science & Engineering, Notre Dame, IN, 46556, USA. E-mail: nchawla@nd.edu. Da: University of Notre Dame, Mendoza College of Business, Notre Dame, IN, 46556, USA. E-mail: zda@nd.edu. Xu: University of Notre Dame, Department of Computer Science and Engineering, IN, 46556, USA. E-mail: i@jianxu.net. Ye: University of Illinois at Urbana-Champaign, College of Business, Champaign, IL, 61820, USA. E-mail: maoye@illinois.edu. We thank Robert Battalio, Tim Burch, Shane Corwin, David Hirshleifer, Thierry Foucault, Pengjie Gao, Tim Loughran, Rik Sen, Paul Tetlock, Charles Trzcinka, Bonnie Van Ness, and seminar participants at the University of Miami, the University of Notre Dame, the NFA 2014 meeting, the MFA 2015 meeting, the Research in Behavioral Finance Conference at Erasmus, the Central Bank Workshop on Market Microstructure in Rome for helpful comments. We thank Sam Achelpohl for access to TD Ameritrade data through their academic collaboration program. We thank Zi Chao and Mehrnoush Shahhosseini for excellent research support. Ye acknowledges the support of the National Science Foundation. This work also uses the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. We thank Robert Sinkovits of the San Diego Supercomputer Center for assistance with supercomputing, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program. All errors are our own.

1. Introduction

Even with today's information dissemination technology, news does not grab everyone's attention at the same time. Instead, it diffuses gradually across networks of investors.[1] An important reason for such gradual diffusion is limited investor attention.[2] Attention constraints are more likely to affect retail investors. Since retail investors rarely short stocks, news that commands their attention will on average lead to retail purchases and positive price pressure, as argued by Barber and Odean (2008). Da, Engelberg, and Gao (2011) find supporting evidence for such price pressure using weekly search frequency with Google as a direct measure of retail attention, but this is a static measure. To the best of our knowledge, the literature has not produced a direct way to measure how one agent's attention is triggered by another's, even though such transmission plays a key role in theoretical diffusion in asset pricing models.

In this paper we measure the diffusion of retail attention by tracking how financial news on Twitter is retweeted during trading hours. Our approach offers several advantages. First, high-frequency analysis provides a more powerful testing environment for a theory of retail attention. For example, the strong pricing overshooting and subsequent reversals on a daily frequency we document is more consistent with retail-attention-triggered price pressure rather than alternative explanations based on low-frequency discount rate variations. Second, we are able to link retail attention to specific news, media outlets, Twitter users, and intraday measures of retail trading, thus shedding new light on the drivers of retail attention. Finally, by borrowing insights from the computer science literature on cascades in social networks, we find that the amount of retail attention the news generates on Twitter can be predicted. The predicted retail attention can then be used as an instrument to alleviate concerns about reverse causality and endogeneity.

[1] Examples include Shiller (1984), Hong and Stein (1999), Shive (2010), Hong, Hong, and Ungureanu (2012), Han and Yang (2013), and Manela (2014).
[2] See Kahneman (1973), Merton (1987), Huberman and Regev (2001), Sims (2003), Hirshleifer and Teoh (2003), Peng and Xiong (2006), and Hong and Stein (2007), among others.

With monthly averages of 288 million active users and 500 million tweets (text messages of no more than 140 characters) posted per day in Q4'14, Twitter is one of the biggest online social media outlets in the world.[3] Thanks to its broad audience, Twitter registered 182 billion timeline views (collections of users' news feeds) in Q4'14, making it one of the ten most-visited websites in the world.[4] News that triggers Twitter users' attention via retweets spreads like wildfire. For example, when Paul Walker's staff tweeted the passing of the actor on December 1, 2013, the tweet was retweeted more than 400,000 times as the world came together in mourning. Importantly, by tracking the total number of such retweets and the number of followers for each retweeting account at intraday frequencies, we can directly measure the diffusion of retail attention on specific news over time. Since Twitter users usually choose to retweet financial news that they consider to be of interest to their followers, it stands to reason that the retail attention we capture via retweets is likely to generate trading. We then align the retweet data with various intraday trading data to study the impact of retail attention on retail trading volume, asset prices, and liquidity.

During our one-year sample period of 2013/11 through 2014/10, we track the complete history of tweets and retweets posted by 78 major media outlets and famous analysts (e.g., @WSJ and @CNBC), 56 active Twitter accounts of S&P1500 CEOs and CFOs (e.g., Elon Musk, CEO & Chief Product Architect of Tesla Motors), and 143 official Twitter accounts of S&P500 companies (e.g. @TysonFoods). We focus on tweets about companies in the Russell 3000 index. We keep track of a given tweet's entire retweet history, including the number of retweets and various characteristics of the retweet account and the original tweet itself. Our final sample contains 1,261 tweets from 115 Twitter accounts covering 178 distinct stocks. The tweets in our sample cover a wide range of news (mergers and acquisitions, earnings announcements, product launch, independent research, etc.).

While the total number of retweets in our sample at the end of the first hour is only 68 on average, the number of potential investors the original tweet can reach is much higher, more than 3 million people

[3] https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321
[4] http://www.alexa.com/topsites

on average, because many Twitter accounts we tracked have thousands or even millions of followers. Importantly, we observe large cross-sectional variations in attention diffusion speed across the tweets in our sample.

We first link attention diffusion speed to trading intensity during the first hour after an initial tweet. Attention diffusion speed is measured as the percentage of first-hour retweets occurring during the first 10 minutes. Likewise, trading intensity is measured as the percentage of first-hour trading that takes place during the first 10 minutes. We find a very strong link between retweet speed and trading intensity: tweets that are retweeted quickly are correlated with immediate trading, even after controlling for time of day, recent turnover and volatility, past and contemporaneous returns, and various fixed effects. For example, a 1% increase in diffusion speed is associated with a 0.14% increase in trading intensity when trading is measured using total volume.

The link between attention diffusion speed and retail trading intensity is 50% greater than the link to total trading intensity. We first proxy intraday retail trading using TAQ volume from Trade Reporting Facilities (TRFs). A recent paper by Battalio, Corwin, and Jennings (2015) finds that some brokers tend to route retail orders to market makers but not to public exchanges, and that 5 out of 10 popular retail brokers route all their non-direct limit orders and market orders to market makers.[5] These trades are reported to TRFs with an exchange symbol D in the TAQ dataset. Trading in TRFs is the best measure we can find using publically available data. We find that a 1% increase in diffusion speed is associated with a 0.21% increase in retail trading intensity after controlling for other factors. The strong link between retweets and trading intensity, especially retail trading intensity, supports the notion that counting retweets is a good measure of the diffusion of retail attention in general.

To measure retail trading even more directly, we take advantage of the unique brokerage account dataset from TD Ameritrade (TDA) that records 331 million de-identified transactions made by 2.8 million clients from June 1, 2010 through June 10, 2014. We focus on a smaller merged sample during

[5] TD Ameritrade is not one of the five identified in Battalio, Corwin, and Jennings (2015), who find that TD Ameritrade routes orders to stock exchanges.

the overlapping period from 2013/11 through 2014/06 that covers 331 tweets and 35,443 distinct individual TDA accounts that trade the corresponding stock at least once during the first three hours after an initial tweet.[6]

Our analysis using the TDA merged sample confirms an even stronger link between retweet speed and retail trading intensity. Not surprisingly, the link is strongest among TDA investors who have greater stock holdings in their accounts. They have a stronger financial incentive to be attentive to financial news. Interestingly, the link is weakest among TDA investors who are under 35, possibly because they invest less on average and are less likely to trade immediately following a tweet due to their work responsibilities. The link is slightly stronger among female TDA investors. Overall, the TDA analysis provides direct support for the proposition that attention diffusion speed measured using retweets is strongly related to retail trading and thus serves as a good measure of retail attention at high frequency.

We then examine how retail attention diffusion affects asset prices and stock liquidity after an initial tweet. We measure the level of retail attention triggered by the tweet using the total number of Twitter users the tweet can reach after three hours. In computing this number, we account for the number of followers of each Twitter user who retweets. In other words, if an influential Twitter user with 5,000 followers retweets, the number of Twitter users the tweet can reach will go up by 5,000.

We find a strong positive contemporaneous relationship between attention diffusion and stock returns. The more users a tweet reaches after three hours, the higher the stock returns on that day (from 10 minutes after the tweet to the end of the day), but the higher returns is completely reverted on the next day. The price overshooting and reversal pattern provides unique support for the findings of Barber and Odean (2008), who argue that retail attention leads to positive price pressure on average since retail investors rarely short stocks. As further support for the retail trading interpretation, we find a greater decrease in the bid–ask spread for stocks whose tweets reach more users. The spread decrease is consistent with lower adverse selection risk as retail trading picks up. In contrast to existing low-frequency attention measures of a stock, our high-frequency news-specific measure provides a tighter link between retail attention and

[6] The identity information for these individual accounts is masked for confidentiality.

price pressure and is more powerful in detecting the subsequent near-future reversal of the effects. Not surprisingly, the price overshooting and reversal pattern is driven mostly by the smaller stocks in our sample.

Among the 1,261 tweets in our sample, there is a wide variation in attention diffusion rates. An emerging body of computer science literature has studied why identical content may diffuse at very different speeds on the same social network and found several reasons based on network characteristics (see Cheng et. al. (2014), Jenders et. al. (2013), Petrovic et. al. (2011) and Suh et. al. (2010), among others). Borrowing insights from this literature, we conduct a predictive exercise wherein we try to predict future attention diffusion rates on Twitter using information observable 10 minutes after an initial tweet. Specifically, the future attention diffusion rate is defined as the growth rate in the total number of Twitter users the tweet can reach from 10 minutes after the tweet to three hours after the tweet.

We find that the attention diffusion rate is driven by several tweet-related characteristics, and our findings are consistent with our intuitions. For example, if a tweet is retweeted in the first 10 minutes by users with more followers, the tweet will diffuse more rapidly afterwards. If the tweet comes from an active Twitter account that posts many new tweets per day, its attention diffusion rate will be lower as multiple tweets from the same account compete with each other for retail attention. Tweets with pictures or hashtags diffuse more rapidly as they grab users' attention. On the other hand, a tweet with a URL link is predicted to diffuse more slowly as it takes time to read the linked article. Finally, tweets sent out earlier during the day or from the West Coast seem to diffuse more quickly. Since these characteristics are independent of future returns and not directly related to trading or valuation, we use them to instrument our retail attention measure.

Specifically, we compute the predicted attention diffusion rate using Twitter characteristics. We then multiply the predicted diffusion rate by the total number of Twitter users the tweet can reach after 10 minutes and use this product in our analysis. We find the same price overshooting and reversal pattern that we found in our other analyses. Interestingly, we do not detect this pattern when using only the total number of Twitter users a tweet can reach after 10 minutes. In other words, the predicted attention diffusion after the first 10 minutes is crucial for correctly measuring the total retail attention the tweet can generate during that day.

Finally, we conduct a predictive out-of-sample exercise. We use only data from the first six months of our sample period (2013/11 through 2014/04) to run the predictive regression and then apply the regression coefficients to the next six months (2014/05 through 2014/10) in computing the attention diffusion rate. The predicted retail attention measure is therefore free of forward-looking bias and can be computed in real time. We then detect positive price pressure and the subsequent reversal of the retweeting effects for the second half of our sample.

Our paper also adds to a growing body of literature that relates news and media activity to investor attention and asset pricing (see Tetlock (2007), Cohen and Frazzini (2008), Corwin and Coughenour (2008), Fang and Peress (2009), Loughran and McDonald (2010 and 2014), Da, Engelberg and Gao (2011), Engelberg and Parsons (2011), Gurun and Butler (2012), Agarwal, Kumar, Leung, and Konana (2014), Peress (2014), and Peress and Schmidt (2014), among others). While most existing papers focus on static and low-frequency investor attention measures, we examine the dynamic diffusion of retail attention and its impact on asset pricing at high frequency.

Our paper is also related to a broad body of literature that studies how information is incorporated into pricing. Most prior studies focus on learning from pricing or trading (Grossman and Stiglitz (1980), Kyle, (1985) and Glosten and Milgrom (1985)). However, learning also comes from more direct channels such as diffusion. Hong and Stein (1999) and Manela (2014) show that *slow* information diffusion can affect asset pricing within a monthly horizon.[7] As Twitter becomes a more and more popular outlet for breaking financial news in the future, we can also track retweets of such "breaking" news as a direct way of measuring the diffusion of genuine information.

[7] If we count research without specific diffusion models, there are even larger bodies of evidence that suggest that public information diffuses gradually through the investor population and that this gradual diffusion affects prices. See, e.g., Peress , 2014; Chan, 2003; Hou, Peng and Xiong, 2006; Cohen and Frazzini, 2008; Hirshleifer, Lim and Teoh, 2009; and Peress, 2008.

Our paper also speaks to the emerging literature that studies social influence, communication, and information sharing on a network (see DeMarzo, Vayanos, and Zweibel (2003), Ivkovich and Weisbenner (2007), Brown, Ivkovich, Smith, and Weisbenner (2008), Colla and Mele (2010), Ozsoylev and Walden (2011), Han and Hirshleifer (2012) and Bildik, Ozsoylev, Walden, and Yavuz (2013), among others). We provide an example wherein characteristics of the social network help predict future information diffusion speed.

Finally, our paper adds to the recent literature that uses information from Twitter. Bollen et al. (2011) apply textual analysis of tweets to gauge investor sentiment. Blankespoor et al. (2014) find that companies using Twitter to disseminate information experience improvement in stock liquidity. Cheng et al. (2014) examines the impact of having a Twitter-active CEO. Bhagwat and Burch (2014) examine firms' tweeting behaviors around earnings announcements. Giannini et al. (2013) measure investor opinion using Twitter messages on Stocktwits.com. Our paper is the first to focus on retweets, which allows us to trace out the dynamic nature of retail attention diffusion.

The rest of the paper is organized as follows. Section 2 describes the data. Section 3 contains the empirical analysis. Section 4 concludes.

2. Data description

Using Twitter's Streaming API,[8] we track the complete history of tweets and retweets of 78 major media outlets and famous analysts (e.g., @WSJ and @CNBC), 56 active Twitter accounts of S&P 1500 CEOs and CFOs (e.g., Elon Musk, CEO & Chief Product Architect of Tesla Motors),[9] and 143 official Twitter accounts of S&P 500 companies (e.g. @TysonFoods). Having well-established reputations and being updated frequently, these accounts deliver news that is likely to attract retail investor attention. The data collection spans a one-year period, from November 1, 2013 through October 31, 2014. For this time

[8] https://dev.twitter.com/docs/api/1.1/post/statuses/filter
[9] See Chen, Hwang, and Liu (2013).

period, we captured any original tweets posted by any of these 277 accounts and all retweets of any tweets they posted.

Table 1 reports the summary statistics for these 277 accounts. The 78 Twitter accounts from media outlets tend to have more followers, with a mean of 888,545 and a median of 100,446. For example, @nytimes has more than 11 million followers and @WSJ has more than 4 million followers. The 143 official Twitter accounts of S&P 500 companies also have many followers, with a mean of 601,931 and a median of 125,521. Both @Google and @Starbucks have more than 5 million followers apiece. Firm CEOs and CFOs have fewer followers. The mean is 54,576 and the median is only 621. @ericschmidt, @RalphLauren, and @MichaelDell attract the most followers (779K, 672K, and 628K, respectively).

Insert Table 1 about Here

Table 1 also reports that the average number of years since inception is 5.7 among media outlet accounts, 5.3 among company accounts, and 4.3 among CEO/CFO accounts. In terms of account activity measured by the total number of tweets per year, Twitter accounts maintained by media outlets are the most active with almost 7,488 tweets per year per account, followed by S&P 500 company accounts (3,334 tweets per year per account). The Twitter accounts of CEOs and CFOs are the least active, with only 264 tweets per year per account.

From the collected tweets, we then identify those meeting the following conditions:

1) Having been retweeted more than 50 times.
2) Having been posted during extended trading hours (4:00 a.m. to 8:00 p.m. EST).
3) Mentioning at least one company that is in the Russell 3000 index.
4) Being about the most important event for each company on each day: if multiple events happened to the same company, we pick the most important one; if multiple tweets about the same event are captured, we pick the one that was published the earliest.

In the selection process described above, steps (1) and (2) were carried out automatically by a computer script; steps (3) and (4), however, needed to be performed manually. The selection process leaves us with 1261 tweets. These tweets originate from 115 Twitter accounts and cover 178 distinct stocks. Table 2 contains a few sample tweets we analyzed in the paper. They cover a wide range of news (mergers and acquisitions, earnings announcements, product launches, independent research, etc.) from the Twitter accounts we track.

Insert Table 2 about Here

Of the 115 distinct Twitter accounts, @WSJ generates the most tweets in our sample (270), followed by @Forbes (129), and @CNBC (83). Of the 178 distinct stocks, Apple (AAPL) appears most frequently (92 times), followed by Facebook (FB, 88 times), Google (GOOG, 82 times), Twitter (TWTR, 81 times), Microsoft (MSFT, 67 times), and Tesla (TSLA, 64 times). Table 3 presents summary statistics on the stocks in our sample. The average stock size is at the 90th percentile of the CRSP universe. The average institutional ownership is also large, at 57.7%, corresponding to the 80th percentile of the CRSP universe. The volatility of the average stock in our sample is similar to that of an average stock in the CRSP universe but trades with slightly greater frequency.

Insert Table 3 about Here

Figure 1 summarizes our proxy for the diffusion of retail attention over time. The left figure plots the total number of retweets over time during the first hour after an original tweet, the total number in the median case of retweets, the total number in the 5th percentile of retweets, and the total number in the 95th percentile of retweets. Each time interval represents 10 minutes. On average, a tweet in our sample will be retweeted 68 times by the end of the first hour. The 5th percentile is 23 retweets while the 95th percentile is 235 retweets. The small number of 68 retweets, however, reaches 3 million more people, because many accounts that retweet the news also have a large number of followers.

Insert Figure 1 about Here

The intraday trading data are taken from the NYSE Daily Trade and Quote (DTAQ) database to construct the complete NBBO quotes. DTAQ provides two files that contain official NBBO quotes. If a single exchange has both the best bid and the best offer, then the official NBBO quotes will be recorded in the DTAQ Quotes File. Otherwise, the NBBO quotes will be recorded in the DTAQ NBBO file. Following the procedure proposed by Holden and Jacobsen (2014), we combined the NBBO quotes from both files to construct the Complete Official NBBO. We exclude quotes with abnormal quote conditions (A, B, H, O, R, and W). We delete any quote with a bid that is greater than or equal to the ask. We also delete cases in which the quoted spread is greater than $5.00. We then compute bid–ask spreads and intraday returns using midpoints.

As market-wide intraday retail trading volume data is not directly available, we base our analysis on a market-wide proxy for retail trading volume using the trading volume from the TRFs, and the results are supplemented by a proprietary dataset on retail trading from TD Ameritrade. The market-wide proxy is constructed based on the empirical finding of Battalio, Corwin, and Jennings (2015) that non-direct limit and market orders are seldom routed to public exchanges but are often internalized by broker–dealers. Therefore, we use TRF volume (exchange symbol D in the TAQ dataset) as our proxy for market-wide retail trading. We are aware that this measure has two limitations. First, TRF volume also contains volume from dark pools (Kwan, Masulis and McInish (2014)). Second, Battalio, Corwin, and Jennings (2014) find that some retail brokers route orders to public exchanges, including TD Ameritrade. Therefore, we supplement our market-wide proxy of retail trading with a proprietary dataset from TD Ameritrade. This dataset records 331 million de-identified transactions made by 2.8 million clients from June 1, 2010 through June 10, 2014.

3. Empirical Analysis

With a direct measure of the diffusion of retail attention at high frequency, we ask three empirical questions. First, how does the diffusion of retail attention affect trading intensity, especially among retail investors? Second, how does the resulting retail attention affect asset prices and stock liquidity? Finally, can the attention diffusion rate be predicted?

3.1 Retweet speed and trading intensity

We first examine how the attention diffusion rate, as measured by retweeting speed, is related to trading intensity. The graphic illustration, Figure 2, shows cumulative numbers of retweets and total trading volumes for each of the six 10-minute intervals during the first hour after a tweet. Both variables are normalized by their totals during the first hour. By construction, the cumulative retweet rate and trading volume, after normalization, will all be one at the end of the first hour, as it would be in a cumulative distribution function (CDF). Such a normalization is motivated by the theoretical literature on information diffusion. The driver of the asset-pricing dynamics in information diffusion models is the proportion of agents who know the information earlier than others, which is characterize by the CDF function (Hong, Hong, and Ungureanu (2012)). Yet the empirical literature does not include a dynamic proxy for the CDF function, and our paper fills this gap. We classify a diffusion process as fast if more than 60% of total first-hour retweets occur in the first 10 minutes, and we classify a diffusion process as slow if less than 40% of total first-hour retweets occur in the first 10 minutes.

Insert Figure 2 about Here

We find that relatively more trades by volume take place early when information diffuses faster. For example, 25.0% of the first-hour trades take place in the first 10 minutes for the fast diffusion case. In contrast, if only a small fraction of retail investors paid attention to the news in the first 10 minutes, there would not be a lot of trading. Indeed, only 13.4% of the first-hour trading takes place in the first 10 minutes for the slow diffusion case. In addition, trading intensity seems to lag behind the diffusion rate by about 10 minutes, possibly due to the time required by retail investors to act after the news triggers their attention.

There are many reasons to explain the observed correlation between the diffusion rate and trading intensity. The correlation may be driven by the time of day. For example, a tweet posted in the early morning may not immediately generate retweets, which would reflect the lack of trading during that time. The correlation may also be driven by a common factor. For example, extreme returns immediately following a tweet could trigger both retweets and trading. The panel regression presented in Table 4 Panel A controls for these factors.

The dependent variable, the percentage of first-hour total trading that occurs in the first 10 minutes, measures trading intensity. The main independent variable, diffusion, measures the percentage of first-hour retweets that occurs in the first 10 minutes. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); abs past 1h ret (absolute stock returns over the market in the past hour); abs 10m ret (absolute stock returns over the market in the first 10 minutes after the tweet). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. The sample covers 1,261 tweets during one year from 2013/11 through 2014/10.

Insert Table 4 about Here

The first two columns in Panel A confirm the strong unconditional correlation between diffusion rate and trading intensity observed in Figure 2. A 1% increase in the diffusion rate leads to a 0.3% increase in trading intensity with a t-value of 5.13. Once we control for time of day in columns 3 and 4, the effect attenuates to 0.17% but is still highly significant. Finally, columns 5 and 6 also control for absolute returns around the tweet and other stock characteristics. The coefficient on the diffusion rate falls to 0.14 but remains significant (t-value = 2.25).

To the extent that Twitter users are more likely to be retail investors, we would expect to see an even stronger link between retweet speed and trading intensity for retail investors. While retail investor trading volume is not directly available, we focus on volumes from the TRFs, which are believed to be more highly correlated with retail investors. We then compute trading intensity using only TRF volume and use it as the dependent variable in the regressions. We report the results in Panel B of Table 4.

We find the link between attention diffusion speed and retail trading intensity to be 50% stronger for retail investors than for all investors. A 1% increase in diffusion speed is associated with a 0.21% increase in retail trading intensity, after controlling for other factors. The strong link between retweets and trading intensity, especially retail trading intensity, supports the notion that retweets on Twitter provide a good general measure of the diffusion of retail attention.

3.2 Evidence from brokerage account data

Even TRF volume is a noisy measure of retail trading. To get at retail trading more directly, we take advantage of a unique brokerage account dataset from TDA that records 331 million transactions made by 2.8 million clients from June 1, 2010 through June 10, 2014. The data are provided generously by TDA through an academic data collaborative agreement. The TDA dataset does not identify individual clients, but it includes demographic characteristics such as age and gender for each anonymous ID. We are also able to track the history of trading from an individual through this unique ID. While trades in the TDA data represent a subset of all trades, it is a relatively clean subset of retail trading.

We merge our tweet sample with TDA brokerage-account-level transaction data during the overlapping period from 2013/11 through 2014/06.[10] We focus only on stock trades from "individual" accounts in TDA.[11] Since investors in TDA rarely trade during after-hour sessions, we focus on tweets during market trading hours (9:30 a.m. to 4:00 p.m. EST). We examine only accounts that trade the corresponding stock at least once during the first three hours after a tweet. Altogether, these selection criteria result in our merged dataset that contains 331 tweets and trades from 35,443 individual TDA accounts. Panel A of Table 5 provides summary statistics on this merged sample.

Insert Table 5 about Here

The top half of Panel A reports the characteristics averaged across the 35,443 distinct TDA accounts in the merged sample. The average account holder is about 49 years old in 2014. The first age quartile is 38 and the third quartile is 58. Their median stock holdings with TDA are worth $20K with 25% holding stocks worth more than $74K. When they trade during the first three hours after a tweet, they are more likely to buy. Both the mean and median of the net trade variable are positive (with t-value > 5). This finding provides direct support for Barber and Odean (2008), who argue that retail attention leads to

[10] No effort was made to cross- reference TDA accounts with Twitter accounts. The data set enables us to compare the behavior of TDA traders and Twitter users, but cannot indicate whether an individual trader did or did not have access to Twitter.
[11] TDA also records trading in options, bonds, warrants, mutual funds and other securities and these transactions represent less than 30% of all trades. Other account types include "Joint Tenants WROS," "IRA," "Rollover IRA," "Trust," "Roth IRA," etc. Trades from "individual" accounts represent almost half of all trades.

positive price pressure on average since retail investors are less likely to short. On average, 20% of all first-hour trades take place during the first 10 minutes following the tweet.

The bottom half of Panel A reports the characteristics averaged across the 331 tweets in the merged sample. On average, we observe trades from almost 200 TDA accounts following a tweet in the merged sample. Seventy-two percent of the trades come from male account holders. The average account holder is about 51 years old, holding about $91K in stocks, and is more likely to buy stocks (rather than sell them). Twenty percent of all first-hour trades take place during the first 10 minutes following a tweet.

Panel B then repeats the regressions in Table 4 for our merged sample of tweets. We measure trading intensity using (I) all trades from TAQ; (II) all TDA trades; (III) all TDA trades of female investors; (IV) all TDA trades of young investors (age <35); and (V) all TDA trades of "rich" investors (whose stock holding is greater than $100K). We include the same set of control variables as those in Table 4. Since we focus on tweets during the trading hour, the pre-market and post-market dummies drop out.

The results of regression I confirm the strong correlation between diffusion speed and trading intensity measured using all trades, in this smaller merged sample of 331 tweets. Even with other controls, a 1% increase in the diffusion rate leads to a significant 0.16% increase in trading intensity. The coefficient of 0.16 is higher than the corresponding coefficient of 0.14 using the full sample (see Panel A of Table 4), possibly because we focus on trading hours.

The results of regression II suggest a much stronger link between diffusion speed and retail trading intensity measured using all TDA trades. The coefficient on the diffusion variable increases from 0.16 to 0.23. The results of regression III suggest an even stronger link between diffusion speed and retail trading intensity among female investors, who account for less than 30% of all TDA investors.

The strongest link between diffusion speed and retail trading intensity is found among the 20% of TDA investors with the largest stock holdings. For them, a 1% increase in the diffusion rate leads to a 0.36% increase in trading intensity. This is not surprising because traders with higher investment in stocks should be more attentive to financial news and thus react more quickly to that news.

Interestingly, the weakest link between diffusion speed and retail trading intensity is found among the 20% of TDA investors who are younger than 35. For them, a 1% increase in the diffusion rate leads to only a 0.18% increase in trading intensity, for two possible reasons. First, younger investors have fewer financial resources and therefore fewer stock investment assets and lower investment value, which means they may be constrained by commissions or other fixed transaction costs. Attention is one such cost. Investors with less valuable investments may have a weaker incentive to follow a particular firm. Second, compared with average TDA investors, who are close to retirement age, younger investors have to focus more attention on work during trading hours and thus may not trade immediately after a tweet.

Overall, the TDA data provide direct support for the idea that attention diffusion speed measured using retweets is strongly related to retail trading.

3.3 Diffusion of retail attention, stock returns and liquidity

Insofar as attention diffusion speed measured using retweets seems to be related to retail trading intensity, we then examine how it affects prices and dollar bid–ask spreads.

We measure contemporaneous stock returns (in excess of the market) from 10 minutes after an initial tweet until the end of the day (labeled as *CAR%[10m, close d0]*). We skip the first 10 minutes after the tweet for two reasons. First, for tweets in our sample that represent genuine breaking news, most of the information should have been incorporated into prices by the end of the first ten minutes. Subsequent returns are more likely to capture attention-driven price pressures. Second, we predict future attention diffusion using information observable in the first 10 minutes and then link predicted attention to stock returns. Skipping the first 10 minutes in the returns measurement thus avoids mechanical correlation. We also measure stock returns (in excess of the market) on the next trading day (labeled as *CAR%[close d0, close d1]*). We then examine the change in stock liquidity as the average dollar bid–ask spread during the three hours after a tweet minus the average dollar bid–ask spread during the hour before the tweet.

We measure the level of retail attention triggered by a tweet using the total number of Twitter users the tweet can reach after three hours (labeled as *diffusion_3hr*). In computing this number, we account for the number of followers of Twitter users who retweet. In other words, if an influential Twitter user with 5,000 followers retweets, the number of Twitter users the tweet can reach will increase by 5,000. Focusing on a three-hour horizon makes the measure comparable across tweets. We find similar results when we measure the level of retail attention until the end of the day, as most of the tweets take place during the first three hours.

We then regress contemporaneous stock returns, future stock returns, and the dollar spread change on the level of retail attention triggered by a tweet in panel regressions. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); past 1h ret (stock returns over the market in the past hour); abs past 1h ret (absolute value of past 1h ret); 10m ret (stock returns over the market in the first 10 minutes after the tweet); abs 10m ret (absolute value of 10m ret); and Isbreaking (a dummy variable equal to 1 if the tweet contains "breaking"). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. The results are reported in Table 6.

Insert Table 6 about Here

Panel A reports the regression results from the full sample. We observe a positive and significant association between retail attention diffusion and contemporaneous returns. A one-standard-deviation increase in our retail attention measures (*diffusion_3hr*) leads to a 23-basis-point increase in contemporaneous-day returns.[12] Interestingly, the higher returns seem to vanish revert completely the next day. Such temporary price overshooting and subsequent reversal is consistent with the retail-attention-triggered price pressure found by Barber and Odean (2008). When retail investors are buying, they can choose from a large set of available alternatives. However, when they are selling, they can sell only what they own since they rarely short sell. This means that an increase in retail attention should lead, on average, to net buying by these uninformed traders.

[12] We multiply the coefficient on *diffusion_3hr* (21 bps) by its standard deviation (1.09).

Our high-frequency news-specific retail attention measure gives us more power to detect such price pressure and the subsequent reversal of the effects of a tweet. For example, Da, Engelberg, and Gao (2011), using the weekly Google search frequency on stock tickers, detect significant price pressure, of a similar magnitude, during the first two weeks, but cannot detect a statistically significant reversal in the long run due to a short sample. The low-frequency analysis in prior literature is also vulnerable to alternative risk-based interpretations. In contrast, the price overshooting and reversal we detect at daily frequency is unlikely to be driven by time-varying discount rates.

When more retail investors are trading once their attention gets triggered, the adverse selection risk decreases, and we therefore expect a reduction in the bid–ask spread. This is exactly what we find. We observe a negative and significant link between retail attention diffusion and the change in the dollar spread.

Not surprisingly, the return and liquidity results are much more pronounced among smaller stocks in our sample (whose market cap is below the median), as reported in Panel B. A one-standard-deviation increase in our retail attention measures (*diffusion_3hr*) leads to a much higher 46-basis-point increase in contemporaneous-day returns.[13] Again, the price pressure completely reverted the next day. Finally, we observe an even greater decrease in the dollar spread after a tweet among smaller stocks. It is important to note that the average market cap of smaller stocks in our sample is still at the 82th percentile of the CRSP universe, which represents large stocks by traditional measures.

3.4 Predicting future attention diffusion rates

Having established strong links between retail attention diffusion on Twitter, retail trading, price pressure, and stock liquidity, we then examine factors driving cross-sectional variation in attention diffusion. In particular, we are interested in predicting future attention diffusion rates on Twitter.

A large number of computer science studies have developed advanced machine-learning techniques for predicting information cascades on large social networks. Prior work relies mostly on the content and source of a tweet (see Jenders et. al. (2013), Petrovic et. al. (2011) and Suh et. al. (2010),

[13] We multiply the coefficient on *diffusion_3hr* (39 bps) by its standard deviation among smaller stocks (1.17).

among others). Cheng et al. (2014), however, suggest that how information diffuses in the first few minutes (also known as temporal features) and the characteristics of people who have retweeted are also crucial factors in predicting information cascades. We borrow these insights and first conduct a simple predictive regression exercise using the full sample. The results based on more sophisticated machine-learning techniques provide stronger statistical power, and they are available upon request.[14]

The dependent variable of interest is the future attention diffusion rate on Twitter. Specifically, the growth rate is calculated as log(diffusion_3hr) – log(diffusion_10m) where diffusion_10m and diffusion_3hr are the number of Twitter users a tweet potentially reaches after 10 minutes and after three hours, respectively.

The predictive variables are information observed 10 minutes after an initial tweet: Total number of tweets (the average daily number of tweets sent by that Twitter account); log(# of followers of recent retweeters)—the total number of followers in log, of the most recent 5 Twitter accounts that retweeted the tweet; Speed of recent retweets (inverse of the average time lapse between the most recent 5 retweets); Hour (calendar hour of the tweet); IsWest (a dummy variable equal to 1 if the tweet is sent from the West Coast); IsCEO (a dummy variable equal to 1 if the tweet is sent by the CEO of the company); HasPicture (a dummy variable equal to 1 if the tweet contains a picture); HasURLs (a dummy variable equal to 1 if the tweet contains URL links); HasHashtags (a dummy variable equal to 1 if the tweet contains Hashtags). The results from this predictive regression are reported in Table 7.

Insert Table 7 about Here

We find that the attention diffusion rate is driven by tweet-related characteristics in a very intuitive way. For example, if an initial tweet is retweeted quickly by users with more followers, the tweet will diffuse more quickly afterwards. In addition, if recent retweets are posted in rapid-fire fashion, the initial tweet will diffuse faster. The predictive power of these temporal features is consistent with the findings in Cheng et al. (2014).

[14] We have experimented using more sophisticated machine-learning techniques such as Support Vector Machine, neural networks, and decision tree-based algorithms. Nevertheless, we choose to report the simple OLS results as we can see more clearly how each predictive variable is related to the future attention diffusion rate.

If the tweet comes from an active Twitter account with many tweets per day, its attention diffusion rate will be lower because distinct tweets from the same account are competing with each other to grab retail attention, consistent with the "driven-to-distraction" hypothesis of Hirshleifer, Lim, and Teoh (2009).

Tweets with pictures or hashtags diffuse faster because they grab users' attention. On the other hand, a tweet with a URL link should diffuse more slowly as it takes time to read the linked article. Finally, tweets posted earlier in the day or from the West Coast seem to diffuse more quickly.

These predictive variables are all observable during the first ten minutes after an initial tweet, and thus are independent of future returns measured after 10 minutes. In addition, most of these variables are not directly related to the valuation content of the tweet. We therefore use them to instrument our retail attention measure. If the predicted attention measure is still strongly linked to price overshooting and reversal, then our conclusion is less likely to be affected by reverse causality (future return causes future attention) or other endogeneity concerns.

Specifically, we first compute the predicted attention diffusion rate from the predictive regression. We then multiply the predicted diffusion rate by the total number of Twitter users a tweet can reach after 10 minutes and use this product in our analysis. Intuitively, this product measures the expected number of users the tweet can reach using the information set available 10 minutes after the initial tweet. We then link the predictive retail attention to contemporaneous and future stock returns using the same panel regressions that we use for Table 6. The results are reported in Table 8.

Insert Table 8 about Here

We find the same price overshooting and reversal pattern here as in our other analyses. Interestingly, we do not detect this pattern when using only the total number of Twitter users the tweet can reach after 10 minutes. In other words, the predicted attention diffusion after the first 10 minutes is crucial for correctly measuring the total retail attention the tweet can generate during that day.

Finally, we conduct a predictive out-of-sample exercise. We use only data from the first six months of our sample period (2013/11 through 2014/04) to run the predictive regression and then apply the regression coefficients to the next six months (2014/05 through 2014/10) in computing the attention diffusion rate. The predicted retail attention measure is therefore free of forward-looking bias and can be computed in real time. We then link the predictive retail attention to contemporaneous and future stock returns using only the second half of our sample. The results are reported in Table 9.

Insert Table 9 about Here

We find that the predicted retail attention forecasts the positive price pressure and subsequent reversal out of sample, thus providing even stronger supporting evidence that retail attention is causing the temporary price pressure.

4. Conclusion

In this paper, we track the diffusion of retail attention to specific investment-related news in real time by monitoring how such news is retweeted on Twitter. Using a unique set of brokerage-account-level data, we find the diffusion of retail attention to be highly correlated with intraday retail trading patterns. The resulting retail attention leads to lower bid–ask spreads and positive price pressure on the news day, but these effects are completely reverted the next day. The amount of retail attention the news generates on Twitter can be predicted using characteristics of users, accounts, and tweets. The fact that predicted retail attention generates similar results helps to alleviate concerns about reverse causality and endogeneity.

As one of the first studies to construct a dynamic measure of information diffusion, the research we report in our paper can be extended in several ways. First, a more direct test could be designed to test the differential impacts of learning from trading and learning from diffusion using retweets of breaking news. Second, we could examine people with local bias in retweeting insofar as they are more likely to retweet local news. The literature on word of mouth has a long history, but the development of social media provides a unique opportunity for researchers to focus on and examine this world directly. Using data such as ours to test the implications of social network theory could prove very fruitful.

Reference:

Agarwal, Ashish, Alok Kumar, Alvin Leung, and Prabhudev Konana, 2014, Online Search and Return Comovement, Working paper, University of Miami.

Angel, James J., Lawrence E. Harris, and Chester S. Spatt, 2011, Equity trading in the 21st century, *The Quarterly Journal of Finance* 1, 1-53.

Barber, Brad M., and Odean Terrance, 2008, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21, 785-818.

Battalio, Robert H., Shane A. Corwin, and Robert H. Jennings, 2015, Can brokers have it all? on the relation between make take fees & limit order execution quality, Working paper, University of Notre Dame.

Bhagwat, Vineet, and Timothy R. Burch, 2014, Pump It Up? Tweeting to Manage Investor Attention to Earnings, Working Paper, University of Oregon and University of Miami.

Bildik, Recep, Han N. Ozsoylev, Johan Walden, and M. Deniz Yavuz, 2013, Investor Networks in the Stock Market, *Review of Financial Studies,* forthcoming.

Blankespoor, Elizabeth, Brian P. Miller and Hal D. White, 2014, Initial evidence on the market impact of the XBRL mandate, *Review of Accounting Studies* 19, 1468-1503.

Bollen, Johan, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of Computational Science* 2, 1-8.

Breiman, Leo, 2001, Random forests, *Machine learning* 45, 5-32.

Brown, Jeffrey R., Zoran Ivkovich, Paul A. Smith, and Scott Weisbenner, 2008, Neighbors matter: Causal community effects and stock market participation, *Journal of Finance* 63, 1509-1531.

Chan, Wesley S., 2003, Stock price reaction to news and no-news: Drift and reversal after headlines, *Journal of Financial Economics* 70, 223-260.

Chen, Hailiang, Byoung-Hyoun Hwang, and Baixiao Liu, 2013, The Economic Consequences of Having “Social” Executives, Working paper, City University of Hong Kong, Cornell University and Florida State University.

Cheng, Justin, Lada A. Adamic, P. Alex Dow, Jon Kleinberg and Jure Leskovec, 2014, Can cascades be predicted?, *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee.

Cohen, Lauren, and Andrea Frazzini, 2008, Economic links and predictable returns, *Journal of Finance* 63, 1977-2011.

Colla, Paolo, and Antonio Mele, 2010, Information Linkages and Correlated Trading, *Review of Financial Studies* 23, 203-246.

Corwin, Shane A., and Jay F. Coughenour, 2008, Limited attention and the allocation of effort in securities trading, *Journal of Finance* 63, 3031-3067.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461-1499.

DellaVigna, Stefano, and Joshua M. Pollet, 2009. Investor Inattention and Friday Earnings Announcements, *Journal of Finance 64, 709-749.*

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel. 2003. Persuasion Bias, Social Influence, and Unidimensional Opinions, *Quarterly Journal of Economics* 118, 909-968.

Duffe, Darrel, and Gustavo Manso, 2007. Information percolation in large markets, *American Economic Review* 97, 203-209.

Duffe, Darrel, Semyon Malamud, and Gustavo Manso, 2009. Information percolation with equilibrium search dynamics, *Econometrica* 77, 1513-1574.

Engelberg, Joseph E., and Christopher A. Parsons, 2011, The causal impact of media in financial markets, *Journal of Finance* 66, 67-97.

Fang, Lily and Joel Peress, 2009, Media Coverage and the Cross-section of Stock Returns, *Journal of Finance* 64, 2023–2052.

Gai, Jiading, Dong Ju Choi, David O'Neal, Mao Ye, and Robert S., Sinkovits, 2014, Fast Construction of Nanosecond Level Snapshots of Financial Markets, *Concurrency and Computation: Practice & Experience*.

Giannini, Robert Charles, Paul J. Irvine, and Tao Shu, 2013, The convergence and divergence of investors' opinions around earnings news: Evidence from a social network, *Asian Finance Association (AsFA) 2013 Conference.*

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders, *Journal of Financial Economics* 14, 71-100.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393-408.

Gurun, Umit G., and Alexander W. Butler, 2012, Don't believe the hype: Local media slant, local advertising, and firm value, *Journal of Finance* 67, 561-598.

Han, Bing, and David Hirshleifer, 2013, Self-Enhancing Transmission Bias and Active Investing, Working Paper, University of Toronto and University of California-Irvine.

Han, Bing, and Liyan Yang, 2013, Social Networks, Information Acquisition, and Asset Prices, *Management Science* 59, 1444-1457.

Hirshleifer, David and Siew Hong Teoh, 2003, Limited attention, information disclosure, and financial reporting, *Journal of Accounting and Economics* 36, 337-386

Hirshleifer, David, and Siew Hong Teoh, 2009, The Psychological Attraction Approach to Accounting and Disclosure Policy, *Contemporary Accounting Research* 26, 1067-1090.

Hong, Dong, Harrison Hong, and Andrei Ungureanu, 2012, An Epidemiological Approach to Opinion and Price-Volume Dynamics, Working paper, Princeton University and Singapore Management University.

Hong, Harrison, and Jeremy Stein, 2007, Disagreement and the Stock Market, *Journal of Economic Perspectives* 21, 109-128.

Hong, Harrison, and Jeremy Stein, 1999, A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets, *Journal of Finance* 54, 2143-2184.

Hong, Harrison, and Jeremy Stein, 2006, Disagreement and the Stock Market, *Journal of Economic Perspective* 21, 109-128.

Hou, Kwei., Lin Peng, and Wei Xiong, 2006, A tale of two anomalies: the implication of investor attention for price and earnings momentum, Working paper, Princeton University.

Huberman, Gur and Tomer Regev, 2001, Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar, *Journal of Finance* 56, 387–396.

Ivkovich, Zoran., and Scott Weisbenner, 2007. Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices, *Review of Financial Studies* 20, 1327-1357.

Jenders, Maximilian, Gjergji Kasneci, and Felix Naumann, 2013, Analyzing and predicting viral tweets, *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee.

Kahneman, Daniel, Amos Tversky, 1973, On the psychology of prediction, *Psychological Review* 80, 237–251.

Kwan, Amy, Ronald W. Masulis, and Thomas H. McInish, 2014, Trading rules, competition for order flow and market fragmentation, *Journal of Financial Economics*, forthcoming.

Kyle, Albert S., 1985, Continuous Auctions and Insider Trading, *Econometrica* 53, 1315-1336.

Loughran, Tim and Bill McDonald, 2014, Measuring Readability in Financial Disclosures, *Journal of Finance* 69, 1643–1671.

Loughran, Tim, and Bill McDonald, 2011, When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance* 66, 35-65.

Manela, Asaf, 2014, The Value of Diffusing Information, *Journal of Financial Economics*, forthcoming.

Merton, Robert C., 1987, A Simple Model of Capital Market Equilibrium with Incomplete Information, *Journal of Finance* 42, 483–510.

Ozsoylev, Han N., and Johan Walden, 2011, Asset Pricing in Large Information Networks, *Journal of Economic Theory* 146, 2252-2280.

Peng, Lin, and Wei, Xiong, 2006, Investor attention, overconfidence and category learning, *Journal of Financial Economics* 80, 563-602.

Peress, Joel, 2008, Media coverage and investors' attention to earnings announcements, Working paper, INSEAD.

Peress, Joel, 2014, The media and the diffusion of information in financial markets: evidence from newspaper strikes, *Journal of Finance* 69, 2007-2043.

Peress, Joel, and Daniel Schmidt, 2014, Glued to the TV: The Trading Activity of Distracted Investor, Working Paper, INSEAD, HEC Paris.

Petrovic, Sasa, Miles Osborne, and Victor Lavrenko, 2011, RT to Win! Predicting Message Propagation in Twitter, ICWSM.

Shiller, Robert J., 1984, Stock prices and social dynamics, *Brookings Papers on Economic Activity*, 457-510.

Shive, Sophie, 2010, An epidemic model of investor behavior, *Journal of Financial and Quantitative Analysis* 45, 169-198.

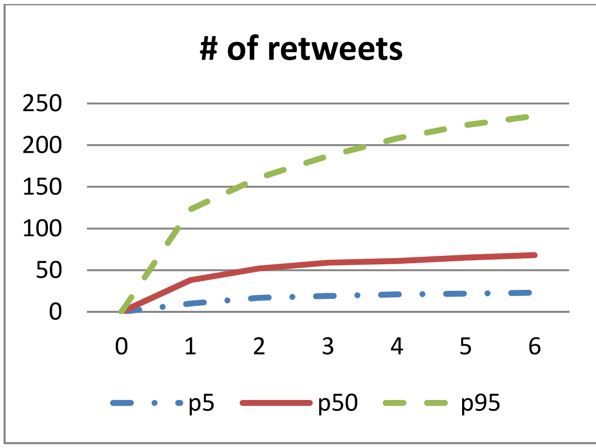
Sims, Christopher A., 2003, Rational inattention and monetary economics, *Journal of Monetary Economics* 50, 665–690.

Suh, Bongwon, Lichan Hong, Peter Pirolli, and Ed H. Chi, 2010, Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network, *Social Computing (socialcom)*, 2010 IEEE second international conference on Privacy, Security, Risk and Trust.

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *Journal of Finance* 62, 1139-1168.

Figure 1: Retweets during the first hour

The left figure plots the total number of retweets during the first hour after the original tweet, for the median case, for the 5th percentile, and for the 95th percentile. In the figure on the right, we also account for the number of followers of each Twitter account that posts the original tweet or the retweet. As a result, the number measures the number of potential users the tweet can reach in the first hour. Each time interval represents 10 minutes.



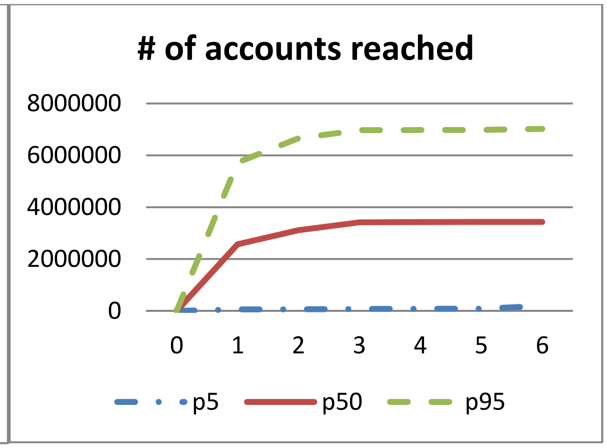


Figure 2: Fast and Slow Diffusion: Retweet Data

Panel (a) and (b) plot the cumulative numbers of retweets and trading volumes for each of the six 10-minute intervals during the first hour following a tweet. Both variables are normalized by their totals during the first hour, so the plot resembles a cumulative distribution function (CDF). Rapid diffusion occurs when more than 60% of total first-hour retweets occur in the first 10 minutes; slow diffusion occurs when less than 40% of total first-hour retweets occur in the first 10 minutes.

(a) diffusion (b) trading volume

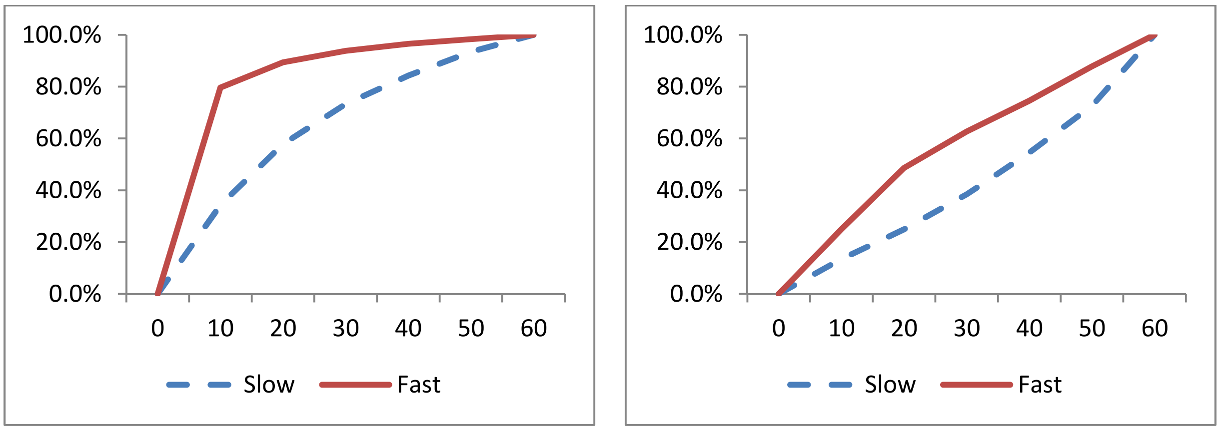


Table 1: Summary Statistics on Twitter Accounts in the Sample

This table reports summary statistics on the 277 Twitter accounts we monitored during the four months from 2013/11 through 2014/10. They include 78 major media outlets and famous analysts (e.g., @WSJ and @CNBC), 56 active accounts of S&P 1500 CEOs and CFOs (e.g., Elon Musk, CEO & Chief Product Architect of Tesla Motors), and 143 official Twitter accounts of S&P 500 companies (e.g. @TysonFoods).

	N	Mean	Std dev	Q1	Median	Q3
			Total number of followers			
CEO/CFO	56	54,576	173,192	167	621	7034
Media	78	888,545	1,789,724	17,802	100,446	923,497
SP500	143	601,931	1,222,074	42,249	125,521	467,134
			Number of years since inception			
CEO/CFO	56	4.3	1.8	3.1	4.5	5.4
Media	78	5.7	1.7	5.1	5.6	6.6
SP500	143	5.3	1.7	4.9	5.4	6.0
			Number of Tweets per year			
CEO/CFO	56	264	703	6	45	186
Media	78	7,488	6,312	2,157	6,076	11,821
SP500	143	3,334	6,987	788	1,413	2,985

Table 2: Examples of Tweets in Our Sample

This table contains examples of tweets in our sample. We report the date, source, and content associated with each tweet. The relevant tickers are also identified.

Date	Source	Tweet	Ticker
11/12/2013	@WSJ	AirWatch expresses interest in buying service division of Blackberry: http://t.co/R9vTFvfHkD	BBRY
11/14/2013	@FordTrucks	@Ford?F-150 EcoBoost hits 400,000 sales, saving 45 million gallons of gas annually:?http://t.co/xYRgWVGoph?#BuiltFordTough	F
11/22/2013	@paradimeshift	Western Union and tradition bank wire transfers are dead! 11/23/13 $147 Million transferred for 37 CENTS! #bitcoin	WU
12/9/2013	@ABC	Just in: American Airlines/US Airways merger complete says company - @ABCaviation	AAL
12/19/2013	@DavidJBarger	Very cool @JetBlue's SJU Team welcomed N903JB, our first A321, "Bigger, Brighter, Bluer" to the airline! http://t.co/IU7JFJt9Y4	JBLU
1/9/2014	@EMCcorp	Congratulations to David Goulden - new CEO of #EMC. Joe Tucci will continue as Chairman & CEO of EMC Corporation http://t.co/no4P9BYOwT	EMC
1/29/2014	@BreakingNews	Facebook earnings: Q4 EPS $0.31 ex-items v. $0.27 estimate; revenues $2.59 billion v. $2.33 billion estimate - @CNBC http://t.co/sNqDbtfyzv	FB
2/5/2014	@ReutersBiz	Twitter reports revenue of $243 million, up 116 percent year-over-year	TWTR
2/19/2014	@businessinsider	TESLA EXPECTS 55% VEHICLE DELIVERY GROWTH IN 2014 http://t.co/aXQZAqHd0z	TSLA
3/4/2014	@CNET	2015 Lamborghini Huracan debuts with Nvidia-powered digital dashboard http://t.co/j7bvnt9JuH http://t.co/XlfBKsU85Q	NVDA

Table 3: Summary Statistics of Firms in our Sample

This table reports summary statistics on the stocks in our final sample. Market capitalization is measured in millions of dollars. Turnover and daily return volatility are measured over the past 30 days. Institutional ownership (IO) is measured using the 13f filing at the most recent quarter. The last row reports the average percentiles of the entire CRSP universe. Our sample covers 178 distinct stocks during the one-year period from 2013/11 through 2014/10

	Mkt Cap (M$)	Turnover	Volatility	IO
Mean	136,668	4.20	0.022	0.577
Median	85,186	2.05	0.016	0.602
Std dev	144,755	4.80	0.019	0.175
CRSP percentile	89.9	62.5	50.5	80.0

Table 4: Diffusion speed and trading intensity

This regression links retweets to trading during the first hour after a tweet. The dependent variable is the percentage of first-hour trading that occurs in the first 10 minutes. Panel A examines total trading volume in TAQ. Panel B focuses on trading volume from TRFs (exchange symbol D from the TAQ dataset). The main independent variable, diffusion, measures the percentage of first-hour retweets that occur in the first 10 minutes. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); abs past 1h ret (absolute stock returns over the market in the past hour); abs 10m ret (absolute stock returns over the market in the first 10 minutes after the tweet). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. The sample covers 1,261 tweets during one year from 2013/11 through 2014/10.

Panel A: Total trading volume

	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value
Intercept	0.02	0.65	0.11	3.27	0.01	0.12
diffusion	0.30	5.13	0.17	3.07	0.14	2.25
pre-market			-0.12	-9.88	-0.12	-7.70
afternoon			-0.02	-1.58	-0.01	-1.02
post-market			0.07	3.67	0.07	2.86
size					0.00	1.00
turn					0.00	-0.37
volatility					1.43	10.09
bm					0.00	-0.53
abs past 1h ret					-0.46	-2.41
abs 10m ret					2.71	3.10
stock FE	Yes		Yes		Yes	
account FE	Yes		Yes		Yes	
R-square	0.031		0.123		0.134	

Panel B: TRF trading volume

	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value
Intercept	0.12	2.64	0.24	5.95	0.33	2.37
diffusion	0.44	5.84	0.25	3.69	0.21	2.91
pre-market			-0.24	-11.81	-0.22	-8.23
afternoon			-0.02	-1.51	-0.02	-1.17
post-market			0.14	4.89	0.14	3.81
size					0.00	-0.61
turn					0.00	1.29
volatility					-1.87	-1.19
bm					0.00	1.49
abs past 1h ret					-0.81	-3.35
abs 10m ret					2.21	1.97
stock FE	Yes		Yes		Yes	
account FE	Yes		Yes		Yes	
R-square	0.036		0.215		0.248	

Table 5: Analysis of TD Ameritrade Brokerage Account Data

We merge our tweet sample with the TD Ameritrade (TDA) brokerage-account-level transaction data during the overlapping period from 2013/11 through 2014/06. We focus on tweets during market hours and retail accounts that trade the corresponding stock at least once during the first three hours after a tweet. The merged sample contains 331 distinct tweets and 35,443 distinct TDA accounts. Panel A reports descriptive statistics across both accounts and tweets. To compute net trades, one buy (sell) is counted as 1 (-1). Trading intensity is again measured as the percentage of first-hour trading that occurs in the first 10 minutes. Panel B repeats the regressions of Table 4 for our merged sample of tweets. We measure trading intensity using (I) all trades from TAQ; (II) all TDA trades; (III) all TDA trades of female investors; (IV) all TDA trades of young investors (age <35); and (V) all TDA trades of "rich" investors (with stock holdings >$100K). Other control variables include afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); abs past 1h ret (absolute stock returns over the market in the past hour); abs 10m ret (absolute stock returns over the market in the first 10 minutes after the tweet). We include stock and Twitter account fixed effects. The standard error is clustered by ticker.

Panel A: Descriptive statistics of the merged sample

Across 35,443 accounts

Avg char	Mean	Std dev	Q1	Median	Q3
Age	48.7	14.3	38.0	48.0	58.0
Stock holding ($)	78,063	167,449	3,481	20,146	74,288
Net trade	0.153	1.023	-1.000	0.167	1.000
Trade intensity	20.0%	36.6%	0.0%	0.0%	25.0%

Across 331 tweets

Avg char	Mean	Std dev	Q1	Median	Q3
# of accounts	194.4	288.9	19.0	69.0	242.0
% of Male	72.1%	12.2%	67.9%	71.7%	76.2%
Age	51.0	5.0	48.8	50.4	52.9
Stock holding ($)	91,004	51,164	68,843	84,461	102,741
Net trade	0.060	0.430	-0.202	0.045	0.333
Trade intensity	20.4%	17.1%	10.2%	18.5%	26.1%

Panel B: The link between diffusion speed and trading intensity

	All Trades (I)		All TDA Trades (II)		TDA, Female (III)		TDA, Young (IV)		TDA, Rich (V)	
	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value
Intercept	-0.07	-0.54	-0.06	-0.26	0.01	0.02	-0.08	-0.19	0.25	0.89
diffusion	0.16	2.89	0.23	2.06	0.27	2.02	0.18	1.04	0.36	2.71
afternoon	-0.01	-0.44	0.01	0.62	0.04	1.39	0.00	0.01	0.06	2.03
size	0.01	1.38	0.01	0.50	0.00	-0.08	0.00	0.21	-0.01	-0.78
turn	0.00	0.79	0.00	0.87	0.00	-0.31	0.00	-0.80	0.00	1.49
volatility	0.01	0.00	-1.76	-0.50	1.78	0.32	8.65	1.02	-8.46	-1.69
bm	0.00	1.87	0.00	2.45	0.00	0.35	0.00	-0.43	0.00	1.75
abs past 1h ret	-0.17	-0.15	-0.10	-0.07	0.50	0.27	-1.45	-0.91	2.14	0.90
abs 10m ret	3.41	1.41	7.76	1.71	6.79	1.73	6.85	1.39	2.88	0.89
stock FE	Yes		Yes		Yes		Yes		Yes	
account FE	Yes		Yes		Yes		Yes		Yes	
R-square	0.068		0.077		0.055		0.037		0.090	

Table 6: Retail attention, stock returns, and change in dollar spread

The dependent variables are stock returns (in excess of the market and by percentage) 10 minutes after a tweet until the end of the day (CAR%[10m, close d0]); stock returns (in excess of the market and by percentage) on the next trading day (CAR%[close d0, close d1]); and the change in the average dollar spread from the one hour before the tweet to the one hour after. The main independent variable is diffusion_3hr, which measures the log number of users the tweet can potentially reach three hours after the tweet. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); past 1h ret (stock returns over the market in the past hour); abs past 1h ret (absolute value of past 1h ret); 10m ret (stock returns over the market in the first 10 minutes after the tweet); abs 10m ret (absolute value of 10m ret); Isbreaking (a dummy variable equal to 1 if the tweet contains "breaking"). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. Panel A covers all 1,261 tweets during one year from 2013/11 through 2014/10. Panel B covers only tweets on firms with market capitalization below the median of all stocks in our sample.

Panel A: All stocks

	CAR%[10m, close d0]		CAR%[close d0, close d1]		Spread_chg	
	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value
Intercept	-4.17	-2.98	5.27	3.1	0.13	0.22
diffusion_3hr	0.21	3.19	-0.21	-2.74	-0.05	-1.89
pre-market	0.13	0.57	-0.23	-0.99	-0.78	-9.47
afternoon	-0.03	-0.15	-0.24	-0.93	0.34	7.43
post-market	0.02	0.12	-0.12	-0.5	0.23	4.65
size	0.07	1.11	-0.03	-0.45	0.02	0.85
turn	0.00	0.16	0.00	1.85	0.00	-1.82
volatility	-18.32	-0.31	-137.66	-2.3	13.23	2.19
bm	0.01	0.28	0.06	2.1	-0.01	-1.63
past 1h ret	-4.12	-0.46	-6.56	-1.09	-1.63	-1.26
abs past 1h ret	15.97	1.53	1.63	0.2	-1.52	-0.89
10m ret	-6.11	-0.29	-38.37	-1.68	-4.80	-1.89
abs 10m ret	64.43	3.08	20.25	0.64	1.96	0.57
Isbreaking	-0.20	-0.96	-0.88	-2.98	-0.07	-0.71
stock FE	Yes		Yes		Yes	
account FE	Yes		Yes		Yes	
R-square	0.117		0.081		0.341	

Panel B: Smaller stocks

	CAR%[10m, close d0]		CAR%[close d0, close d1]		Spread_chg	
	Coeff	*t*-value	Coeff	*t*-value	Coeff	*t*-value
Intercept	-7.61	-2.55	7.81	2.03	0.81	1.00
diffusion_3hr	0.39	3.31	-0.38	-2.98	-0.06	-2.12
pre-market	0.33	0.79	-0.64	-1.52	-0.89	-8.88
afternoon	0.01	0.03	-0.30	-0.61	0.37	7.00
post-market	0.13	0.43	-0.51	-1.11	0.21	4.01
size	0.10	0.68	-0.01	-0.03	0.00	-0.02
turn	0.00	0.08	0.00	1.72	0.00	-1.03
volatility	-13.01	-0.2	-150.15	-2.18	8.71	1.20
bm	0.01	0.42	0.07	2.28	-0.01	-1.58
past 1h ret	-6.33	-0.67	-7.78	-1.25	-1.80	-1.41
abs past 1h ret	17.45	1.57	5.06	0.59	-0.92	-0.53
10 ret	-5.95	-0.25	-44.23	-1.79	-4.87	-1.86
abs 10m ret	67.24	2.97	23.34	0.68	2.18	0.61
Isbreaking	-0.23	-0.54	-1.44	-2.60	-0.10	-0.87
stock FE	Yes		Yes		Yes	
account FE	Yes		Yes		Yes	
R-square	0.148		0.113		0.368	

Table 7: Predicting diffusion

In this OLS regression, we use Twitter characteristics observable 10 minutes after a tweet to predict the growth rate in diffusion from 10 minutes to 3 hours after the tweet. The growth rate is defined as log(diffusion_3hr) – log(diffusion_10m), where diffusion_10m and diffusion_3hr are the number of Twitter users the tweet potentially reaches after 10 minutes and three hours, respectively. The Twitter characteristics include Total number of tweets (the average daily number of tweets sent by that Twitter account); log(# of followers of recent retweeters)—the total number of followers in log, of the most recent 5 Twitter accounts that retweeted the tweet; Speed of recent retweets (inverse of the average time lapse between the most recent 5 retweets); Hour (calendar hour of the tweet); IsWest (a dummy variable equal to 1 if the tweet is sent from the West Coast); IsCEO (a dummy variable equal to 1 if the tweet is sent by the CEO of the company); HasPicture (a dummy variable equal to 1 if the tweet contains a picture); HasURLs (a dummy variable equal to 1 if the tweet contains URL links); HasHashtags (a dummy variable equal to 1 if the tweet contains Hashtags). The sample covers all 1,261 tweets during one year from 2013/11 through 2014/10.

Variable	Coeff	*t*-value
Intercept	-4.15	-10.87
Total # of tweets	-0.01	-3.71
log(# of followers of recent retweeters)	0.02	2.17
Speed of recent retweets	4.31	5.16
Hour	-0.05	-4.95
IsWest	0.75	5.80
IsCEO	1.24	2.35
HasPicture	0.63	6.28
HasURLs	-0.37	-2.52
HasHashtags	0.41	2.46
R-Square	0.190	

Table 8: Predicted retail attention and stock returns

The dependent variables are stock returns (in excess of the market and by percentage) 10 minutes after a tweet until the end of the day (CAR%[10m, close d0]) and stock returns (in excess of the market and by percentage) on the next trading day (CAR%[close d0, close d1. In Panel A, the main independent variable is predicted diff, which measures the log number of users the tweet is predicted to reach three hours after the tweet. It is computed by summing the log number of Twitter users the tweet reaches after 10 minutes (diffusion_10m) and the predicted log growth rate from the regression shown in Table 6. In panel B, the main independent variable is simply diffusion_10m. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); past 1h ret (stock returns over the market in the past hour); abs past 1h ret (absolute value of past 1h ret); 10m ret (stock returns over the market in the first 10 minutes after the tweet); abs 10m ret (absolute value of 10m ret); Isbreaking (a dummy variable equal to 1 if the tweet contains "breaking"). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. The sample covers all 1,261 tweets during one year from 2013/11 through 2014/10.

Panel A: Using predicted diffusion

	CAR%[10m, close d0]		CAR%[close d0, close d1]	
	Coeff	*t*-value	Coeff	*t*-value
Intercept	-3.97	-2.88	4.93	2.89
predicted diff	0.20	3.16	-0.17	-2.25
pre-market	0.21	0.96	-0.22	-0.94
afternoon	-0.02	-0.08	-0.26	-1.00
post-market	0.02	0.14	-0.06	-0.24
size	0.06	1.05	-0.04	-0.64
turn	0.00	0.21	0.00	1.90
volatility	-16.90	-0.28	-141.06	-2.34
bm	0.00	0.06	0.06	2.08
past 1h ret	-3.95	-0.19	-38.31	-1.66
abs past 1h ret	-2.52	-0.28	-6.10	-1.00
10 ret	16.63	1.6	1.49	0.18
abs 10m ret	61.75	2.96	19.15	0.60
Isbreaking	-0.24	-1.12	-0.92	-3.11
stock FE	Yes		Yes	
account FE	Yes		Yes	
R-square	0.126		0.081	

Panel B: Using diffusion after 10 minutes

	CAR%[10m, close d0]		CAR%[close d0, close d1]	
	Coeff	*t*-value	Coeff	*t*-value
Intercept	-3.24	-2.25	3.37	1.78
diffusion_10m	0.10	1.58	-0.04	-0.53
pre-market	0.04	0.16	-0.16	-0.68
afternoon	-0.11	-0.5	-0.14	-0.54
post-market	0.08	0.47	-0.03	-0.12
size	0.09	1.4	-0.06	-0.85
turn	0.00	0.32	0.00	1.96
volatility	-17.16	-0.31	-134.22	-2.28
bm	0.01	0.41	0.04	1.43
past 1h ret	-9.67	-0.38	-40.76	-1.85
abs past 1h ret	-10.09	-1.09	-8.30	-1.36
10 ret	9.12	0.79	3.56	0.49
abs 10m ret	72.95	2.84	20.33	0.71
Isbreaking	-0.24	-0.87	-1.00	-3.11
stock FE	Yes		Yes	
account FE	Yes		Yes	
R-square	0.081		0.067	

Table 9: Predicted retail attention and stock returns: Out-of-sample

We break our one-year sample into an in-sample period (2013/11 through 2014/04) and an out-of-sample period (2014/05 through 2014/10). We estimate the predictive regression from Table 6 during the in-sample period only. We then take the estimated coefficients and apply them to the out-of-sample period to compute predicted diff. In other words, predicted diff is observable 10 minutes after a tweet. We then link predicted diff to future returns in the out-of-sample period. The dependent variables are stock returns (in excess of the market and by percentage) 10 minutes after the tweet until the end of the day (CAR%[10m, close d0]) and stock returns (in excess of the market and by percentage) on the next trading day (CAR%[close d0, close d1. Other control variables include pre-market (a dummy variable equal to 1 if the tweet takes place before 9:30 a.m. EST); afternoon (a dummy variable equal to 1 if the tweet takes place between 12:30 p.m. and 4:00 p.m. EST); post-market (a dummy variable equal to 1 if the tweet takes place after 4:00 p.m. EST); size (log market capitalization); turn (turnover); volatility (daily returns volatility in the past 30 days); bm (book-to-market ratio); past 1h ret (stock returns over the market in the past hour); abs past 1h ret (absolute value of past 1h ret); 10m ret (stock returns over the market in the first 10 minutes after the tweet); abs 10m ret (absolute value of 10m ret); Isbreaking (a dummy variable equal to 1 if the tweet contains “breaking”). We include stock and Twitter account fixed effects. The standard error is clustered by ticker. The regression uses tweets during the out-of-sample period from 2014/05 through 2014/10.

	CAR%[10m, close d0]		CAR%[close d0, close d1]	
	Coeff	*t*-value	Coeff	*t*-value
Intercept	-1.17	-0.69	4.14	1.85
Predicted diff	0.15	1.88	-0.22	-2.18
pre-market	0.56	1.91	-0.44	-1.37
afternoon	0.13	0.62	-0.70	-2.19
post-market	0.15	0.83	-0.40	-1.24
size	-0.05	-0.65	0.07	0.77
turn	0.00	0.39	0.00	2.07
volatility	-40.60	-0.53	-159.74	-2.24
bm	0.01	0.28	0.05	0.90
past 1h ret	28.41	0.65	-122.08	-2.90
abs past 1h ret	-3.96	-0.4	-1.47	-0.22
10 ret	14.61	1.28	1.54	0.17
abs 10m ret	77.40	1.52	99.94	1.88
Isbreaking	-0.16	-0.66	-1.18	-2.90
stock FE	Yes		Yes	
account FE	Yes		Yes	
R-square	0.133		0.187	