Working Paper: April 2024

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Network through Social Media Connections

Huaixin Wang
University of Macau

Weichen Zhang
Singapore Management University



SIM KEE BOON INSTITUTE FOR FINANCIAL ECONOMICS

LEE KONG CHIAN SCHOOL OF BUSINESS

Network through Social Media Connections

by *

Huaixin Wang Weichen Zhang

April 29, 2024

Abstract

Using text data from the Reddit platform, we construct inter-firm linkages based on shared threads and shared authors within social media. We find that firms linked via social media exhibit correlated fundamentals across various characteristics. The returns of Reddit peer stocks positively predict focal stocks' future returns, suggesting sluggish dissemination of latent information within the social media network. This lead-lag effect is also robust to controlling for other firm characteristics and alternative inter-firm connections. Our findings suggest that social media activities contain collective perceptions of connectedness between firms, thereby providing an implicit representation of the financial network.

JEL classification: G11, G12, G14

Keywords: Social media, connected firms, comovements, return predictability

*We thank Michael Clemens (discussant), Bing Han, Grace Xing Hu, Weijie Hu (discussant), Mengyao Kang (discussant), Bo Li, Shen Lin, Ke Wu, Jingda Yan, Jianfeng Yu, Fudong Zhang, Linti Zhang, Xiaoyan Zhang, Dexin Zhou, Zhen Zhou, and conference and seminar participants at CFAM 2023, EFMA 2023, AsianFA 2023, FMCG 2023, Tsinghua PBCSF workshop, and Renmin University of China for helpful comments and suggestions. All errors are our own.

Author affiliation/contact information: Huaixin Wang and Weichen Zhang are at PBC School of Finance, Tsinghua University. E-mail address: wanghx.19@pbcsf.tsinghua.edu.cn (Huaixin Wang), zhangwch.18@pbcsf.tsinghua.edu.cn (Weichen Zhang).

Electronic copy available at: https://ssrn.com/abstract=4536567

1 Introduction

Firms are interconnected in many different ways and form various networks. This could include supply chains, business partners, and geographic proximity. While many of these connections are easily identifiable through objective measures, others emerge from subjective cognition. For example, Tesla, renowned for its electric vehicles, and Google, a leader in internet services, do not share a direct financial connection; however, they are often considered interconnected through less visible aspects, such as strategic collaborations, technological innovations, and shared future visions. These elements are challenging to quantify using traditional metrics such as production networks or industry classifications. In this paper, we show that user activities on social media are synonymous with investors' information structure, which can help uncover these implicit firm networks.

Our approach is motivated by the concept of the "wisdom of crowds," which posits that a group of individuals can surpass individual biases to generate novel insights and effective behavior. Given that discussions on social media reflect a collection of individual insights, we use the network feature of social media to illuminate economic linkages underlying firms. In particular, we aim to examine the implications of social media for financial networks, which prompts two crucial questions guiding our study: First, can social media provide valuable information about inter-firm connections? Second, does the market price effectively reflect such information?

Specifically, we compile decentralized insights of social media users with *Reddit* as our research subject. As the renowned "front page of the Internet," Reddit stands as a hub for news dissemination and vibrant discussions. As of January 2024, this platform ranks as the 9th most-visited website globally and the 3rd most-visited within the United States.[1] Figure 1 offers a brief overview of Reddit's structure. Typically, Reddit consists of millions of "subreddits," each devoted to distinct topics.[2] Within subreddits, users start discussions by creating submissions, or "posts." Each submission initiates a "thread," which includes all subsequent comments and replies, forming a structured dialogue. Users participate in threads by commenting directly on the original post or on others' comments, fostering intricate "reply-to" connections and a multi-layered dialogue. Our focus is the *r/wallstreetbets* subreddit, where users discuss equity and options trading. r/wallstreetbets is also well-known for its emphasis on highly

[1] For the latest statistics and details, please refer to `https://www.semrush.com/website/reddit.com/overview/`

[2] For example, a topic can be economic news, technology, politics, sports, or personal hobbies.

Electronic copy available at: https://ssrn.com/abstract=4536567

speculative trading strategies (Bradley et al., 2023) and the "GameStop Short Squeeze" that occurred in January 2021.[3] Within this subreddit, users wield the flexibility to post a wide array of content, ranging from news about specific firms to other materials they believe would captivate the community's interest, including textual posts, images, videos, and hyperlinks to external web resources.

Using comprehensive text data from the Reddit platform, we start our analysis by defining two types of "Reddit peers" for individual firms. First, two firms are identified as Reddit thread-linked each month if there was at least one thread on Reddit mentioning both firms in the past six months. This construction captures the collective perception of interconnection between these entities. Second, we defined two firms as Reddit author-linked if at least one Reddit user published comments on both firms in the past six months. This definition reflects the aggregated preference of Reddit users for potentially connected firms. By construction, we essentially formulate two types of Reddit networks, with firms representing nodes and shared threads or users constituting edges. Figure 2 presents a graphical depiction of the Reddit linkages.

The first part of our findings shows that Reddit-linked firms are economically connected. Specifically, we document positive and significant fundamental comovements between focal firms and their Reddit peers along a battery of characteristics. These encompass profitability, valuation, growth potential, and investment. Additionally, we find that the fundamentals of Reddit-linked peer firms also serve as good predictors for focal firms' future fundamentals along several dimensions, including, for example, return on assets, return on equity, the earnings-to-price ratio, and sales growth. These results suggest that firm linkages identified through Reddit activities indeed help to capture real economic connections, highlighting the significant role of social media networks in revealing implicit information between firms.

While we find that discussions across different firms on Reddit are unlikely to be random, these connections might simply reflect stale information that has already been incorporated into stock prices. Therefore, we proceed to examine our second question regarding market efficiency in the social media network. To this end, we test the cross-firm return predictability in the context of Reddit connections. If investors fail to fully integrate network information into their trading, then the focal stock's price may lag in reflecting fundamental news conveyed by Reddit peers' returns. Consistent with this hypothesis, we find that the stock returns of Reddit peers positively and significantly

[3]The short squeeze caused GameStop's share price to rise by 1,500% over two weeks, reaching an all-time intraday high of $483.00 on the NYSE on January 29.

Electronic copy available at: https://ssrn.com/abstract=4536567

predict the future returns of the focal stock. The predictive ability of Reddit peer returns is less pronounced for firms with higher analyst coverage, more media coverage, or higher institutional ownership. We also find that the predictive ability is stronger for stocks with more Reddit peers. This result supports the hypothesis that investors underreact to peer stocks' price movements, due to their limited capability of processing value-relevant information in social media networks. We further examine the predictive ability of Reddit peer returns for focal stocks' returns on days with earnings announcements or news releases. The underreaction story suggests that investors' expectations about the focal stock tend to be too pessimistic (optimistic) relative to the rational benchmark in the presence of favorable (unfavorable) news, thus giving rise to the return predictability. Consequently, the anomaly returns should be more pronounced when firm-specific information is released and investors correct their beliefs (Engelberg et al., 2018). We find evidence consistent with this hypothesis.

An important concern regarding our findings is that the Reddit network could be a repackage of other existing inter-firm linkages. For example, it is well-known that industry portfolios also display a lead-lag effect (Moskowitz and Grinblatt, 1999; Hou, 2007). More recently, Ali and Hirshleifer (2020) demonstrate that shared analyst coverage captures most network information of economically connected firms and provides a potentially unified measure of cross-firm return predictability. We empirically show that our Reddit network is distinct from these previously studied connections, as we find that Reddit-implied links only exhibit little overlap with them. More importantly, we control for the industry lead-lag effect and the shared analyst coverage return in regressions. We find that the predictive ability of Reddit peer returns remains highly significant. In robustness tests, we further consider a wide range of alternative economic linkages as additional controls and still find consistent results. These findings indicate that Reddit linkages provide incremental information beyond existing inter-frim linkages.

Overall, we find that the collective view on social media helps uncover economic linkages, and market prices do not reflect such information efficiently. In a context where individual attention and knowledge are inherently limited, our results show that the "wisdom of crowds," emerging from communications and interactions, sheds light on the broad picture of cross-firm associations.

This paper contributes to the emerging literature that studies the impact of social media on financial markets. Pedersen (2022) provides a theoretical framework of how investment ideas transmitted through a social network can affect investor behavior and market prices. The belief spillover in social network interactions implied by Hosseini et al.

Electronic copy available at: https://ssrn.com/abstract=4536567

(2020) is closely related to our study. Using data from *Seeking Alpha*, Chen et al. (2014) is one of the earliest papers to discuss how investor opinions are propagated through social media. Hu et al. (2023) show that Reddit activities have predictive power for future stock returns. Dim (2023) discusses how social media investment analysts, the influencers on social media platforms, affect the form of others' beliefs. Hosseini et al. (2020) use Twitter data to investigate the social media risk premium in stocks and bonds. These theoretical and empirical findings reveal the informational role of social media platforms and online activities, thus motivating us to construct inter-firm social media linkages.

Our work is also closely related to the literature on social networks. Kuchler and Stroebel (2021) provide a comprehensive review of recent contributions to this field. Studies in this area typically discuss the impact of social networks on financial decisions; for example, Bailey et al. (2018a) show how Facebook connections influence house price expectations. Recent studies also use social media networks to infer connections in real society. Bailey et al. (2018b) use Facebook administrative data and create the Social Connectedness Index to measure social connectedness between two locations. Furthermore, Al Guindy and Riordan (2019) use Twitter data to map the network of U.S. firms, shedding light on the dissemination of information and noise within these networks.

This paper also adds to the growing literature on discovering economic connections and studying their asset pricing implications. These connections manifest in multiple ways, including, for example, the supply chain (Cohen and Frazzini, 2008; Menzly and Ozbas, 2010), conglomerate firms (Cohen and Lou, 2012), co-searches (Lee et al., 2015), innovation similarities (Lee et al., 2019), shared directors (Burt et al., 2020), geographic links (Parsons et al., 2020), and shared analyst coverage (Ali and Hirshleifer, 2020). The prevailing methodology in studying firm connections often hinges on objective criteria. Our research seeks to enrich the understanding of dynamic and highly time-varying firm connectedness by introducing a subjective approach inspired by the "wisdom of crowds." An example that illustrates our motivation is thematic investing, an important method for fundamental analysis—especially among institutional investors. Thematic investing emphasizes the identification and selection of firms sharing common themes. Market themes often transcend specific industries, regions, or production relationships, making conventional classification insufficient for capturing inter-firm connectedness timely. We attempt to show that harnessing the collective intelligence gleaned from social media could enable us to reveal and comprehend stock clusters.

Compared with previous studies, our contributions are threefold. First, we provide

Electronic copy available at: https://ssrn.com/abstract=4536567

a novel perspective to map firm networks through social media discourse. We uncover incremental information on connections between firms that is seldom visible in documented network forms, demonstrating these links are not efficiently incorporated into market prices. Second, we shift focus from whether social media content is informational or noise, to how it is categorized and perceived by users. Even in cases of misinformation, such as the spread of fake news about a stock, our study finds that the networks we identify on platforms like Reddit maintain their integrity, suggesting the resilience of these connections against misinformation. Third, we position social media as a venue for sharing opinions, using its network structure to reveal economic relationships. This approach distinguishes our work from prior studies by integrating a collective view of how firms are interconnected, thus proposing a new dimension through which social media influences the flow, compilation, and integration of market news.

The rest of the paper is organized as follows: Section 2 describes the Reddit data and the construction of Reddit linkages. Section 3 examines the fundamental comovement of Reddit-linked firms. Section 4 presents the results on the lead-lag returns relationship among Reddit-linked firms. Section 5 provides additional robustness tests. Section 6 concludes.

2 Data and variable construction

Reddit is made up of over 3 million subreddits, each of which is named "r/subreddit_name" and usually focuses on a certain field. The subreddit itself is composed of different threads, each of which focuses on a more specific topic launched by a Reddit member. We rely on the *r/wallstreetbets* subreddit in later analysis.

2.1 Reddit data

We collect text-form comment data from the *r/wallstreetbets* subreddit via Pushshift, a social media data collection, and archiving platform. Pushshift has collected Reddit data since 2015. A more detailed description of the dataset is provided by Baumgartner et al. (2020). To identify comments that are related to stocks, we first tokenize the body of all comments in *r/wallstreetbets* and then cross-reference these tokenized texts with the list of tickers to infer which might be connected to stock tickers. The list of tickers consists of common stocks (share codes 10 or 11) traded on NYSE/AMEX/NASDAQ. We obtain stock-related data from CRSP and accounting information from COMPUSTAT. We describe our approach in more detail in Appendix A.

Electronic copy available at: https://ssrn.com/abstract=4536567

We choose Reddit as our research subject for several reasons. First, a growing body of research provides evidence on the informativeness of Reddit activities,[4] while the "GameStop Short Squeeze" event demonstrates how consensus and coordination emerge from online discussions and its implications for stock prices (Mancini et al., 2022). Our construction of social media linkages is firmly supported by this branch of studies. Second, the meteoric ascent of Reddit's user base enables us to gain insights into how people use social media and the intricate interactions taking place on the platform. Third, in contrast to other peer-reviewed websites, *r/wallstreetbets* moderators often refrain from taking an active role in the discussion of each thread. Here, registered Reddit users can upvote or downvote posts and comments, which affects their ranking position. Posts must receive votes and comments to stay at the top of the page; otherwise, they might quickly disappear in the flood of fresher posts. This setting helps to prevent some social media platform regulators from having the power to censor speech and steer users in certain directions. Fourth, the extensive scope of Reddit data presents an opportunity for investigating connections in a variety of forms and at multiple levels of granularity. Unlike Reddit, Twitter has limitations due to the volume and bias inherent in its API. For instance, the Twitter API v2 imposes a monthly limit of 10 million tweets and a rate limit of 100 requests every 15 minutes.[5]

Following prior literature, we impose several filters on the Reddit dataset: (1) drop comments that are deleted and removed by the author or Reddit; (2) drop comments missing information about the author, the content, or the timestamp when comments are created; (3) drop comments that do not mention any ticker symbol;[6] (4) drop comments with the author being "AutoModerator" or have been deleted;[7] (5) keep only one comment for duplicated author-time-content observations. In addition, we exclude GameStop (stock ticker: "GME") from our sample due to its abnormal frequency of mentions on Reddit.[8] Our final sample consists of 5,559,973 text-form observations, written by 541,395 Reddit users under 269,075 threads, and ranges from July 1, 2018, to December 31, 2022. [9]

[4]See, for example, Bradley et al. (2023) and Hu et al. (2023).

[5]For further information on the Twitter API, please refer to https://developer.twitter.com/en/docs

[6]The definition of mentioning stock tickers is based on whether the comment contains any string of a ticker, and we keep tickers with at least 3 letters to avoid ambiguities in the matching process. For example, if a comment describes "AAPL is soaring," we match it to *Apple Inc*. By contrast, we do not associate a comment like "BA raises its airline capacity" with *Boeing Co*.

[7]"AutoModerator" refers to the Reddit bot to maintain the community order.

[8]Our analysis is not affected by this filter and we get quantitatively similar results when including GameStop in the sample.

[9]In each thread on Reddit, the original submission is initiated by a user about a particular topic (the Level 0 comment). Then users could comment on this submission about this topic (Level 1 comments).

Electronic copy available at: https://ssrn.com/abstract=4536567

2.2 Definition of Reddit peers

Our primary work is to identify stock linkages from social media activities. As described previously, a thread on Reddit is a collection of comments on a specific topic, where comments are made by Reddit authors (i.e., subscribed users). Therefore, we consider two forms of Reddit connection: the Reddit *thread* link and the Reddit *author* link. Specifically, we define two stocks as Reddit thread (author)-connected if there was at least one thread (author) on Reddit mentioning both firms in the past six months. The co-mention of firms in the same thread reflects the collective interest of Reddit crowds about the underlying stocks; the co-track of firms by the same author reflects Reddit users' aggregated attention and knowledge about the stocks featured.

We use a screenshot of a real Reddit thread to help better understand the construction of Reddit peers, as shown in Figure 3. The poster expressed its viewpoint on INTC (Intel Corp.) in the original submission. Another user "u/soc****cious" in the same thread mentioned AAPL (Apple Inc.), AMD (Advanced Micro Devices, Inc.), and MSFT (Microsoft Corp.). Therefore, by our construction, we identify six stock pairs, and each of them is defined as Reddit thread-connected: {INTC, AMD}, {INTC, AAPL}, {INTC, MSFT}, {AMD, AAPL}, {AMD, MSFT}, and {AAPL, MSFT}. Similarly, we also separately aggregate each Reddit author's comments using a six-month rolling window and extract co-mention relationships to define Reddit author links.

By construction, the Reddit thread (author) peer for each focal firm is the set of firms that are also mentioned in the same thread (by the same author) in the preceding six months. We identify and update Reddit connections each month. Therefore, our Reddit peer data ranges from December 2018 to December 2022.

The distinction between the two types of Reddit connections could help better understand the structure of social media networks. Users, who are the primary content-generating units of social media networks, function as both information senders and receivers. They express their opinions and share information on the platform, and the threads serve as collections of these views and reflect the collective perception of the users. The construction of Reddit author linkages inherently encompasses more personal sentiments and opinions, while Reddit thread linkages can be seen as the aggregation of information and reveal firms' economic connectedness based on the opinions of different

They could also comment on existing Level 1 comments (Level 2 comments). The depth of the comments in the thread could extend indefinitely. In our study, we do not differentiate between initial submissions and various levels of comments and treat them all as equal. Our sample period is constrained by data quality, since Reddit users in our dataset can be identified precisely and uniquely only after July 1, 2018.

Electronic copy available at: https://ssrn.com/abstract=4536567

users. In the subsequent part of this paper, we examine both the proxy of the collective knowledge of users, represented by Reddit thread linkages, and the proxy of the private perspectives of Reddit users, represented by Reddit author linkages.

2.3 Summary statistics

Table 1 presents summary information on the composition of our Reddit comments data. On average, each thread in our sample mentions 5 stocks, collectively by 21 replies from 13 Reddit users. Each author (comment) in our sample on average mentions 4 (1) stocks. Panel B shows that 14% of comments, 52% of threads, and 51% of authors in our sample mention at least two stocks, which illustrates the feasibility of the construction of Reddit linkages. Furthermore, the distribution of users, comments, and mentioned stocks in each thread is not uniform and is likely to follow a power-law trend. The phenomenon of "superstars," discussed in Rosen (1981), seems to be at play similarly in Reddit, where relatively small numbers of authors and opinions earn enormous amounts of attention and dominate the topics in which they engage. Combined with some views that believe Reddit user Keith Patrick Gill (often referred to as "u/Deep******Value" on Reddit) triggered the "GameStop Short Squeeze," we can gain further insight into Reddit's power-law-like pattern.

Table 2 reports summary statistics of Reddit thread connections (Panel A) and the Reddit author connection (Panel B). The first three rows of each panel report distributions of the number and size of Reddit-linked stocks across the sample period. The remaining two rows of each panel report the time-series averages of cross-sectional statistics of the corresponding variables. On average, our Reddit thread (author) connection sample covers 56.1% (54.2%) of the stock universe in terms of number and 79% (78.5%) in terms of market capitalization, which indicates that the discussion on the Reddit forum frequently focuses on large firms. Each focal firm on average links to 576 firms through shared Reddit threads, and the most central focal firm has 2042 peer firms. Each focal firm connected by Reddit authors on average links to 277 firms, while the most central focal firm has 1608 peer firms. This demonstrates that Reddit thread-connected firms often have more connections than those connected by authors. Each thread-connected stock pair has an average of 7 common threads, and each author-connected stock pair has an average of 7 common authors.

We also rank stocks based on market capitalization and assign them to deciles using NYSE breakpoints. Figure 4 illustrates the size distribution of stocks with Reddit linkages. The average size rank of firms with Reddit connections is around 4.6, and the overall

Electronic copy available at: https://ssrn.com/abstract=4536567

distribution aligns with the market deciles. This result demonstrates that our sample is not dominated by particularly large or small companies.

To further inspect the composition of Reddit linkages, we report the overlap of Reddit links with alternative economic connections in Figure 5. Specifically, we consider industry links, geographic links, shared analyst coverage (Ali and Hirshleifer, 2020), and supply chain links (Cohen and Frazzini, 2008).[10] On average, we find that around 14% to 16% of Reddit-linked stock pairs are in the same industry (classified by one-digit SIC codes), and about 10% to 12% are located in the same area (identified by one-digit ZIP codes). The overlaps become much smaller when examining industry and geographic relationships at more granular levels. In addition, 1.61% (2.18%) of Reddit thread (author)-linked pairs are also covered by at least one common analyst, and only 0.06% (0.08%) of Reddit thread (author) stock pairs belong to the customer-supplier relationship. Overall, while Reddit linkages indeed capture explicit economic connections to some extent, our sample suggests that communications on social media platforms such as Reddit potentially reveal inter-firm relations beyond traditional definitions.

As a last check of the sample, in Figure 6 we plot the time-series average of cross-sectional percentiles of the number of Reddit connections for stock pairs based on shared threads and shared authors, respectively. Consistent with the pattern that we mentioned previously, the distribution of shared Reddit threads and shared Reddit authors at the stock-pair level resembles a power-law distribution. While most of the observations are 1 or 2 below the 80th percentile, those that are above the 80th percentile start to show a significant increase. Given the high skewness of the data, we use the logarithm transformation of the number of shared Reddit threads/authors, instead of the raw number, as the weight in calculating Reddit peer firms' variables in later analyses.

3 Fundamental comovement of Reddit-linked firms

We begin by testing whether Reddit-linked firms have correlated fundamentals. Reddit users are often associated with retail investors. However, several studies have already discussed how collective judgment, or the "wisdom of crowds," is different from personal opinions, and how the coordination among small participants impacts the whole

[10] Following Ali and Hirshleifer (2020), a stock is identified as covered by an analyst if the analyst issues at least one FY1 or FY2 earnings forecast for the underlying firm over the past year. The analyst forecast data is obtained through the International Brokers Estimate System (IBES). We follow Cohen and Frazzini (2008) and identify the customer-supplier links using data from the Compustat customer segment files.

Electronic copy available at: https://ssrn.com/abstract=4536567

market.[11] Studies based on Reddit, such as Hu et al. (2023) also provide supporting evidence that variables derived from Reddit activities are informative. Pedersen (2022) develops a theoretical model and shows how social networks influence prices before the information is fully known. These studies support our use of Reddit information to discover economic connections between firms.

We consider a broad set of characteristics that capture the profitability, valuation, growth potential, and investment activity of a firm: return on assets (ROA), return on equity (ROE), gross profit (GP), earnings-to-price ratio (EP), book-to-market ratio (BM), cash flow-to-price ratio (CP), sales-to-price ratio (SP), sales growth (SG), profit growth (PG), revenue growth (RG), asset growth (AG), leverage ratio (LEV), asset turnover (AT), and R&D expense-to-sales ratio (RDS). A detailed description of the fundamental variables is provided in the Appendix.

3.1 Contemporaneous comovement

Following Parsons et al. (2020) and Peng et al. (2023), we conduct the following panel regression:

$$F_{i,j,t} = \beta_1 F^R_{i,t} + \beta_2 F^I_{j,t} + \gamma_t + \varepsilon_{i,j,t} \tag{1}$$

in which $F_{i,j,t}$ is a fundamental variable of firm i in industry j at quarter t.[12] The independent variable of interest is the weighted average of Reddit peers' corresponding fundamentals, $F^R_{i,t}$, and the weights are determined by the number of common Reddit threads (authors) shared with a focal firm i:

$$F^R_{i,t} = \frac{1}{\sum_{p=1}^{N} w_{p,t}} \sum_{p=1}^{N} w_{p,t} F_{p,t}, \quad w_{p,t} = \log(1 + \#\text{Reddit connection}). \tag{2}$$

In regressions, we also control for the average fundamental of industry peers $F^I_{j,t}$ and include time-fixed effects (γ_t).

Panel A of Table 3 shows that focal firms comove positively and significantly with their Reddit thread peers along 13 out of 14 dimensions, with the only exception of profit growth. The economic magnitudes are also significant. For example, a one standard deviation increase in Reddit thread peer's ROA and sales growth is associated with an

[11] For example, Chen et al. (2014) discuss how investors' opinions transmission impact the stock market, and Allen et al. (2023) discuss how investor coordination on social media platforms may have fueled a series of short squeezes.

[12] Industries are defined based on two-digit SIC codes.

Electronic copy available at: https://ssrn.com/abstract=4536567

increase in the focal stock's contemporaneous ROA and sales growth of about 0.86% and 3.66%, respectively. The fundamental comovement of Reddit author peers is slightly stronger. Panel B of Table 3 shows that focal firms have significantly similar fundamentals to their Reddit author peers along all examined dimensions. On average, the economic magnitude of the Reddit author peers' fundamental comovement is larger relative to that of the Reddit thread peers.

3.2 Intertemporal relationships

We next examine whether Reddit peers' fundamentals serve as a useful predictor of focal firms' future performance. Specifically, we conduct similar regressions as equation (1) by replacing the dependent variable with one-quarter ahead fundamental variables. We also use focal firms' contemporaneous fundamentals as an additional control variable.

Table 4 reports the results. For the Reddit thread links, we find a strong predictive ability of peer fundamentals in terms of ROA, ROE, EP, SG, PG, and RG. For example, a one standard deviation increase in Reddit peer sales growth is associated with an increase of 1.83% in *future* firm sales growth. As a comparison, the estimated coefficient on industry sales growth is insignificant. We find similar, albeit weaker, results using Reddit author links. Panel B of Table 4 shows that the fundamentals of Reddit author-based peers are a good predictor in terms of profitability (ROA, ROE, GP) and valuation (EP, CP). Although the predictive power based on SG, PG, RG, AG, and LEV is weak, we still find a positive relationship intertemporally among Reddit-linked firms, and the industry fundamental is less important in the prediction.

Overall, the result suggests that Reddit linkages help to capture fundamental connectedness between firms. The network identified through Reddit is not likely to be a simple retrospective of historical firm performance but provides forward-looking information about future fundamentals and business interactions.

4 Lead-lag returns relationship of Reddit-linked firms

This section examines the asset pricing implications of Reddit linkages. Specifically, we investigate whether market participants sufficiently explore the information underlying the inter-firm connections implied by social media. To this end, we first test the cross-period relationship of Reddit-linked stocks' returns. Prior studies document significant lead-lag effects among economically connected firms, starting with the early work of

Electronic copy available at: https://ssrn.com/abstract=4536567

Moskowitz and Grinblatt (1999) and Cohen and Frazzini (2008). Following this literature, if investors fully recognize the Reddit connection and that the pricing of Reddit-linked stocks is efficient, then the Reddit peer's price movement should be unrelated to the focal stock's future return. Instead, if investors fail to incorporate the information implied by Reddit peers' returns, then we expect to see a lead-lag returns relationship between stocks that are Reddit-connected.

4.1 Portfolio analysis

To examine whether returns spillover among Reddit-linked firms, we construct a return signal for each type of Reddit linkage:

$$REDDIT\ RET_{i,d} = \frac{1}{\sum_{p=1}^{N} w_{p,m}} \sum_{p=1}^{N} w_{p,m} Ret_{p,d}, \quad w_{p,m} = \log(1+\#\text{Reddit connection}), \quad (3)$$

where $Ret_{p,d}$ is stock p's daily return on day d, month m. We use the most recent linkage information to calculate $REDDIT\ RET_{i,d}$, hence the weight for each Reddit peer ($w_{p,m}$) remains constant throughout month m. For example, consider any d in July 2021. In this case, month $m-1$ is June 2021, and Reddit linkages are identified using Reddit comments data from January 2021 to June 2021. Each day, stocks are sorted into quintile portfolios based on REDDIT RET. We then calculate value-weighted and equal-weighted returns for each portfolio on the next day. To alleviate microstructure issues, we exclude stocks with a share price below $1 in any of the past five days before portfolio formation. We also use the average market value in the past five days when calculating value-weighted returns, which helps mitigate the influence of short-term fluctuations in share prices.

Panel A of Table 5 reports the result of the Reddit thread link. We find that the Reddit thread peer's return positively predicts the focal stock's future return. A long-short strategy based on REDDIT RET generates a return of 5.24 basis points (t=2.68) per day. The return remains significant after adjusting for the five-factor model (Fama and French, 2015) or the momentum-augmented model (FF6). Panel B presents the result of the Reddit author link, which suggests a similar lead-lag effect. For example, the value-weighted strategy based on REDDIT RET earns a return of 5.23 bps (t=2.56) on average, and the equal-weighted return is 8.91 bps and highly significant (t=4.65).

In a related study, Peng et al. (2023) utilize the Social Connectedness Index (SCI) from Facebook, and construct "social ties" among industry peer firms. They find that the thus-defined socially-weighted industry peer return is a strong predictor for focal stocks'

Electronic copy available at: https://ssrn.com/abstract=4536567

future returns. Our design and results are different from Peng et al. (2023) in several aspects. First, the SCI measure captures social connectedness between counties, whereas our Reddit network is silent on such *friendship* links; we emphasize that the commonality in individual users' social media activities reveals economically meaningful information. Second, while Peng et al. (2023) focus on within-industry connections, the Reddit network covers both cross- and within-industry links, as illustrated in Figure 5. Finally, we would show in the robustness section that the Reddit lead-lag effect remains significant after controlling for a large set of alternative economic links.

4.2 Regression analysis

Next, we control for other firm characteristics in regressions and examine the predictive ability of REDDIT RET. Given the constrained volume of Reddit r/wallstreetbets threads and data availability, our dataset is limited to a four-year length. Consequently, we align with the methods of Chen et al. (2014) and Hu et al. (2023) to estimate the following panel regression:

$$Ret_{i,d+1} = \beta_0 + \beta_1 REDDIT\ RET_{i,d} + \beta_2 Controls_{i,d} + \gamma_d + \varepsilon_{i,d+1}. \tag{4}$$

The dependent variable $Ret_{i,d+1}$ is focal stock i's return on day $d+1$. The main dependent variable of interest is $REDDIT\ RET_{i,d}$, the one-day lagged Reddit peer returns. We include day fixed effects (γ_d) and control for focal stock's past performance, including the return on the day d, the cumulative return during $[d-20, d-2]$, and the cumulative return during $[d-120, d-21]$. We also control for characteristics such as illiquidity, idiosyncratic volatility, the log of market capitalization, and the log of the book-to-market ratio measured at the end of last month.[13]

Table 6 reports the estimated coefficients. Consistent with the portfolio results, we find significant daily predictive ability under the Reddit linkages. Columns (1) and (5) of Table 6 indicate that 6.03% (3.95%) of the prior day's Reddit thread (author) peer return carries over into the following day's return of the focal stock, after controlling for other firm characteristics. Although we have shown that Reddit linkages do not overlap heavily with other economic connections, an important concern of our result is that the Reddit lead-lag effect is driven by alternative linkages. In columns (2) and (6), we control for

[13] In our main analysis, we use equal-weighted regressions, cluster standard errors at the firm level, and include day-fixed effects. Our results remain robust to using value-weighted regressions, two-way clustering by both firm and date and firm fixed effects. These results are provided in the Appendix Table A1 and A2.

Electronic copy available at: https://ssrn.com/abstract=4536567

the industry peer return (IND RET), calculated as the value-weighted one-day return of other stocks with the same two-digit SIC codes. It turns out that the estimated Reddit momentum remains highly significant.

We further include co-analyst peer return (CF RET) in columns (3)-(4) and columns (7)-(8) to control for the effect of shared analyst coverage, which is shown to potentially unify most economic linkages (Ali and Hirshleifer, 2020). We follow the same method in Ali and Hirshleifer (2020) in calculating CF RET. Each month, we define two stocks as connected if at least one analyst issued an FY1 or FY2 earnings forecast for both stocks in the past 12 months. For each focal stock, the co-analyst peer return is computed as the average of linked stocks' one-day return, weighted by the number of analysts shared with the focal stock. We use the most recent month's analyst linkage information for calculations. Notably, while the effect of industry peer return is largely reduced by CF RET, the predictive ability of REDDT RET remains highly significant. For example, column (4) of Table 6 shows that the estimated coefficient on Reddit thread peer return is 3.940 (t=4.96), whereas the coefficient on industry peer return is 1.737, which is about 38% the size of the coefficient in column (2). In sum, our result suggests that the Reddit linkage helps to capture incremental information beyond industry classification and analyst coverage.

We also examine the predictive ability of REDDIT RET for returns over longer horizons. Specifically, we use the regression specification of equation (4) and replace the dependent variable $Ret_{i,d+1}$ with cumulative returns during $[d+1, d+n]$. Figure 7 shows the estimated coefficients graphically. We find that peer stock returns from Reddit linkages continue to predict the focal stock's long-term returns significantly, up to 20 trading days. The estimated coefficient on REDDIT RET is comparably smaller for the Reddit author link than for the Reddit thread link. Nevertheless, the two types of Reddit peer returns exhibit persistent predictive ability for long-horizon returns and do not display significant reversals. This result is consistent with the story that the lead-lag returns relationship among Reddit-linked firms is mainly driven by the slow diffusion of information instead of transient price pressure from investor sentiment.

4.3 Tests of the limited attention hypothesis

The prevailing explanation for the cross-firm return predictability is the underreaction hypothesis. Many studies suggest that, due to limited attention, investors often underreact to news conveyed by the stock returns of peer firms, which leads to prices not completely incorporating recent information. Consequently, the information contained in Reddit peers' price movements would be sluggishly disseminated into the market, which

Electronic copy available at: https://ssrn.com/abstract=4536567

gives rise to the lead-lag returns relationship. This section tests the limited attention hypothesis in explaining the lead-lag effect of Reddit-linked firms.

We use analyst coverage and media coverage as attention proxies.[14] If the underreaction to the information contained in Reddit peer's return generates the cross-firm predictability, then the predictive power of REDDIT RET should be weaker when investor attention is high. We also consider the effect of institutional ownership.[15] The prediction is that underreaction-induced mispricing should be mitigated when there are more sophisticated investors trading the stock. In addition, the predictive ability of REDDIT RET would be stronger when the number of Reddit peer firms is high, since, in that case, investors require more attention to comprehend the linkage information. Throughout all specifications, we control for the interaction effect of size because larger firms tend to attract more attention, have more institutional investor participation, and are connected to more Reddit peers.

Table 7 reports the estimated coefficients of regressions. Panel A presents the result of the Reddit thread link and panel B shows the Reddit author link. When the attention to the focal stock is high, measured by a high level of analyst coverage or media coverage, the estimated effect on REDDIT RET is 24%-29% weaker for the Reddit thread link and 20%-28% weaker for the Reddit author link. This result is consistent with the Limited Attention Hypothesis. The estimated coefficient on the interaction term of REDDIT RET and the variable indicating high institutional ownership (Inst High) is negative, suggesting that the Reddit peer momentum is less evident for firms with more institutional holdings. We also find a stronger predictive ability of Reddit peer returns when the number of Reddit linkages is high. The interaction of REDDIT RET and the dummy indicating a high linkage number is positively significant. The stock return of Reddit peers carries over into the focal stock's future return by 1.48% (2.30%) more if the number of Reddit thread (author) linkages is above the medium. Overall, the results are consistent with the explanation that investors underreact to peer firms' price movements, which leads to sluggish information diffusion among Reddit-linked firms.

[14]For media coverage, we use news data from the RavenPack database. Following previous literature, we require an Event Novelty Score (ENS) and a relevance score of 100. Each month, the media coverage of a stock is the number of news stories mentioning the stock.

[15]We obtain quarterly data on institutional investor (13F) holdings from Thomson-Reuters.

Electronic copy available at: https://ssrn.com/abstract=4536567

4.4 Returns around earnings and news days

To further inspect the mechanism underlying the cross-firm return predictability from Reddit linkages, we examine focal stocks' returns around their information release events, such as earnings announcements or news stories. For each stock, we first identify the day with an earnings announcement or news release.[16] Following the method of Engelberg et al. (2018), we examine the focal stock's trading volume (scaled by the market trading volume) for a three-day window centered on the event date and define an earnings day (*Eday*) or news day (*Nday*) as the day with the highest trading volume. Then, we perform the following regression:

$$\begin{aligned} Ret_{i,d+1} =& \beta_0 + \beta_1 REDDIT\ RET_{i,d} + \beta_2 REDDIT\ RET_{i,d} \times Eday_{i,d+1} \\ &+ \beta_3 REDDIT\ RET_{i,d} \times Nday_{i,d+1} + \beta_4 Eday_{i,d+1} + \beta_5 Nday_{i,d+1} \\ &+ \delta Controls + \gamma_d + \varepsilon_{i,d+1}. \end{aligned} \tag{5}$$

As specified, the variable $Eday_{i,t+1}$ ($Nday_{i,t+1}$) is a dummy variable equal to one on the focal stock i's earnings (news) days and zero otherwise. The set of control variables includes lagged values for each of the past ten days of stock returns, return squared, and trading volume. The initial mispricing story implies that the focal stock's price deviates from the rational benchmark as the information conveyed by peer stocks' returns diffuses slowly. The release of information helps correct the biased expectation about the focal stock's fundamentals; as a result, the underpricing (overpricing) of high (low) REDDIT RET stocks is mitigated, and prices converge to the fundamental value, resulting in higher returns on these days. Therefore, we would expect the estimated coefficients β_2 and β_3 to be positive.

Table 8 presents our estimation results. We find that subsequent returns on earnings or news days are significantly larger for stocks with higher REDDIT RET. Column (1) shows that the effect of Reddit peer returns is about 1.95 times higher on earnings days than on non-earnings days. In column (3), the estimated coefficient on $REDDIT\ RET \times Nday$ is 5.045, suggesting the effect is about 1.26 times stronger on news days than on non-news days. Overall, we find higher predictive returns on both earnings days and news days, although the statistical significance varies depending on specifications. This result is consistent with the findings of Engelberg et al. (2018) and supports the underreaction story of the lead-lag effect among Reddit-linked firms.

[16]The earnings announcement dates are obtained from the Compustat fundamental quarterly database. We obtain news story dates from the RavenPack database.

Electronic copy available at: https://ssrn.com/abstract=4536567

5 Additional robustness tests

5.1 The information content of Reddit peer returns

We have shown that Reddit-linked firms are fundamentally connected and the Reddit linkage generates cross-firm return predictability. The underlying premise of this finding is that Reddit peers' price movements contain useful information about the focal firm's future performance. In this section, we examine the information content of Reddit peer returns using focal firms' standardized unexpected earnings (SUE), which captures unanticipated changes in cash flow. We calculate SUE as unexpected earnings scaled by the standard deviation of unexpected earnings over the preceding eight quarters. The unexpected earnings are measured by year-over-year changes in quarterly earnings before extraordinary items.

Table 9 presents the result of regressing the focal firm's SUE on lagged Reddit peer quarterly return (REDDIT QRET). Panel A shows that Reddit thread peer return has a significant predictive ability for the focal firm's future SUE. In the univariate regression, column (1) shows that a one standard deviation increase in REDDIT QRET implies an increase of SUE in the next quarter of 0.028. The estimated coefficient remains significant after controlling for the focal firm's lagged SUE. Panel B reports the result of the Reddit author link, and we again find a strong predictive power of REDDIT QRET. This result suggests that the stock return of Reddit peer firms does contain fundamental information about the focal firm.

5.2 Uncertainty and Reddit Momentum

As examined in previous sections, the lead-lag effect of Reddit peer firms is likely to be driven by investors' underreaction to peer firms' news, which generates a delayed price reaction of the focal stock. In this section, we examine the time variation of the predictive power of peer firms' returns. Specifically, we use the Chicago Board Options Exchange Volatility Index (VIX) to predict the return of the Reddit momentum strategy. Since investors are likely to be distracted in periods of enhanced uncertainty (Jiang et al., 2021), the Reddit momentum effect should be stronger following a high level of VIX.

Table 10 reports the time-series regression results. Column (1) shows that a one standard deviation increase in the last day's VIX implies that the Reddit peer momentum return increases by 4.79 bps. This effect is robust to adjusting for the six-factor model. For the Reddit author link, column (3) shows that the following day's strategy return

Electronic copy available at: https://ssrn.com/abstract=4536567

increases by 8.35 bps when VIX increases by one standard deviation. Overall, we find a positive and significant relationship between uncertainty and the Reddit lead-lag effect, which in turn supports the Limited Attention Hypothesis.

5.3 Order imbalance, Robinhood, and Reddit peer returns

In addition to the underreaction hypothesis, researchers also document significant lead-lag relationships that are driven by *continued overreaction*. For example, Chen et al. (2023) present the "attention spillover" effect of stocks with adjacent listing codes in China's stock market; He et al. (2023) show that stocks with similar characteristics exhibit cross-firm return predictability, which is consistent with investors' categorial thinking or experience effect. Guo et al. (2022) find that joint news coverage leads to attention-driven overvaluation. Under this channel, the predictive ability of Reddit peer returns is the result of excess demand for the focal stock. That is, the predictive ability of Reddit peer firms' returns reflects the buildup of mispricing.

We examine this channel by testing whether Reddit peer returns can predict the order imbalance and the change in Robinhood ownership of the focal stock. If the lead-lag effect of Reddit-linked firms is driven by investors' continued overreaction, then we should observe a significantly positive relationship between Reddit peer returns and subsequent excess demand. We calculate daily order imbalances using data from TAQ. The trade imbalance is calculated by the difference in the number of buys and sells divided by the total number of buys and sells. Similarly, we also calculate the volume imbalance and the dollar imbalance.[17]

Robinhood (RH) is an online retail brokerage company founded in 2013 and registered with the U.S. Securities and Exchange Commission. RH is a FINRA-regulated broker-dealer and a member of the Securities Investor Protection Corporation. RH is devoted to offering a simple and cheap channel for small investors to participate in the stock market, and its customers are widely believed to be a new generation of young, computer-savvy, but novice investors (Welch, 2022). We downloaded data on the number of RH users who held a particular stock from *Robintrack.net*. If Reddit communications lead to attention spillovers among connected stocks and generate unjustified demand, then Reddit peer returns should be positively associated with future change in Robinhood ownership.

[17]Specifically, the volume imbalance is shares of buy trades minus shares of sell trades divided by the total volume of buys and sells. The dollar imbalance is the difference in the dollar value of buys and sells divided by the total dollar value of buys and sells. A buyer-initiated or seller-initiated trade is determined based on the Lee and Ready (1991) test. The data on total buys, total sells, retail buys, and retail sells is available through the Wharton Research Data Services (WRDS) Millisecond Intraday Indicators.

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 11 reports the regression result that uses REDDIT RET to predict one-day ahead order imbalance and change in Robinhood ownership. Across all specifications, we do not detect a significant predictive ability of Reddit peer returns for future excess demand. Although we cannot completely rule out the overreaction hypothesis, this result suggests that the predictive ability from Reddit linkage should not be simply ascribed to the continued irrational trading of the focal stock. Instead, we find supporting evidence in previous analyses that limited attention is more likely to be the source of the cross-firm predictability of Reddit-linked firms.

5.4 Controlling for additional lead-lag effects

We have shown that the Reddit lead-lag returns relationship is primarily driven by investors' underreaction and that the effect is robust to controlling for industry momentum (Moskowitz and Grinblatt, 1999; Hou, 2007) and shared analyst coverage (Ali and Hirshleifer, 2020). However, concerns persist as to whether the predictive ability of REDDIT RET is due to comovement with other existing cross-firm return predictors. Therefore, we consider a large set of economic link variables to further assess the robustness of our result. Specifically, we consider the text-based peers (Hoberg and Phillips, 2016, 2018), technological links (Lee et al., 2019), geographic links (Parsons et al., 2020), the customer-supplier relationship (Cohen and Frazzini, 2008), and conglomerate firms (Cohen and Lou, 2012). In Appendix Table A3 and A4, we repeat our regression analysis by controlling for these additional inter-firm linkage variables. We find that the predictive power of Reddit peer returns remains significant across different specifications.

6 Conclusion

Social media platforms are undoubtedly becoming an increasingly influential venue for the dissemination of opinions and information in today's world. Analogous to a company's geographical location or industry classification, a firm's place in social media may also symbolize its position in the financial network. If two firms are connected in the context of social media, it implies a potential financial interaction. This study demonstrates the possibility of identifying economic linkages from activities within the domain of social media.

In sum, the results in this paper suggest that the collective wisdom of Reddit crowds helps to capture comprehensive and dynamic information about financial connections. It turns out that market participants do not fully comprehend such implicit linkage

Electronic copy available at: https://ssrn.com/abstract=4536567

relationships, thereby leading to cross-firm return predictability. Given the ongoing attention to social media's role in information propagation in modern society, future research may benefit from incorporating multiple levels of social media data, including the author level and the topic level, to provide a comprehensive overview of the nested networks within social media.

Electronic copy available at: https://ssrn.com/abstract=4536567

References

Al Guindy, M., and R. Riordan. 2019. The social internetwork and stock returns. *Available at SSRN 3501915* .

Ali, U., and D. Hirshleifer. 2020. Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics* 136:649–675.

Allen, F., M. Haas, E. Nowak, M. Pirovano, and A. Tengulov. 2023. Squeezing Shorts Through Social Media Platforms. *Working Paper* .

Bailey, M., R. Cao, T. Kuchler, and J. Stroebel. 2018a. The Economic Effects of Social Networks: Evidence from the Housing Market. *Journal of Political Economy* 126:2224–2276.

Bailey, M., R. Cao, T. Kuchler, J. Stroebel, and A. Wong. 2018b. Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives* 32:259–280.

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. 2020. The Pushshift Reddit Dataset. *Available at arXiv:2001.08435* .

Bradley, D., J. Hanousek Jr, R. Jame, and Z. Xiao. 2023. Place your bets? The market consequences of investment research on Reddit's Wallstreetbets. *Working Paper)* .

Burt, A., C. Hrdlicka, and J. Harford. 2020. How much do directors influence firm value? *The Review of Financial Studies* 33:1818–1847.

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

Chen, X., L. An, J. Yu, and Z. Wang. 2023. Attention spillover in asset pricing. *The Journal of Finance* .

Cohen, L., and A. Frazzini. 2008. Economic links and predictable returns. *The Journal of Finance* 63:1977–2011.

Cohen, L., and D. Lou. 2012. Complicated firms. *Journal of financial economics* 104:383–400.

Dim, C. 2023. Social Media Analysts' Skills: Insights from Text-implied Beliefs. *Available at SSRN 3813252* .

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

Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model. *Journal of financial economics* 116:1–22.

Guo, L., L. Peng, Y. Tao, and J. Tu. 2022. Joint news, attention spillover, and market returns. *Working Paper* .

He, W., Y. Wang, and J. Yu. 2023. Similar stocks. *Available at SSRN 3815595* .

Hoberg, G., and G. Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124:1423–1465.

Hoberg, G., and G. M. Phillips. 2018. Text-based industry momentum. *Journal of Financial and Quantitative Analysis* 53:2355–2388.

Electronic copy available at: https://ssrn.com/abstract=4536567

Hosseini, A., G. Jostova, A. Philipov, and R. Savickas. 2020. The Social Media Risk Premium. *Available at SSRN 3514826* .

Hou, K. 2007. Industry information diffusion and the lead-lag effect in stock returns. *The Review of Financial Studies* 20:1113–1138.

Hou, K., C. Xue, and L. Zhang. 2020. Replicating anomalies. *The Review of Financial Studies* 33:2019–2133.

Hu, D., C. M. Jones, V. Zhang, and X. Zhang. 2023. The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Available at SSRN 3807655* .

Jiang, H., S. Z. Li, and H. Wang. 2021. Pervasive underreaction: Evidence from high-frequency data. *Journal of Financial Economics* 141:573–599.

Kuchler, T., and J. Stroebel. 2021. Social Finance. *Annual Review of Financial Economics* 13:37–55.

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

Lee, C. M., and M. J. Ready. 1991. Inferring trade direction from intraday data. *The Journal of Finance* 46:733–746.

Lee, C. M., S. T. Sun, R. Wang, and R. Zhang. 2019. Technological links and predictable returns. *Journal of Financial Economics* 132:76–96.

Li, F. 2022. Retail Trading and Asset Prices: The Role of Changing Social Dynamics. *Available at SSRN 4236966* .

Mancini, A., A. Desiderio, R. Di Clemente, and G. Cimini. 2022. Self-induced consensus of Reddit users to characterise the GameStop short squeeze. *Scientific Reports* 12:13780.

Menzly, L., and O. Ozbas. 2010. Market segmentation and cross-predictability of returns. *The Journal of Finance* 65:1555–1580.

Moskowitz, T. J., and M. Grinblatt. 1999. Do industries explain momentum? *The Journal of Finance* 54:1249–1290.

Parsons, C. A., R. Sabbatucci, and S. Titman. 2020. Geographic lead-lag effects. *The Review of Financial Studies* 33:4721–4770.

Pedersen, L. H. 2022. Game On: Social Networks and Markets. *Forthcoming in Journal of Financial Economics* p. 54.

Peng, L., S. Titman, M. Yönaç, and D. Zhou. 2023. Social Ties, Comovements, and Predictable Returns. *Working Paper* .

Rosen, S. 1981. The Economics of Superstars. *The American Economic Review* Publisher: American Economic Association.

Welch, I. 2022. The wisdom of the Robinhood crowd. *The Journal of Finance* 77:1489–1527.

Electronic copy available at: https://ssrn.com/abstract=4536567

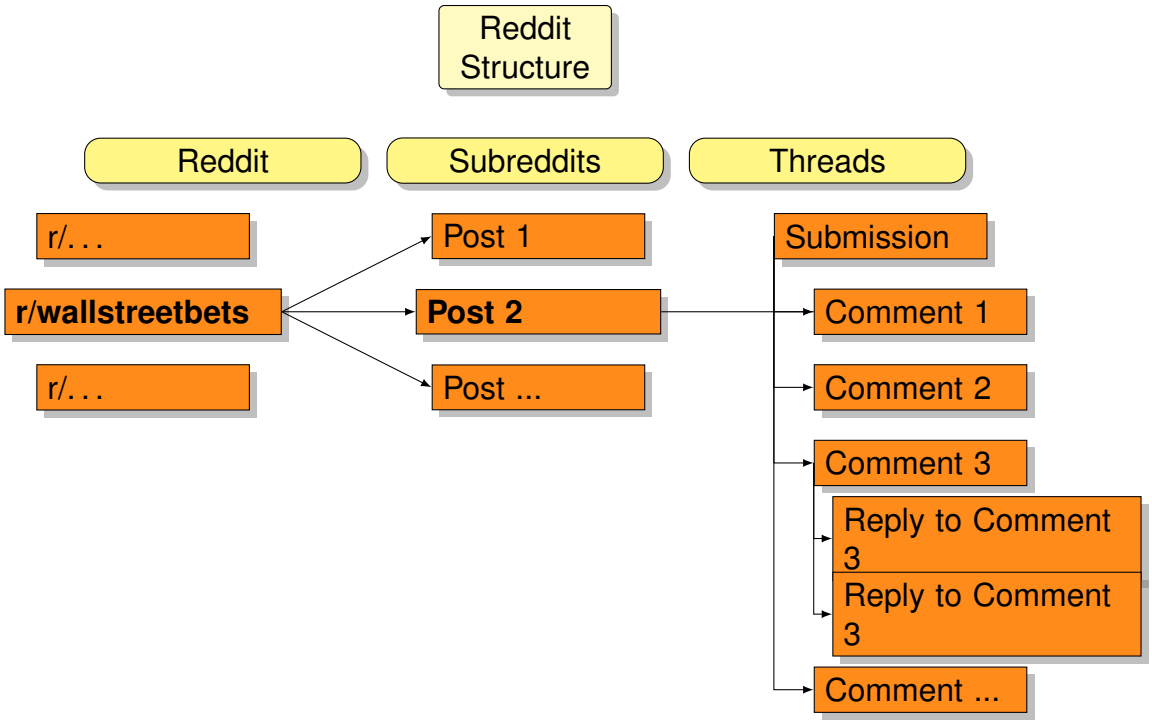


Figure 1. Structure of Reddit.
This figure illustrates the brief structure of Reddit, a popular social media aggregation containing over three million communities known as "subreddits", discussing a wide range of topics. Each subreddit is dedicated to a particular topic, ranging from science to entertainment. On Reddit, "submissions" refer to the content users post within these subreddits, which may include text, links, or images. These submissions initiate discussion threads, where other users can engage by posting comments. Comments are organized into a hierarchical structure, allowing for direct responses to both the original submission and individual comments, creating a multi-layered dialogue.

Electronic copy available at: https://ssrn.com/abstract=4536567

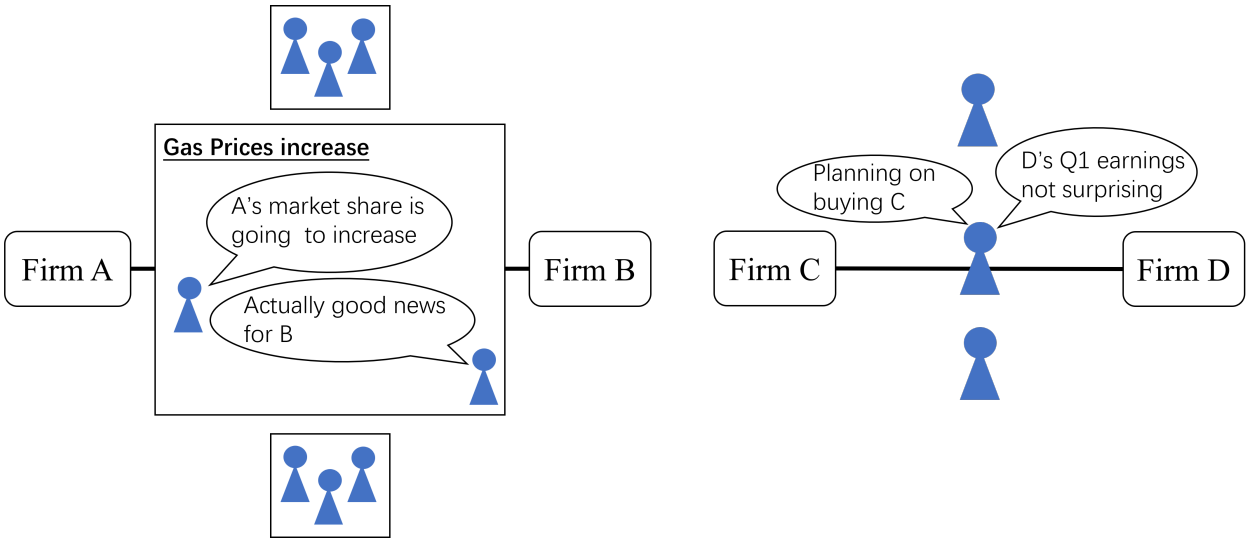


Figure 2. Illustration of Reddit linkages.
This figure illustrates the Reddit thread link (left) and the Reddit author link (right). Firm A and Firm B are connected through shared Reddit threads because they are both referenced in a discussion regarding gas prices. Firm C and Firm D are connected through shared Reddit authors since a Reddit user has published comments about the firms.

Electronic copy available at: https://ssrn.com/abstract=4536567

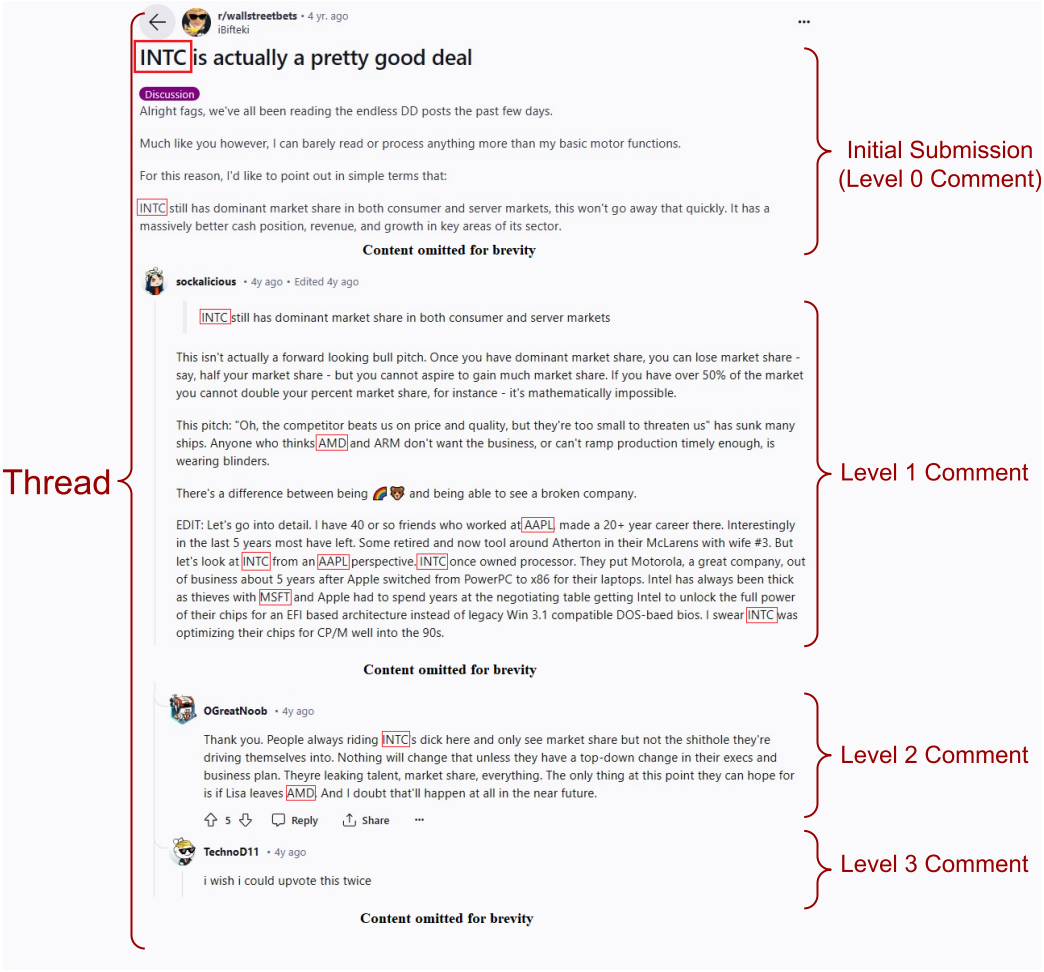


Figure 3. Example of Reddit *r/wallstreetbets* comments.
This figure presents a screenshot capturing selected comments from a Reddit thread, available at `https://www.reddit.com/r/wallstreetbets/comments/i45ao7/intc_is_actually_a_pretty_good_deal`. The discussion is initiated by the original poster with a focus on Intel Corp., and the subsequent comments extend this conversation to other companies, including Advanced Micro Devices, Inc., Apple Inc., and Microsoft Corp. Ticker symbols mentioned within these comments are highlighted with red boxes. We use red brackets to illustrate the terms we used in this paper. Red brackets are used to emphasize key terms defined in this paper. The term "thread" is defined as the entire sequence of the original submission and all the responses. This figure illustrates a four-level reply hierarchy within the comments.

Electronic copy available at: https://ssrn.com/abstract=4536567

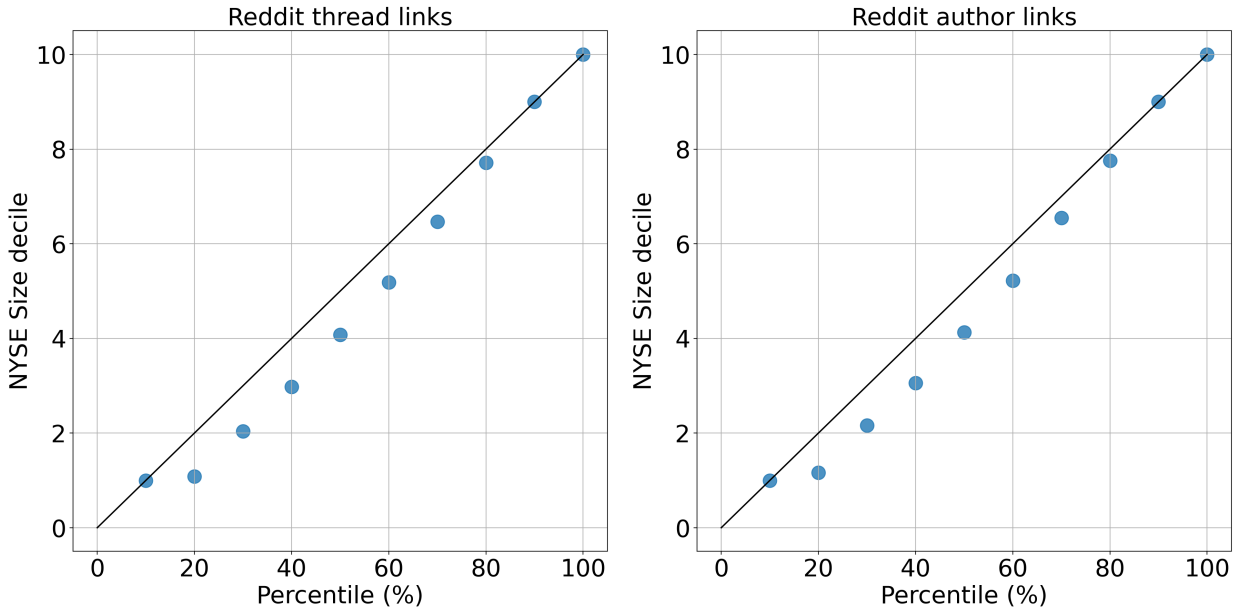


Figure 4. Size distributions of firms with Reddit linkages.
These two graphs plot the time-series averages of the cross-sectional distribution of firm size. Each month, stocks are assigned to deciles based on their market capitalization, using NYSE breakpoints as the standard. We then calculate the size rank distribution of stocks with at least one shared Reddit thread link (left) or shared Reddit author link (right). The x-axis of each graph represents the percentiles of this distribution, showing the range from the smallest to the largest firms. The y-axis corresponds to the NYSE size deciles, categorizing firms based on their relative size. A 45-degree line is plotted for reference. If the plotted points align with the 45-degree line, it indicates that the distribution is not skewed toward larger or smaller firms, suggesting a balanced representation across all size categories. The sample period is from December 2018 to December 2022.

Electronic copy available at: https://ssrn.com/abstract=4536567

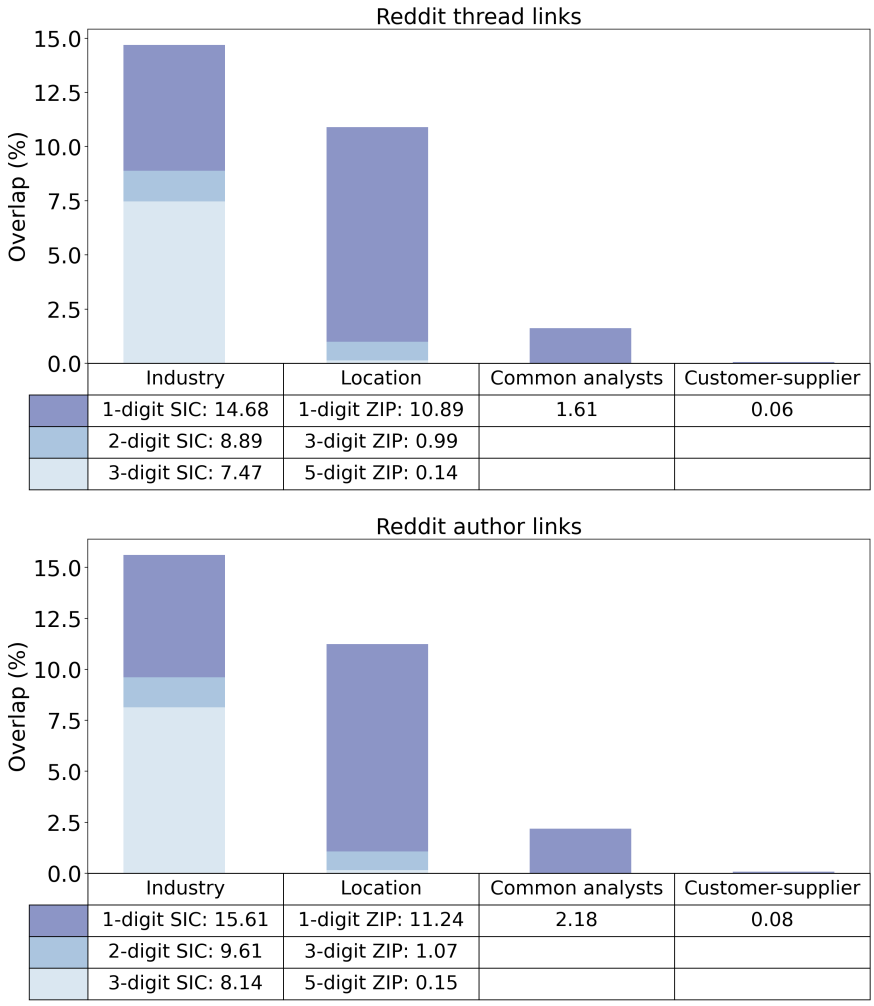


Figure 5. Reddit linkages and economic connections.
These figures report the overlap of Reddit linkages with economic connections from industry classification, headquarters location, shared analyst coverage (Ali and Hirshleifer, 2020), and supply chain (Cohen and Frazzini, 2008). Each month, we first identify stock pairs with Reddit linkages and then calculate the proportion of pairs with alternative economic links. The figure reports the time-series average of proportions. For industry classifications, we consider clusters based on one-digit SIC codes, two-digit SIC codes, and three-digit SIC codes, respectively; for firms' headquarters locations, we consider geographic clusters based on one-digit ZIP codes, three-digit ZIP codes, and five-digit ZIP codes, respectively. Common analyst coverages are constructed following the method of Ali and Hirshleifer (2020). We also identify whether a stock pair belongs to the customer-supplier relationship using the definition of Cohen and Frazzini (2008).

Electronic copy available at: https://ssrn.com/abstract=4536567

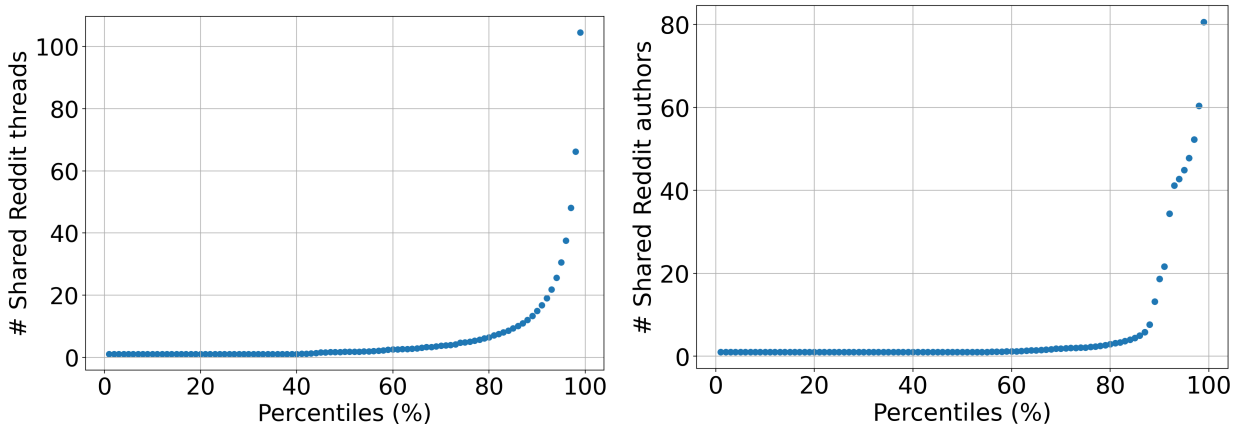


Figure 6. Distribution of shared Reddit threads and shared Reddit authors.
These two graphs plot the time-series averages of cross-sectional percentiles of the number of shared Reddit threads and the number of shared Reddit authors, measured at the stock-pair level. We do not show the maximum value in graphs because of the existence of outliers. For example, the last point in the left-hand-side graph shows the monthly time series average of the cross-section 99th percentile of the number of common threads for each stock pair. The sample period is from December 2018 to December 2022.

Electronic copy available at: https://ssrn.com/abstract=4536567

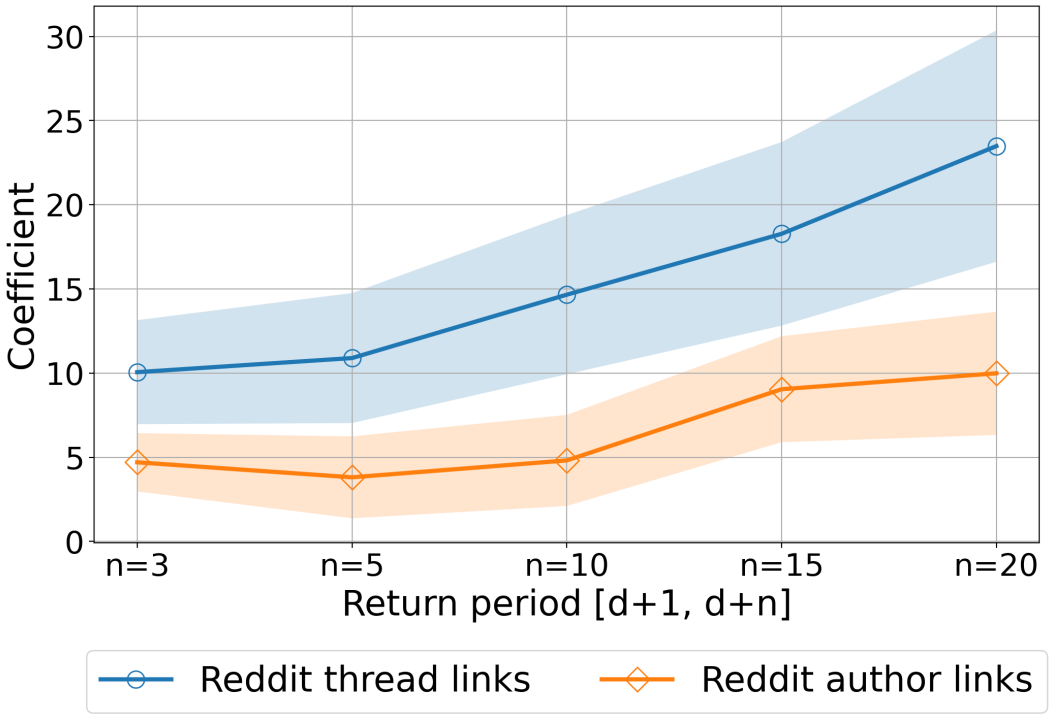


Figure 7. Long-term return prediction.
This figure shows the estimated regression coefficients and 95% confidence intervals predicting future cumulative returns (in percent) during $[d+1, d+n]$ using Reddit peer returns on day d. For each focal stock, the Reddit peer return is the weighted average one-day return of stocks that are linked through Reddit threads (blue) or Reddit authors (orange). Control variables include the focal stock's one-day return, one-month return (skipping the most recent day), cumulative 100-day return (skipping the most recent month), as well as illiquidity, idiosyncratic volatility, the log of market capitalization, and the log of the book-to-market ratio measured at the end of last month. Day-fixed effects are included in regressions and t-statistics, with standard errors clustered at the firm level. The sample period is from January 2019 to December 2022.

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 1. Summary statistics of the Reddit data.

This table reports the summary statistics of the pooled Reddit observation. Our list of stock tickers includes common shares on NYSE/AMEX/NASDAQ. Comments in our sample are related to at least one stock ticker. Panel A reports the statistics of the number of authors and comments in each thread and the stocks mentioned in each thread, author, and comment. Panel B reports the percentage of threads, authors, and comments of the whole sample mentioning at least two stock tickers. The sample is from July 1, 2018, to December 31, 2022.

Panel A. Reddit and stock mentions								
	N	Mean	Std.Dev	Min	P10	Median	P90	Max
# of authors each thread	269075	12.7	120.7	1	1	2	10	9370
# of comments each thread	269075	20.7	234.3	1	1	2	11	19096
# of stocks mentioned each thread	269075	4.9	21.4	1	1	2	7	843
# of stocks mentioned each author	541395	4.2	13.8	1	1	2	8	2552
# of stocks mentioned each comment	5559973	1.2	3.0	1	1	1	2	437

Panel B. Co-mention on Reddit	
	Percentage
threads mentioning at least two stocks	52.1
Authors mentioning at least two stocks	51.0
Comments mentioning at least two stocks	14.0

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 2. Summary statistics of Reddit linkages.
This table reports summary statistics of the Reddit thread link and the Reddit author link. Our list of stock tickers includes common shares on NYSE/AMEX/NASDAQ. Comments in our sample are related to at least one stock ticker. Each month, two stocks are defined as Reddit-linked if there was at least one thread (Panel A) or author (Panel B) on Reddit mentioning both stocks in the past six months. In each Panel, the first two rows report monthly time-series statistics of the respective variables across the sample period. The remaining four rows in each Panel report the time-series averages of cross-sectional statistics of the respective variables. The sample period is from December 2018 to December 2022.

Panel A. Reddit thread linkage								
	N	Mean	Std.Dev	Min	P10	Median	P90	Max
# of stocks with Reddit linkage	49	2207	492.3	1491	1564	2309	2922	3118
% of total number of stocks covered	49	56.1	11.4	41.3	43.4	53	72.7	76.7
% of total market capitalization covered	49	79	3.2	73	74.3	80.3	82.5	82.8
# of Reddit linkage for each focal firm	108157	576	471.5	1	96	438	1301	2042
# of common threads for each stock pair	34429967	7	21.7	1	1	2	15	1612
Panel B. Reddit author linkage								
	N	Mean	Std.Dev	Min	P10	Median	P90	Max
# of stocks with Reddit linkage	49	2135	506.7	1446	1485	2157	2862	3083
% of total number of stocks covered	49	54.2	11.9	40.1	41.3	50.4	71.5	75.5
% of total market capitalization covered	49	78.5	3.4	72.2	73.4	79.6	82.4	82.7
# of Reddit linkage for each focal firm	104597	277	291.8	1	8	164	679	1608
# of common authors for each stock pair	16001187	7	26.9	1	1	1	19	2865

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 3. Fundamental comovement.
This table reports the fundamental comovements between the focal firm and its Reddit peers:

$$F_{i,j,t} = \beta_1 F^R_{i,t} + \beta_2 F^I_{j,t} + \gamma_t + \varepsilon_{i,j,t}.$$

Each month, two stocks are defined as Reddit-linked if there was at least one thread (Panel A) or author (Panel B) on Reddit mentioning both stocks in the past six months. The independent variable ($F_{i,j,t}$) is the focal firms' quarterly fundamental, including return on assets (ROA), return on equity (ROE), gross profit (GP), earnings-to-price ratio (EP), book-to-market ratio (BM), cash flow-to-price ratio (CP), sales-to-price ratio (SP), sales growth (SG), profit growth (PG), revenue growth (RG), asset growth (AG), leverage ratio (LEV), asset turnover (AT), and R&D expense-to-sales ratio (RDS). The independent variable ($F^R_{i,t}$) is the corresponding fundamentals of Reddit-connected firms. We control for average fundamentals of industry peers ($F^I_{j,t}$) and include time-fixed effects (γ_t). All independent variables are winsorized at the 1% and the 99% levels in each cross-section and standardized to have zero mean and unit variance. Standard errors are clustered by both time and firm. t-statistics are reported in parentheses. The sample period is from Q4 2018 to Q4 2022.

	Panel A. Reddit peer: thread					Panel B. Reddit peer: author				
	Reddit	Industry	Time FE	#Obs.	Adj.R^2	Reddit	Industry	Time FE	#Obs.	Adj.R^2
ROA	0.859	0.618	Yes	21983	0.03	1.028	0.610	Yes	20913	0.03
	(5.97)	(5.03)				(7.22)	(5.00)			
ROE	1.166	0.942	Yes	20490	0.01	1.672	0.826	Yes	19479	0.01
	(3.45)	(4.10)				(5.34)	(3.77)			
GP	0.496	1.058	Yes	20814	0.03	0.750	1.041	Yes	19798	0.04
	(5.08)	(7.12)				(9.61)	(6.94)			
EP	0.850	1.336	Yes	22046	0.04	0.785	1.326	Yes	20979	0.04
	(6.48)	(3.75)				(7.91)	(3.70)			
BM	1.797	2.180	Yes	20621	0.03	5.459	2.004	Yes	19603	0.04
	(1.94)	(1.42)				(5.32)	(1.31)			
CP	0.223	0.437	Yes	16056	0.02	0.406	0.440	Yes	15257	0.03
	(3.76)	(2.40)				(5.55)	(2.37)			
SP	1.763	2.240	Yes	21374	0.02	3.506	2.219	Yes	20365	0.02
	(2.84)	(1.80)				(8.48)	(1.76)			
SG	3.660	6.871	Yes	19929	0.03	4.571	6.946	Yes	18970	0.03
	(3.07)	(1.83)				(3.43)	(1.82)			
PG	0.043	0.586	Yes	19994	0.04	0.136	0.594	Yes	19006	0.04
	(1.61)	(5.05)				(2.99)	(5.14)			
RG	3.833	6.639	Yes	18820	0.03	4.659	6.691	Yes	17917	0.03
	(3.00)	(1.72)				(3.39)	(1.70)			
AG	3.603	8.188	Yes	20620	0.02	5.528	8.556	Yes	19600	0.02
	(2.99)	(3.13)				(2.04)	(3.15)			
LEV	6.755	3.624	Yes	20465	0.00	9.923	3.597	Yes	19447	0.00
	(2.41)	(0.82)				(3.71)	(0.79)			
AT	1.306	1.479	Yes	21983	0.02	1.584	1.461	Yes	20913	0.02
	(7.22)	(5.00)				(8.00)	(4.97)			
RDS	17.583	24.278	Yes	21372	0.00	33.837	22.898	Yes	20362	0.00
	(2.18)	(2.65)				(2.20)	(2.48)			

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 4. Fundamental comovement: cross-period relationships.
This table reports the intertemporal fundamental comovements between the focal firms and their Reddit peers:

$$F_{i,j,t+1} = \beta_1 F^R_{i,t} + \beta_2 F^I_{j,t} + \gamma_t + \varepsilon_{i,j,t}.$$

Each month, two stocks are defined as Reddit-linked if there was at least one thread (Panel A) or author (Panel B) on Reddit mentioning both stocks in the past six months. The independent variable ($F_{i,j,t+1}$) is the focal firms' one-quarter ahead fundamental, including return on assets (ROA), return on equity (ROE), gross profit (GP), earnings-to-price ratio (EP), book-to-market ratio (BM), cash flow-to-price ratio (CP), sales-to-price ratio (SP), sales growth (SG), profit growth (PG), revenue growth (RG), asset growth (AG), leverage ratio (LEV), asset turnover (AT), and R&D expense-to-sales ratio (RDS). The independent variable ($F^R_{i,t}$) is the corresponding fundamentals of Reddit-connected firms. We control for average fundamentals of industry peers ($F^I_{j,t}$) and include time-fixed effects (γ_t). We further control for focal firms' own contemporaneous fundamentals, whose estimated coefficients are not reported for brevity. All independent variables are winsorized at the 1% and the 99% levels in each cross-section and standardized to have zero mean and unit variance. Standard errors are clustered by both time and firm. t-statistics are reported in parentheses. The sample period is from Q4 2018 to Q4 2022.

	Panel A. Reddit peer: thread					Panel B. Reddit peer: author				
	Reddit	Industry	Time FE	#Obs.	Adj.R^2	Reddit	Industry	Time FE	#Obs.	Adj.R^2
ROA	0.212	0.218	Yes	18432	0.40	0.309	0.224	Yes	17848	0.39
	(2.94)	(4.38)				(6.03)	(4.39)			
ROE	0.271	0.265	Yes	17003	0.33	0.615	0.306	Yes	16456	0.33
	(2.20)	(1.84)				(3.69)	(2.06)			
GP	0.039	0.150	Yes	17531	0.75	0.089	0.149	Yes	16973	0.76
	(1.00)	(5.11)				(2.87)	(5.38)			
EP	0.343	0.573	Yes	18495	0.19	0.391	0.591	Yes	17915	0.19
	(3.58)	(2.74)				(5.58)	(2.77)			
BM	0.400	-0.221	Yes	17164	0.81	0.317	-0.280	Yes	16613	0.81
	(1.26)	(-1.01)				(0.93)	(-1.27)			
CP	0.022	-0.005	Yes	12377	0.47	0.091	-0.002	Yes	11988	0.47
	(0.70)	(-0.10)				(2.55)	(-0.05)			
SP	0.129	0.151	Yes	17929	0.79	-0.089	0.173	Yes	17388	0.79
	(0.78)	(0.68)				(-0.47)	(0.64)			
SG	1.828	-1.111	Yes	16652	0.21	2.103	-1.178	Yes	16138	0.21
	(3.70)	(-0.84)				(1.67)	(-0.83)			
PG	0.053	0.259	Yes	16669	0.19	0.056	0.262	Yes	16134	0.20
	(2.86)	(2.39)				(1.22)	(2.39)			
RG	1.636	-1.623	Yes	15809	0.21	2.015	-1.689	Yes	15325	0.21
	(3.09)	(-1.05)				(1.63)	(-1.02)			
AG	0.768	-0.697	Yes	17239	0.35	1.393	-0.762	Yes	16681	0.35
	(1.01)	(-0.54)				(1.68)	(-0.57)			
LEV	0.476	-0.740	Yes	16989	0.82	1.595	-0.854	Yes	16435	0.82
	(0.88)	(-0.62)				(1.80)	(-0.71)			
AT	0.051	0.109	Yes	18433	0.88	0.058	0.099	Yes	17849	0.88
	(0.85)	(2.08)				(0.96)	(1.91)			
RDS	-1.973	0.581	Yes	17924	0.59	-1.111	0.729	Yes	17383	0.58
	(-0.55)	(0.27)				(-0.19)	(0.34)			

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 5. One-sort portfolios by Reddit-peer returns.
This table reports the performance of portfolios based on the returns of Reddit peer firms. Reddit peers are identified by shared Reddit threads (Panel A) or shared Reddit authors (Panel B). Each day, stocks are sorted into quantile portfolios based on $REDDIT\ RET$, the weighted average one-day return of stocks that are linked through Reddit. Value-weighted and equal-weighted returns (in basis points) with a one-day holding period are calculated for each portfolio. The sample includes common stocks listed on NYSE, AMEX, and NASDAQ with a share price of at least $1 as of portfolio formations. The table reports average excess returns (Return) and alphas using the Fama-French five-factor model (FF5) and the momentum-augmented factor model (FF6). t-statistics with Newey-West adjustment are reported in parentheses. The sample period is from January 2019 to December 2022.

	Low	2	3	4	High	High-Low
Panel A. Reddit thread peers						
			Value-weighted			
Return	3.75	5.67	4.90	8.01	8.99	5.24
	(0.80)	(1.36)	(1.16)	(1.91)	(1.90)	(2.68)
FF5	-1.36	0.61	-0.18	2.64	3.76	5.12
	(-1.06)	(0.54)	(-0.15)	(2.53)	(2.69)	(2.58)
FF6	-1.35	0.64	-0.13	2.66	3.76	5.11
	(-1.05)	(0.56)	(-0.11)	(2.56)	(2.68)	(2.57)
			Equal-weighted			
Return	1.63	3.69	5.49	6.44	8.10	6.46
	(0.28)	(0.63)	(0.91)	(1.11)	(1.40)	(3.70)
FF5	-2.41	-0.43	1.20	2.42	4.12	6.53
	(-1.64)	(-0.32)	(0.90)	(1.72)	(2.47)	(3.71)
FF6	-2.56	-0.54	1.12	2.31	4.00	6.55
	(-1.74)	(-0.41)	(0.84)	(1.66)	(2.42)	(3.72)
Panel B. Reddit author peers						
			Value-weighted			
Return	2.69	5.28	4.92	8.25	7.91	5.23
	(0.56)	(1.26)	(1.08)	(1.84)	(1.78)	(2.56)
FF5	-2.25	0.21	-0.17	2.82	2.86	5.11
	(-1.67)	(0.17)	(-0.12)	(2.04)	(2.07)	(2.52)
FF6	-2.22	0.28	-0.11	2.87	2.88	5.09
	(-1.65)	(0.23)	(-0.09)	(2.11)	(2.08)	(2.52)
			Equal-weighted			
Return	0.31	3.95	5.30	7.16	9.22	8.91
	(0.05)	(0.66)	(0.89)	(1.22)	(1.60)	(4.65)
FF5	-3.46	-0.25	1.13	2.95	5.10	8.56
	(-2.26)	(-0.18)	(0.86)	(1.95)	(2.90)	(4.38)
FF6	-3.56	-0.34	1.03	2.84	4.96	8.52
	(-2.32)	(-0.24)	(0.78)	(1.91)	(2.88)	(4.39)

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 6. Momentum spillover among Reddit-linked firms.
This table reports the estimated regression coefficients predicting one-day ahead returns (in percent) using Reddit peer returns (REDDIT RET):

$$Ret_{i,d+1} = \beta_0 + \beta_1 REDDIT\ RET_{i,d} + \beta_2 Controls_{i,d} + \gamma_d + \varepsilon_{i,d+1}.$$

For each focal stock. the Reddit peer return is the weighted average one-day return of stocks that are linked through the Reddit thread (Panel A) or Reddit author (Panel B). Control variables include the focal stock's one-day return, one-month return (skipping the most recent day), cumulative 100-day return (skipping the most recent month), as well as illiquidity, idiosyncratic volatility, the log of market capitalization, and the log of the book-to-market ratio measured at the end of last month. We further control for peer returns from other linkage relationships such as industry (IND RET) and shared-analyst coverage (CF RET) of Ali and Hirshleifer (2020). Day-fixed effects are included in regressions, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2022.

	Panel A. Reddit thread peers				Panel B. Reddit author peers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
REDDIT RET	6.029	5.691	3.820	3.940	3.953	3.673	2.356	2.509
	(6.94)	(6.58)	(4.82)	(4.96)	(7.77)	(7.25)	(5.14)	(5.46)
IND RET		4.563	1.407	1.737		4.489	1.293	1.623
		(12.12)	(3.63)	(4.46)		(11.76)	(3.30)	(4.13)
CF RET			6.634	8.346			6.671	8.388
			(12.95)	(14.26)			(12.86)	(14.11)
Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	0.050	0.042	0.059	-0.011	0.073	0.065	0.062	0.007
	(1.50)	(1.27)	(21.41)	(-0.34)	(2.14)	(1.88)	(22.67)	(0.21)
Observations	1873506	1869505	1711117	1711117	1822362	1818346	1664769	1664769
Adjusted R^2	0.16	0.16	0.19	0.19	0.15	0.15	0.19	0.19

Electronic copy available at: https://ssrn.com/abstract=4536567

SMU SINGAPORE MANAGEMENT UNIVERSITY
Sim Kee Boon Institute for Financial Economics

Table 7. Tests of the limited attention hypothesis.

This table reports the results of regressions that explore the limited attention hypothesis of Reddit peers' momentum spillover effect. The dependent variable is the focal stock's one-day ahead return (in percent). The main dependent variables of interest are Reddit peer return (REDDIT RET) and its interaction terms. We consider interactions using the number of Reddit peers (Link), analyst coverage (Analyst), media coverage (Media), and institutional ownership (Inst). In regressions, we include dummy variables that are equal to 1 if the underlying interaction variable is above the cross-sectional median and 0 otherwise. Due to the availability of news stories data, the sample period for regressions with media coverage is from January 2019 to December 2021. Reddit peers are identified by shared Reddit threads in Panel A and shared Reddit authors in Panel B. In all specifications, we control for the interaction effect from firm size. Other control variables ($Controls$) are defined identically as in Table 6, except that we replace the focal firm's log market firm value with the dummy variable $Size\ High$. We include time-fixed effects, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2022.

	Panel A. Reddit thread peers				Panel B. Reddit author peers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
REDDIT RET	7.914	9.163	9.340	7.848	6.015	7.413	7.172	5.940
	(8.94)	(8.09)	(10.33)	(8.86)	(10.93)	(10.98)	(12.38)	(10.85)
REDDIT RET×Analyst High	-2.321				-1.685			
	(-4.45)				(-3.45)			
REDDIT RET×Media High		-2.279				-1.494		
		(-3.99)				(-2.81)		
REDDIT RET×Inst High			-5.157				-4.320	
			(-11.73)				(-10.27)	
REDDIT RET×Link High				1.480				2.304
				(3.30)				(5.15)
REDDIT RET×Size High	-5.953	-7.812	-5.493	-7.667	-5.490	-7.059	-4.885	-6.788
	(-11.05)	(-13.16)	(-12.52)	(-16.42)	(-11.00)	(-12.77)	(-11.72)	(-15.96)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	0.073	0.092	0.063	0.076	0.074	0.093	0.066	0.081
	(8.33)	(8.81)	(6.71)	(8.81)	(8.27)	(8.77)	(6.91)	(9.00)
Observations	1873506	1457668	1873506	1873506	1822362	1427373	1822362	1822362
Adjusted R^2	0.16	0.14	0.16	0.16	0.15	0.14	0.15	0.15

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 8. Returns around information days.
This table reports results from regressions of one-day ahead returns (in percent), on lagged Reddit peer returns (REDDIT RET), the earnings day and news day dummy variables, and interaction terms:

$$\begin{aligned} Ret_{i,d+1} =& \beta_0 + \beta_1 REDDIT\ RET_{i,d} + \beta_2 REDDIT\ RET_{i,d} \times Eday_{i,d+1} \\ &+ \beta_3 REDDIT\ RET_{i,d} \times Nday_{i,d+1} + \beta_4 Eday_{i,d+1} + \beta_5 Nday_{i,d+1} \\ &+ \delta Controls + \gamma_d + \varepsilon_{i,d+1}. \end{aligned}$$

The dependent variable is multiplied by 100. The control variables include the lagged values for each of the past ten days of stock returns, return squared, and trading volume. The Reddit peer return is the weighted average one-day return of stocks that are linked through Reddit thread (Panel A) or Reddit author (Panel B). Eday (Nday) is a dummy variable equal to one on earnings (news) days and zero otherwise. Following Engelberg et al. (2018), the earnings (news) day is defined as the day with the highest trading volume around the three-day window centered on an earnings announcement (news release). Regressions include day-fixed effects. Standard errors are clustered on the time and t-statistics are in parentheses. The sample period is from January 2019 to December 2021.

	Panel A. Reddit thread peers		Panel B. Reddit author peers	
	(1)	(2)	(3)	(4)
REDDIT RET	5.650	6.417	3.987	4.348
	(3.30)	(3.63)	(4.06)	(4.56)
REDDIT RET×Eday	11.027	11.306	8.323	8.524
	(2.19)	(2.24)	(1.94)	(1.98)
REDDIT RET×Nday	3.923	4.110	5.045	5.205
	(1.81)	(1.91)	(2.43)	(2.53)
Eday	0.231	0.230	0.257	0.256
	(2.99)	(2.98)	(3.33)	(3.31)
Nday	0.522	0.527	0.517	0.523
	(19.91)	(19.84)	(19.73)	(19.69)
Controls	No	Yes	No	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	1597737	1583899	1562182	1548712
Adjusted R^2	0.13	0.13	0.14	0.14

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 9. Reddit peer return and future earnings surprise.
This table reports regressions of the next quarter's standardized unexpected earnings (SUE_t) of Reddit peer's quarterly return (REDDIT $QRET_{t-1}$). Reddit peers are identified by shared Reddit thread in Panel A and shared Reddit author in Panel B. We include firm-fixed effect and time-fixed effect and control for the focal firm's own lagged SUEs. Independent variables are winsorized at the 1% and the 99% level in each cross-section and standardized to have zero mean and unit variance. t-statistics are reported using standard errors clustered on firm and quarter. The sample period is from December 2018 to December 2022.

	Panel A. Reddit post peers		Panel B. Reddit author peers	
	(1)	(2)	(3)	(4)
REDDIT $QRET_{t-1}$	0.028	0.028	0.033	0.041
	(2.20)	(2.58)	(2.79)	(3.29)
SUE_{t-1}		0.224		0.222
		(7.32)		(7.02)
SUE_{t-2}		0.119		0.120
		(5.37)		(5.25)
SUE_{t-3}		0.031		0.032
		(1.25)		(1.31)
SUE_{t-4}		-0.556		-0.558
		(-13.98)		(-14.03)
Firm FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	27216	25347	26391	24607
Adjusted R^2	0.14	0.26	0.14	0.27

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 10. Uncertainty and Reddit peer momentum.
This table reports predictive regressions on Reddit peer momentum on lagged VIX. The Reddit peer momentum is measured by the equal-weighted strategy return (basis points) as implemented in Table 5. Reddit peers are identified using shared Reddit threads or shared Reddit authors. VIX is standardized to have zero mean and unit variance. Control variables include daily factor returns of the market excess return, SMB, HML, CMA, RMW, and Momentum (Fama and French, 2015). *t*-statistics with Newey-West adjusted standard errors are reported in parentheses. The sample period is from January 2019 to December 2022.

	(1)	(2)	(3)	(4)
	Reddit thread		Reddit author	
Lagged VIX	4.788	5.033	8.347	8.484
	(3.07)	(3.27)	(2.32)	(2.35)
Controls	No	Yes	No	Yes
Intercept	6.529	6.628	8.906	8.538
	(3.82)	(3.87)	(4.94)	(4.67)
Observations	1006	1006	1006	1006

Electronic copy available at: https://ssrn.com/abstract=4536567

Table 11. Order imbalance, Robinhood ownership, and Reddit peer returns.
This table reports the estimated regression coefficients predicting one-day ahead order imbalance and changes in Robinhood ownership (RH) using Reddit peer returns. For each focal stock. the Reddit peer return is the weighted average one-day return of stocks that are linked through Reddit thread (Panel A) or Reddit author (Panel B). We consider three types of order imbalance: trade imbalance, volume imbalance, and dollar imbalance. The trade imbalance is calculated by the difference in the number of buys and number of sells divided by the total number of buys and sells; the volume imbalance is shares of buy trades minus shares of sell trades divided by the total volume of buys and sells; the dollar imbalance is the difference in the dollar value of buys and the dollar value of sells divided by the total dollar value of buys and sells. The change in Robinhood ownership of a stock is defined as the daily growth rate of Robinhood users holding the stock. Control variables are defined identically as in Table 6. Day-fixed effects are included in regressions, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2021 for order imbalance and from January 2019 to August 2020 for the Robinhood ownership.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Reddit thread peers							
	Total order imbalance			Retail order imbalance			RH
	Trade	Volume	Dollar	Trade	Volume	Dollar	
REDDIT RET	-0.065	-0.019	-0.020	0.003	0.115	0.113	0.009
	(-2.13)	(-0.50)	(-0.52)	(0.05)	(1.63)	(1.61)	(1.26)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1400590	1400590	1400590	1396818	1396818	1396818	570003
Adjusted R^2	0.02	0.02	0.02	0.01	0.00	0.00	0.05
Panel B. Reddit author peers							
	Total order imbalance			Retail order imbalance			RH
	Trade	Volume	Dollar	Trade	Volume	Dollar	
REDDIT RET	-0.022	-0.013	-0.013	0.036	0.056	0.056	-0.002
	(-1.01)	(-0.55)	(-0.55)	(1.46)	(1.55)	(1.54)	(-0.38)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1370920	1370920	1370920	1367114	1367114	1367114	550601
Adjusted R^2	0.02	0.02	0.02	0.01	0.00	0.00	0.05

Electronic copy available at: https://ssrn.com/abstract=4536567

Appendix to "Network through Social Media Connections"

A Matching between stocks and texts

We discuss in detail the method to match a comment in *r/wallstreetbets* to a stock here. In this paper, if the tokenized text[1] matched our list of stock tickers[2], we would label this comment as a stock-related comment and identify any related stocks. For example, the sentence "INTC still has dominant market share in both consumer and server markets", shown in Figure 3 is tokenized in our method to a list of 12 tokens ['INTC', 'still', 'has', 'dominant', 'market', 'share', 'in', 'both', 'consumer', 'and', 'server', 'markets'], then we cross-reference all these tokens of this list with the ticker list and find "INTC" meeting our requirements. So, we label the comment containing this sentence as stock-related and tag this comment as "Intel Corp." related.

Naturally, this matching method may raise several concerns. First, removing tickers with fewer than 3 letters may result in the omission of many stock-related comments, referred to as false negative matches. Second, some matches may include comments not discussing stocks—the false positive matches. We discuss our opinions about these two concerns respectively. The deletion of tickers makes the sample not contain stocks with less than three letters, and these ticker symbols account for about 3% of the number of stocks in the list. But the inclusion of these tickers with one or two letters will cause much confusion because of the acronym overload and add excessive noise to the construction of our linkage. However, the criterion of at least three letters cannot solve semantic confusion problems completely. These ticker symbols could be used by certain Reddit users to represent something other than stock. For example, they can use "CAKE" to refer to the cake they are now eating today without talking about The Cheesecake Factory Inc., which uses the ticker symbol "CAKE". To reduce the noise that this confusion causes, we use another version of the ticker symbol list as a robustness check, deleting all the 5000 most frequent words according to the Google Web Trillion Word Corpus[3], and we get similar empirical results.

An alternative matching method is to use "$" before the ticker symbol, which means we tag the comment as "Intel Corp." related when "$INTC" emerges in it instead of

[1]Tokenization is the process of substituting text into words, phrases, symbols, or other non-sensitive equivalents, known as tokens.

[2]The list contains stocks in NYSE/AMEX/NASDAQ and with share code 10 or 11. And we keep tickers with at least 3 letters to avoid ambiguities in the matching process.

[3]The word counts data file is obtained from `http://norvig.com/ngrams/`.

Electronic copy available at: https://ssrn.com/abstract=4536567

when "INTC" does. The majority of the semantic misunderstanding might be avoided by using this tighter approach, which is also widely used in literature, like in Hu et al. (2023) and in Li (2022). The main reason we refuse this method is that it may result in a significant lack of comments about stocks and may affect the representativeness of our Reddit linkages. Reddit users are not forced to use "$TICKER", with a dollar sign before the ticker, to refer to the stock, and they also usually use the ticker symbol without dollar signs while discussing the fundamental information of the firm.

B Description of fundamental variables

We use CRSP monthly stock returns and COMPUSTAT quarterly data for the sample period December 2018-December 2022. We drop firm-quarter observations with missing or negative total assets (item atq). We follow Lee et al. (2015) and Hou et al. (2020) in constructing the fundamental variables.

ROA: return on assets. Quarterly net income before extraordinary items (item ibq) scaled by 1-quarter-lagged total assets (item atq).

ROE: return on equity. Quarterly net income before extraordinary items (item ibq) scaled by 1-quarter-lagged book equity. Book equity is shareholders' equity, plus balance sheet deferred taxes and investment tax credit (item txditcq) if available, minus the book value of preferred stock (item pstkq). Depending on availability, we use stockholders' equity (item seqq), or common equity (item ceqq) plus the book value of preferred stock, or total assets (item atq) minus total liabilities (item ltq) in that order as shareholders' equity.

GP: gross profit. Total revenue (item revtq) minus cost of goods sold (item cogsq) divided by 1-quarter-lagged total assets (item atq).

EP: earnings-to-price ratio. Quarterly net income before extraordinary items (item ibq) divided by market equity.

BM: book-to-market ratio. Book equity divided by market equity. Book equity is shareholders' equity, plus balance sheet deferred taxes and investment tax credit (txditcq) if available, minus the book value of preferred stock (pstkq). Depending on availability, we use stockholders' equity (seqq), or common equity (ceqq) plus the book value of preferred stock, or total assets (atq) minus total liabilities (ltq) in that order as shareholders' equity.

Electronic copy available at: https://ssrn.com/abstract=4536567

CP: cash flow-to-price ratio. Quarterly cash flows divided by the market equity. Quarterly cash flows are income before extraordinary items (item ibq) plus depreciation (item dpq).

SP: sales-to-price ratio. Quarterly sales (item saleq) divided by the market equity.

SG: sales growth. Quarterly sales (item saleq) divided by the four-quarter-lagged value, and then minus one. We require both the current sales and the lagged sales to be positive in order to calculate the growth rate.

PG: profit growth. Quarterly operating income before depreciation (item oibdpq) minus its value four quarters ago divided by 1-quarter-lagged total assets (item atq).

RG: revenue growth. Quarterly total revenue (item revtq) divided by its value four quarters ago, then minus one.

AG: asset growth. Total assets (item atq) divided by its value four quarters ago and then minus one.

LEV: leverage ratio. Long-term debt (item dlttq) scaled by total shareholder' equity (item seqq).

AT: asset turnover. Quarterly sales (item saleq) divided by total assets (item atq).

RDS: R&D expense-to-sales ratio. Quarterly R&D expense (item xrdq) divided by quarterly sales (item saleq). Firms with nonpositive sales are excluded. If a firm's R&D expense is missing, we assume the value to be zero.

C Additional robustness checks of the main results

In our main regressions that examine the lead-lag effect of Reddit-linked firms, we follow prior literature to include day-fixed effects and cluster standard errors at the firm level. Table A1 presents market capitalization-weighted regression results and Table A2 presents regressions using two-way clustered standard errors (by firm and day) and including firm-fixed effects. We find that the magnitude of the estimated coefficient on REDDIT RET remains qualitatively similar. Although the t-statistic decreases in general, the result suggests that the predictive ability of REDDIT RET is still significant.

In Table A3 and Table A4, we further control for alternative inter-firm linkages, including text-based peers (Hoberg and Phillips, 2016, 2018), technological links (Lee et al., 2019), geographic links (Parsons et al., 2020), the customer-supplier relationship

Electronic copy available at: https://ssrn.com/abstract=4536567

(Cohen and Frazzini, 2008), and conglomerate firms (Cohen and Lou, 2012). The Reddit lead-lag effect remains evident under additional controls.

Electronic copy available at: https://ssrn.com/abstract=4536567

Table A1. Value-weighted regressions.

This table reports the estimated regression coefficients predicting one-day ahead returns (in percent) using Reddit peer returns (REDDIT RET). In each regression, stocks are weighted by market capitalization. Control variables include the focal stock's one-day return, one-month return (skipping the most recent day), cumulative 100-day return (skipping the most recent month), as well as illiquidity, idiosyncratic volatility, log of market capitalization, and log of book-to-market ratio measured at the end of last month. We further control for peer returns from other linkage relationships such as industry (IND RET) and shared-analyst coverage (CF RET) of Ali and Hirshleifer (2020). Day-fixed effects are included in regressions, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2022.

	Panel A. Reddit thread peers				Panel B. Reddit author peers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
REDDIT RET	6.009	6.199	6.145	6.143	2.570	2.607	2.493	2.497
	(4.20)	(4.37)	(4.18)	(4.21)	(4.44)	(4.24)	(3.91)	(4.02)
IND RET		-0.519	-1.064	-1.022		-0.556	-1.092	-1.039
		(-0.75)	(-1.31)	(-1.25)		(-0.79)	(-1.33)	(-1.26)
CF RET			1.474	1.567			1.388	1.520
			(1.64)	(1.70)			(1.54)	(1.64)
Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	0.085	0.084	0.045	0.084	0.089	0.088	0.047	0.088
	(1.69)	(1.68)	(8.21)	(1.65)	(1.74)	(1.74)	(8.28)	(1.69)
Observations	1873473	1869478	1711090	1711090	1822332	1818322	1664745	1664745
Adjusted R^2	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35

Electronic copy available at: https://ssrn.com/abstract=4536567

Table A2. Cluster and fixed effects in regressions.
This table reports the estimated regression coefficients predicting one-day ahead returns (in percent) using Reddit peer returns (REDDIT RET). For each focal stock. the Reddit peer return is the weighted average one-day return of stocks that are linked through Reddit thread or Reddit author. Control variables are the same as used in Table 6 of the paper. IND RET is the value-weighted industry return and CF RET is the shared-analyst coverage peer firm return (Ali and Hirshleifer, 2020). In Panel A, standard errors are clustered at the firm and day levels. In Panel B, we further include firm-fixed effects in regressions. t-statistics are reported in parentheses. The sample period is from January 2019 to December 2022.

Panel A. Two-way cluster						
	(1)	(2)	(3)	(4)	(5)	(6)
	Reddit thread peers			Reddit author peers		
REDDIT RET	6.029	5.691	3.940	3.953	3.673	2.509
	(4.60)	(4.39)	(3.42)	(5.30)	(5.07)	(3.81)
IND RET		4.563	1.737		4.489	1.623
		(3.62)	(1.78)		(3.53)	(1.63)
CF RET			8.346			8.388
			(4.53)			(4.55)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	No	No
Intercept	0.050	0.042	-0.011	0.073	0.065	0.007
	(0.64)	(0.55)	(-0.13)	(0.94)	(0.83)	(0.08)
Observations	1873506	1869505	1711117	1822362	1818346	1664769
Adjusted R^2	0.16	0.16	0.19	0.15	0.15	0.19
Panel B. Firm-fixed effects included						
	(1)	(2)	(3)	(4)	(5)	(6)
	Reddit thread peers			Reddit author peers		
REDDIT RET	6.052	5.698	3.932	3.995	3.707	2.524
	(4.61)	(4.39)	(3.40)	(5.38)	(5.14)	(3.84)
IND RET		4.730	1.887		4.668	1.786
		(3.77)	(1.94)		(3.69)	(1.79)
CF RET			8.366			8.401
			(4.55)			(4.56)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	5.507	5.504	5.375	5.868	5.878	5.715
	(9.95)	(9.96)	(9.02)	(10.36)	(10.39)	(9.42)
Observations	1873505	1869504	1711117	1822361	1818345	1664769
Adjusted R^2	0.16	0.16	0.19	0.16	0.16	0.19

Electronic copy available at: https://ssrn.com/abstract=4536567

Table A3. Control for alternative spillover effects: Reddit thread links.
This table reports the estimated regression coefficients predicting one-day ahead returns (in percent) using shared Reddit thread-based peer returns (REDDIT RET). Control variables include the focal stock's one-day return, one-month return (skipping the most recent day), cumulative 100-day return (skipping the most recent month), as well as illiquidity, idiosyncratic volatility, log of market capitalization, and log of book-to-market ratio measured at the end of last month. We further control for the momentum spillover effect from other linkage relationships, including the text-based peers $TNIC$ (Hoberg and Phillips, 2016, 2018), technological links $TECH$ (Lee et al., 2019), geographic links GEO (Parsons et al., 2020), major customers CUS (Cohen and Frazzini, 2008), conglomerate firms PC (Cohen and Lou, 2012), industry peers IND, and shared-analyst coverage CF (Ali and Hirshleifer, 2020). Day-fixed effects are included in regressions, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2022. Due to the data availability of constructing return signals, the sample period for regressions with $TNIC$ is from January 2019 to June 2021, and for regressions with $TECH$ is from January 2019 to June 2022.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
REDDIT RET	5.355	4.879	6.130	6.439	5.581	3.904	3.442	4.147	5.467	3.558
	(4.17)	(2.96)	(6.85)	(3.37)	(3.68)	(3.54)	(2.24)	(5.12)	(2.78)	(2.28)
TNIC RET	5.460					2.588				
	(11.63)					(4.91)				
TECH RET		7.365					-0.092			
		(4.24)					(-0.05)			
GEO RET			2.411					0.850		
			(6.98)					(2.72)		
CUS RET				3.681					3.081	
				(4.56)					(4.22)	
PC RET					2.856					1.258
					(4.07)					(1.90)
IND RET						0.323	0.500	1.577	1.458	0.826
						(0.57)	(0.66)	(3.93)	(1.55)	(1.17)
CF RET						8.258	7.194	8.210	4.788	6.335
						(9.16)	(6.27)	(13.75)	(3.64)	(4.90)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	0.343	0.080	0.049	0.169	0.112	0.276	0.015	-0.010	0.062	0.017
	(6.52)	(1.03)	(1.40)	(1.97)	(1.94)	(5.91)	(0.24)	(-0.29)	(0.86)	(0.32)
Observations	1029324	542471	1775670	321581	488250	949769	516925	1630969	303256	446564
Adjusted R^2	0.15	0.16	0.16	0.17	0.20	0.18	0.18	0.19	0.20	0.29

Electronic copy available at: https://ssrn.com/abstract=4536567

Table A4. Control for alternative spillover effects: Reddit author links.
This table reports the estimated regression coefficients predicting one-day ahead returns (in percent) using shared Reddit author-based peer returns (REDDIT RET). Control variables include the focal stock's one-day return, one-month return (skipping the most recent day), cumulative 100-day return (skipping the most recent month), as well as illiquidity, idiosyncratic volatility, log of market capitalization, and log of book-to-market ratio measured at the end of last month. We further control for the momentum spillover effect from other linkage relationships, including the text-based peers $TNIC$ (Hoberg and Phillips, 2016, 2018), technological links $TECH$ (Lee et al., 2019), geographic links GEO (Parsons et al., 2020), major customers CUS (Cohen and Frazzini, 2008), conglomerate firms PC (Cohen and Lou, 2012), industry peers IND, and shared-analyst coverage CF (Ali and Hirshleifer, 2020). Day-fixed effects are included in regressions, and t-statistics with standard errors clustered at the firm level are reported in parentheses. The sample period is from January 2019 to December 2022. Due to the data availability of constructing return signals, the sample period for regressions with $TNIC$ is from January 2019 to June 2021, and for regressions with $TECH$ is from January 2019 to June 2022.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
REDDIT RET	4.621	2.883	3.813	3.221	3.394	3.579	1.649	2.393	3.019	2.298
	(6.57)	(2.90)	(7.23)	(2.81)	(3.03)	(5.41)	(1.91)	(5.06)	(3.03)	(3.00)
TNIC RET	5.301					2.435				
	(11.13)					(4.57)				
TECH RET		7.073					-0.415			
		(4.02)					(-0.21)			
GEO RET			2.351					0.861		
			(6.75)					(2.74)		
CUS RET				3.663					3.061	
				(4.42)					(4.07)	
PC RET					2.815					1.270
					(3.99)					(1.91)
IND RET						0.289	0.347	1.455	1.685	0.794
						(0.50)	(0.46)	(3.60)	(1.77)	(1.11)
CF RET						8.163	7.330	8.287	4.524	6.333
						(8.94)	(6.44)	(13.72)	(3.35)	(4.84)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	0.351	0.086	0.067	0.198	0.115	0.283	0.028	0.004	0.104	0.005
	(6.55)	(1.08)	(1.91)	(2.23)	(1.93)	(5.86)	(0.44)	(0.11)	(1.43)	(0.09)
Observations	1000847	530642	1727862	311760	476366	925056	505390	1586793	293853	436377
Adjusted R^2	0.15	0.16	0.16	0.17	0.20	0.18	0.18	0.19	0.20	0.29

Electronic copy available at: https://ssrn.com/abstract=4536567