Reddit's self-organised bull runs: Social contagion and asset prices

Semenova, Valentina and Winkler, Julian

Institute of New Economic Thinking at the Oxford Martin School, Department of Mathematics, University of Oxford, Department of Economics, University of Oxford

5 May 2021

Reddit's self-organised bull runs: Social contagion and asset prices*

Valentina Semenova†[1,3] and Julian Winkler‡[2,3]

[1]*Department of Mathematics, University of Oxford*
[2]*Department of Economics, University of Oxford*
[3]*Institute of New Economic Thinking at the Oxford Martin School*

February 6, 2022

Abstract

This paper develops an empirical and theoretical case for how 'hype' among retail investors can drive large asset fluctuations. We use text data from discussions on WallStreetBets (WSB), an online investor forum with over eleven million followers as of February 2022, to demonstrate how the adoption of trading strategies spreads among retail investors. Using sentiment analysis, we document that WSB users adopt price predictions about assets (bullish or bearish) in part due to the sentiments expressed by their peers. We, furthermore, document evidence that asset discussions on WSB are self-perpetuating: an initial set of investors in a stock attracts a growing number of followers – a pattern reminiscent of an epidemiological setting. Leveraging these findings, we develop a theoretical case for the impact of social dynamics among retail investors on asset prices. Our framework helps identify components of asset demand stemming from social dynamics, which we predict using WSB data. Our predictions explain significant variation in stock market activity; a consequential result is that weekly changes in sentiments and stock-specific discussions explain 16% of variance in the percent change of nominal trading volumes, within stocks. These findings emphasise the role that social dynamics play in financial markets, amplified by online social media.

JEL codes: D91, G14, G41.

---

*[*] **Acknowledgements**: We are grateful to participants of the Royal Economic Society annual conference, the Alan Turing Institute interest group on economic data science, and the Rebuilding Macroeconomics seminar. We thank Rick Van der Ploeg, J. Doyne Farmer, Xiaowen Dong, Steve Bond, Francis DiTraglia, Kevin Sheppard, Marteen Scholl, John Pougué-Biyong, Ilan Strauss, Jangho Yang, Torsten Heinrich, François Lafond, Matthias Winkler, José Moran, Farshad Ravasan, Pedro Bordalo, Cars Hommes, Stefan Zohren, Mirta Galesic, and Renaud Lambiotte for their helpful comments and insightful questions. We thank Baillie Gifford and the Institute for New Economic Thinking at the Oxford Martin School for funding our work at the University of Oxford.

[†] valentina.semenova@maths.ox.ac.uk

[‡] julian.winkler@economics.ox.ac.uk

1 Introduction

In investigating the stock market crash of May 28, 1962, the Securities and Exchange Commission (SEC) found that: 'investor "psychology" being what it is, the increasing decline in one or several issues can easily spread to others. Once the process becomes generally operative, the stage is set for a serious market break' (1963). The SEC concluded that large institutions acted as a balancing force during the collapse. The report pointed at retail traders as the *key players* behind the widespread panic.

Over half a century later, we are again confronted with the consequences of investors' social behaviour. When online discussants on Reddit's 'WallStreetBets' (WSB) forum drove up the price of GameStop shares in January, 2021, retail investors regained a spotlight on the (virtual) trading floor. A key difference between 1962 and today is the internet, which offers both a coordination platform with an unprecedented following, and a new datasource on investor narratives, interactions and psychology.

This paper sets out to reconcile the behaviours on social media with economic theory. To the extent that humans are social beings, interactions are likely to play a role in financial decision-making – as famously stipulated by Shiller (1984), giving way to a rich literature on information and strategic choice (Barlevy & Veronesi 2000, Hellwig & Veldkamp 2009, Banerjee 1993, Zenou 2016). We propose a model where agents update their sentiments about assets based on a combination of public/private information, as well as the information signals revealed by others. We consider the implications of our model and test them using text data from the WSB forum.

We find that the sentiments expressed by an investor's peers about an asset impact the future sentiment adopted by an investor (net of a set of asset, investor and stock market controls). We test for direct peer effects in investor decision-making in two ways. We evaluate the effect of peers on people who share their outlook about an asset multiple times, and interact with peers in-between. This approach allows us to precisely identify the *change* in individual sentiment due to peer interactions. We also pursue a network approach – leveraging interactions on WSB to identify a peer cohort. Both methods use an Instrumental Variable (IV) approach to differentiate *peer effects* from *contextual effects* and *correlated effects*, as well as to tackle the *common shock* problem (Zenou 2016).

We find that individuals appear to weight the opinions of others almost as heavily as their own prior opinion about an asset – on average, investors on WSB appear to assign a weight of 0.16 to their own prior opinion about an asset and of 0.11 to the sentiments expressed by peers, when updating their outlook about an asset. Asset returns also play a role in opinion formation – an increase in an asset's share price raises investor sentiment. Even though social contagion and the role that narratives play in investor decision-making have been heavily discussed in the literature, to the best of our knowledge, this is the first work documenting a causal relationship between an investor's sentiment and that of his peers, outside a controlled experimental setup, such as Bursztyn et al. (2014).

Our second finding is that asset interest also appears to be contagious among peers. Individuals become interested in a stock not only because their peers are discussing it, but also because the peers have a track record of correctly predicting stock returns. We follow the framework of Banerjee (1993) and Shiller (1984) in our analysis, and observe that this mechanism for information transmission fits the data well, explaining almost 40 percent of the variance in the sum of squares for the log-odds of an asset discussed on WSB in a given week.

We tie together our findings about peer effects, or 'consensus' formation among investors, and interest spread, termed 'contagion', into a single framework to understand the role of social dynamics in financial markets. Following the approach of Black (1986), Shefrin & Statman (1994), De Long et al. (1990), we divide the market into 'hype' investors, susceptible to these social dynamics, and other traders who are not. A key takeaway is that the strength of social influence between investors plays an important role in determining hype investor market impact. Our model also highlights the problem of reverse causality: the impact of social dynamics depends on hype investors' response to changes in asset prices. Our main result centers on an empirical test, identifying variation in consensus and contagion unrelated to current price changes. This identification strategy works well, because of the strong temporal persistence of sentiments and dedication to specific assets, due to peer effects and information transmission. Our estimates are significant in predicting changes in weekly average log returns, as well as changes in volatility and trading volumes. These results provide evidence for a causal relationship between social dynamics, proxied by WSB conversations, and financial markets.

Economists have long deliberated to what extent social dynamics and human psychology play a role in economic decision-making (Shiller 1984, Black 1986), with Hirshleifer (2001) concluding that 'despite many empirical studies, scholarly viewpoints on the rationality of asset pricing have not converged.' Researchers have collected evidence of behaviours that undermine rational expectations theory and the efficient markets hypothesis, such as risk aversion (Kahneman & Tversky 1972, 1973, 1979), the perseverance of formed impressions (Ross et al. 1975, Anderson et al. 1980) and, more recently, peer effects (Bursztyn et al. 2014, Lahno & Serra-Garcia 2015, Banerjee et al. 2013), to name a few. However, many questions still remain unanswered. Do these behaviours persist among investors evaluating investment opportunities in the 'wild?' If so, what are the impacts on the trading markets? Should we be concerned about online forums turbo-charging 'irrational exuberance'? We shed light on these questions, and hope to demonstrate concurrently the usefulness of online text data to understand social dynamics behind economic decisions.

The importance of peers and narratives in forming investor perspectives was first considered in the seminal work of Robert Shiller (Shiller 1984), which discusses how social forces play a role in investor decision-making, as well as provides statistical evidence of the greater volatility of stock prices than warranted by that of dividends. Since then, 'narrative economics' has played an increasingly important role in our understanding of market price moves and investor decision-making (Shiller 2005, 2014, 2017, Banerjee et al. 2013). Economists have proposed impactful models for understanding how investors influence each other, with many studies focusing on information transmission in financial markets, such as Grossman & Stiglitz (1980), Barlevy & Veronesi (2000), Hellwig & Veldkamp (2009), Banerjee (1993), Cont & Bouchaud (2000). Simultaneously, psychologically-founded decision models have been developed to explain deviations from rational expectations and expected utility theory, founded in 'prospect theory' (Kahneman & Tversky 1972, 1973, 1979, Gennaioli & Shleifer 2010, 2018, Bordalo et al. 2019). Despite many important findings stemming from these areas of research, practical difficulties and lack of data have restricted many scientists to controlled laboratory experiments or theoretical research, whose external validity remains unchecked (Barberis 2013). By leveraging new data, our work provides fresh empirical evidence of how heuristics, and peer effects in particular, affect investor decision making.

There has been some carry-over from the peer effects literature to work on investor behaviour, in an attempt to understand how investors behave in the 'real world.' Bursztyn et al. (2014) perform a field experiment with a financial brokerage in Brazil, where they study investment decisions made by peer pairs: the peers are offered a 'high stakes' investment opportunity (minimum investments were R$2,000 – around 50% of the median investor's monthly income) in a certain order to identify the effects of 'social learning' and 'social utility' in financial decision-making. Other related work investigates the diffusion of micro-finance decisions in a social network (Banerjee et al. 2013) and the effect of social networks on saving (Breza & Chandrasekhar 2019). We distinguish ourselves from this and related works, such as Lahno & Serra-Garcia (2015) who study the effect of peers on risk taking in a controlled experiment, by observing a broader set of investors whose behaviours are unconstrained. Our methods focus on exploiting the naturally occurring variation in investor groups to identify peer effects, rather than on the controlled experimental setting. We also tackle the important question of evaluating the impact of social dynamics on financial markets.

Several impactful studies in the peer effects literature leverage naturally occurring variation in peers for their identification strategy. An area which pioneered many of these techniques investigates peer effects in the classroom (see Epple & Romano (2011), Sacerdote (2011) for a general overview, and Duflo et al. (2011) for a prominent example). Social networks are also an active area of study (see Bramoullé et al. (2020) for a recent review). Finally, techniques for correctly identifying peer impacts have shaped recent developments in this area of research (Angrist 2014, Blume et al. 2011). The present paper highlights how to transfer well-established techniques from this literature to new social media data, improving our understanding of investor psychology.

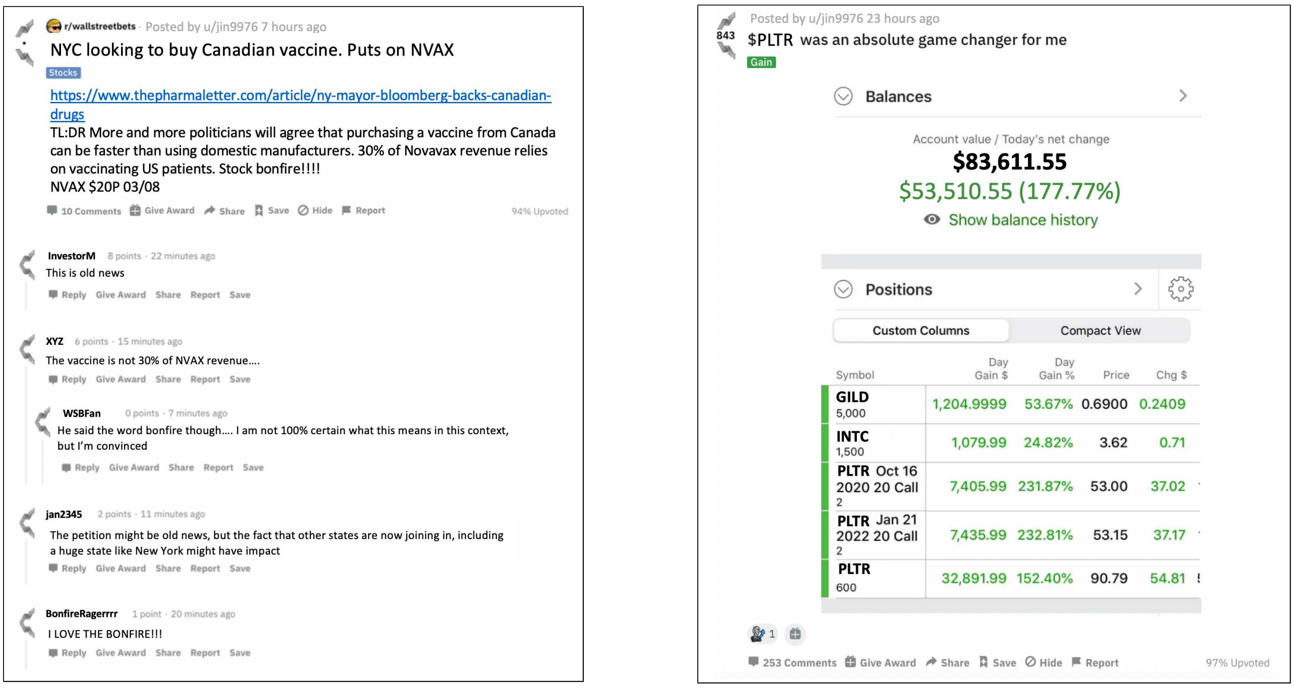
Several other works study the interplay between online forums and financial markets, as well as the spread of information in social networks. This paper differs from studies focusing on the spread of information through friend networks, such as Aral et al. (2009), Aral & Nicolaides (2017), since interactions take place in an anonymous forum, without any explicitly defined friendship links. The anonymity of Reddit is crucial to the prominence of WSB: in contrast to the exercise in Banerjee et al. (2013), where information is transmitted via friendship networks, the mechanism by which information dissipates on WSB is much closer to the homogenous mixing conditions popular in traditional epidemiological models, and therefore closer in spirit to Banerjee (1993). Our work is distinct from Bailey et al. (2016), Sabherwal et al. (2011), Bollen et al. (2011), Kumar & Lee (2006), which focus on identifying a direct relationship between social activity and assets. In contrast, we specify two, distinct channels, by which investors decide strategies based on their peers, plus what they observe in financial markets, and, subsequently, impact financial markets.

We present our results in four sections. The first comprehensively describes the data source and relevant variables. Section 3 presents empirical evidence of investor social dynamics, including the identification of peer effects. Section 4 outlines a simple model of how the observed social dynamics impact asset prices. It also contributes empirical evidence underscoring the effect of retail investors on the markets. Section 5 concludes.

2 What is *WallStreetBets*?

Reddit, launched in 2005, is a social news aggregation, web content rating, and discussion website. It was ranked as the 19th most visited site globally in April 2021,[1] with over 430 million anonymous users by the end of 2019.[2] The website's contents are self-organized by subject into smaller subforums, 'subreddits', to discuss a unique, central topic.

Within subreddits, users publish titled posts (also called 'submissions'), typically accompanied with a body of text or a link to an external website. These submissions can be commented and 'upvoted' or 'downvoted' by other users. A ranking algorithm raises the visibility of a submission with the amount of upvotes it receives, but lowers it with age. Therefore, the first posts that visitors see are i) highly upvoted, and ii) recent. Comments on a post, visible to anyone, are subject to a



NYC looking to buy Canadian vaccine. Puts on NVAX

https://www.thepharmaletter.com/article/ny-mayor-bloomberg-backs-canadian-drugs

TL:DR More and more politicians will agree that purchasing a vaccine from Canada can be faster than using domestic manufacturers. 30% of Novavax revenue relies on vaccinating US patients. Stock bonfire!!!!

NVAX $20P 03/08

InvestorM 8 points · 22 minutes ago

This is old news

XYZ 6 points · 15 minutes ago

The vaccine is not 30% of NVAX revenue….

WSBFan 0 points · 7 minutes ago

He said the word bonfire though…. I am not 100% certain what this means in this context, but I'm convinced

jan2345 2 points · 11 minutes ago

The petition might be old news, but the fact that other states are now joining in, including a huge state like New York might have impact

BonfireRagerrrr 1 point · 20 minutes ago

I LOVE THE BONFIRE!!!

$PLTR was an absolute game changer for me

Account value / Today's net change

$83,611.55

$53,510.55 (177.77%)

Symbol	Day Gain $	Day Gain %	Price	Chg $
GILD 5,000	1,204.9999	53.67%	0.6900	0.2409
INTC 1,500	1,079.99	24.82%	3.62	0.71
PLTR Oct 16 2020 20 Call 2	7,405.99	231.87%	53.00	37.02
PLTR Jan 21 2022 20 Call 2	7,435.99	232.81%	53.15	37.17
PLTR 600	32,891.99	152.40%	90.79	54.81

(a) A Typical Discussion on WSB

(b) A Sample Screenshot of User Profits

Figure 1: **What does WSB look like?** These snapshots display typical discussions on WSB. The exact text, usernames, and conversation details have been modified to protect user identities.

similar scoring system, and can, themselves, be commented on.

The WSB subreddit was created on January 31, 2012, and reached one million followers in March 2020.[3] As per a Google survey from 2016, the majority of WSB users are 'young, male, students that are inexperienced investors utilizing real money (not paper trading); most users have four figures in their trading account'.[4] The conversation guidelines outlined by the moderators of WSB handily demonstrate the financial focus and whimsical tone of discussions:[5]

- Discussion about day trading, stocks, options, futures, and anything market related,
- Charts and technical analysis,
- Shower before posting,
- Some irresponsible risk taking,
- People sharing trades, ideas, observations.

Figure 1a displays a typical exchange on the WSB forum: individuals discuss stock-related news and their sentiments on whether this will affect stock prices in the future. In addition to market discussions, there is ample evidence of users pursuing the investment strategies encouraged in WSB conversations. Users post screenshots of their investment gains and losses, which subreddit moderators are encouraged to verify, as illustrated in Figure 1b. These observations are reminiscent of Shiller (2005) in his definition of an asset bubble (our emphasis):

> A situation in which news of price increases spurs investor enthusiasm which spreads by psychological contagion from person to person, in the process amplifying stories that might justify the price increases and bringing in a larger and larger class of investors, who, despite doubts about the real value of an investment, are drawn to it partly through envy of others' successes and partly through a *gambler's excitement*.

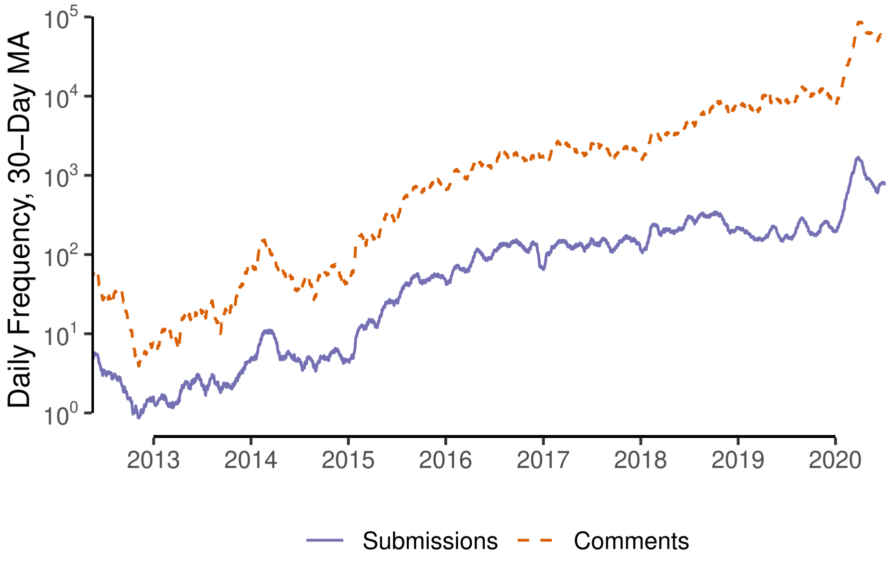


Figure 2: **Daily activity on WSB plotted on a logarithmic scale**; the daily submission and comment counts, averaged over 30 days, demonstrate a persistent exponential increase in activity on the WSB forum from 2015 to 2020, with a substantial jump in early 2020.

All posts made on Reddit, plus their metadata, can be queried via Reddit's API, as well as other sources. For this paper, we downloaded the entire history of WSB data using the PushShift API[6]. The only caveat of PushShift is that submissions are recorded at the time of their creation.

The full dataset consists of two parts. The first is a total of 426,840 submissions, with their authors, titles, text and timestamps. The second is comprised of 16.5 million comments, with their authors, text, timestamp, and the identifier of the comment/submission they reply to. Figure 2 displays the evolution of WSB over time; submission and comment activity has grown exponentially since 2015. Two jumps are notable: a smaller, seemingly idiosyncratic rise in early 2018, and a sharp spike during the COVID-19 pandemic.

Our dataset spans January, 2012 to July, 2020. Importantly, it does not include the events of the 2021 GameStop (GME) short squeeze. The decision to focus on this timeframe is intentional: before the GME short squeeze and widespread popularity of the forum, WSB received less attention from institutional investors, and less bot-activity. As such, our sample tracks retail investor discussions more precisely, without outside influence or manipulation. Furthermore, ample research has emerged focusing exclusively on the GME short squeeze, such as Chohan (2021), Vasileiou et al. (2021), whereas our goal is to characterize investor behaviour, rather than decompose a single event.

The following sections predominantly rely on submissions for text data, since they are substantially richer. Comments are used to trace user activity and, subsequently, the interactions between discussants. In order to understand how users discuss specific assets, we extract mentions of 'tickers' from the WSB submissions' text data. A ticker is a short combination of capital letters, used to identify an asset in the financial markets. For example, 'AAPL' refers to shares in Apple, Inc. Appendix A.1 documents how tickers are extracted from submissions. Table 5 in Appendix A.1 displays the twenty tickers that feature most prominently in WSB conversations. These are typically shares in technology firms, such as AMD or FB. A handful of indices are also present, notably the S&P 500 (SPY) and a gold ETF (JNUG). A small fraction of the 4,650 tickers we extract dominate the discourse on WSB. 90% of tickers are mentioned fewer than 31 times, and more than 60% are mentioned fewer than five times. Appendix A.1 considers the heavy-tailed nature of ticker discussions.

In addition to extracting tickers from submissions, we gauge whether discussions on certain assets express an expectation for their future price to rise, the 'bullish' case, to fall, the 'bearish' case, or to remain unpredictable/stable, the 'neutral' case. Among other alternatives, we identify the sentiment expressed about an asset within a WSB submission using a supervised-learning approach, with a hand-labeled dataset of almost five thousand submissions used for training, validation and testing. The sentiment model outputs a probability for each sentiment category (-1,0,+1), and achieves a 75% labeling accuracy on the hand-labeled test set. Refer to Appendix A.2 for further details on the sentiment model.

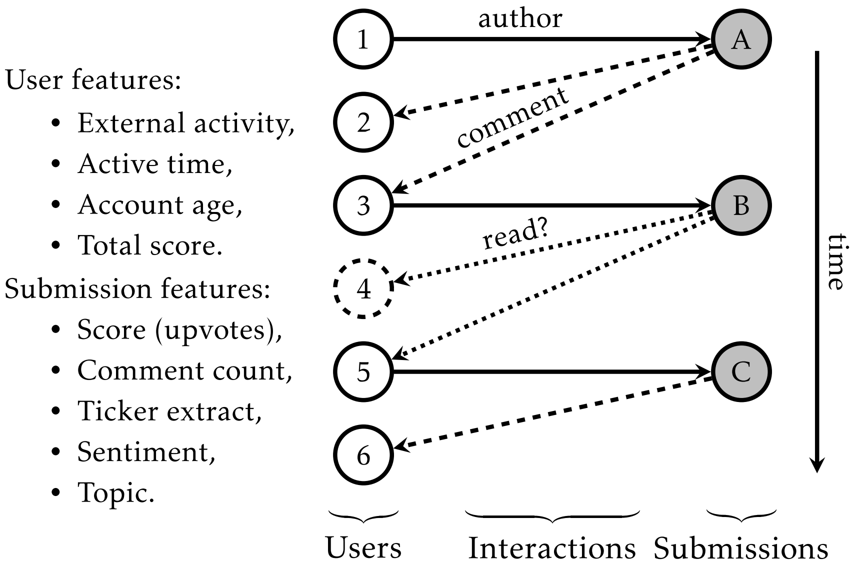


Figure 3: **Diagramatic representation of user interactions on WSB**; we encode how information on a specific asset propagates through WSB in a bipartite graph, with submissions on the right and users on the left. Dotted edges represent possible 'Read' interactions, but take a dashed line if the author comments on the submission. Solid edges denote users authoring submissions. Some of the relevant metadata available for each node type is listed on the left. Only users who author, or comment, on a submission are observed. Other users may not visibly interact with WSB content, but perhaps are exposed to it due to the time they are active on the forum, such as user **4**.

In this paper, we estimate the spread of information among investors using the set of submissions mentioning the same asset. To illustrate the structure of our data, Figure 3 presents a bipartite graph of submissions, on the right, and users, on the left. We encode different types of relationships, such as node **1** authoring submission **A**. Other users, such as users 2 and **3**, comment on submission **A** and may, therefore, be influenced by it. We also consider the potential for authors to read and interact with submissions without commenting, such as the links from submission **B** to authors **4** and **5**. To proxy for these relationships, we consider the time a set of authors are actively discussing the specific asset.

3 Social dynamics among retail investors

We convey our main intuition using a game with strategic complementarities in retail investor decision-making, using established methods (Zenou 2016). Strategic substitutes and complements are first proposed in Bulow et al. (1985) in the context of firms making production decisions – where an increase in production of one firm increases the marginal revenues of competitors. The framework was extended to information acquisition among agents in financial markets by Barlevy & Veronesi (2000), who argue, contrary to Grossman & Stiglitz (1980), that learning among investors can become a strategic complement. Barlevy & Veronesi (2000) suggest that, as the fraction of 'informed' traders in an asset increases and prices, consequently, become more extreme, it may become harder, rather than easier, to identify the asset's payoff. We propose and empirically test a framework with strategic information complementarities among investors on WSB. These complementarities manifest in two manners: i) the sentiments expressed in relation to the future outlook of an asset, and ii) the choice of assets discussed.

3.1 Peer effects due to strategic complementarities

Utility framework with strategic complementarities Suppose that investor i derives the following utility, U, from adopting sentiment ϕ about an asset:

$$U_i(\phi_i) = \phi_i \mathbb{E}_i(r) - \theta \phi_i^2 \mathbb{E}_i \left(r - \mathbb{E}_i(r)\right)^2 + u_i. \tag{1}$$

where $\mathbb{E}_i$ is i's expectation, r is the asset's return, θ is a scalar, and u_i is some idiosyncratic error. Utility is increasing in the asset's expected return, $\mathbb{E}_i(r)$, and decreasing in its expected forecast error. An investor's expectation includes some signal with information both public and private to i, b_i, and also the expressed sentiments of investors they interact with:

$$\mathbb{E}_i(r) = g(b_i) + f(\bar{\phi}_{-i}), \tag{2}$$

where $\bar{\phi}_{-i}$ is the average, revealed peer sentiment on the same asset. We assume that functions $g(\cdot), f(\cdot)$ are monotonic and continuous. Utility can be rewritten as

$$U_i(\phi_i) = \phi_i g(b_i) + \phi_i f(\bar{\phi}_{-i}) - \theta \phi_i^2 \sigma_i^2 + u_i,$$

where σ_i^2 is the variance in the privately-formed signal $g(b_i)$.

Leveraging the framework above, we rationalise the broad following of WSB by Hellwig & Veldkamp's (2009) main result, which we reformulate in Proposition 1. In addition to the quadratic loss assumed in Eq. 1, we remind the reader of the additional assumptions in their framework, adapted to the present context.

Assumption 1. *Investor i can observe another's sentiment, ϕ_j, at a cost $C(\omega_{ij})$, where ω_{ij} takes value one if investor i interacts with investor j, and zero otherwise.*

Assumption 2. *Investors update their sentiments ϕ_i according to Bayes' Law.*

Proposition 1. *If expressed sentiments are complementary, $f'(\bar{\phi}_{-i}) > 0$, the value of additional information is increasing in the information acquisition of other investors.*

Proof. See Hellwig & Veldkamp (2009).

□

Proposition 1 predicts that interactions between investors whose decisions include a social component, a group we call 'hype investors', provides them with additional expected utility, thus increasing their propensity to invest according to their updated sentiment, under costly information acquisition. The result is a statement on the second derivative of utility, with respect to the information acquisition of oneself, and of the investor's peers. The very existence of a website like WSB leverages the desire of retail investors to coordinate their strategies with other like-minded people. One consequence, further developed by Hellwig & Veldkamp (2009), is the emergence of multiple equilibria: we should expect WSB discussions to exhibit self-reinforcement in their content, and sudden regime shifts.

Target independent variable The target independent variable of interest for studying hype investor sentiment is the log-odds of bullish over bearish sentiment:

$$\Phi_{i,t} = \frac{1}{2}\log\left(\frac{P(\phi_{i,t} = +1)}{P(\phi_{i,t} = -1)}\right) = g(b_{i,t}) + f(\bar{\phi}_{-i,(t-1,t)}) + \varepsilon_{i,t}, \tag{3}$$

derived from the utility framework in Appendix B.1. One key addition is the time subscript, t. An author chooses a bullish over bearish strategy depending on: i) a public/private signal, $b_{i,t}$, and ii) the observed sentiments of peers, $\bar{\phi}_{-i,(t-1,t)}$.

Testable propositions The following proposition and corollary state testable implications of investors affecting each other directly, and expressed sentiment complimentarity.

Corollary 1.1. *If expressed sentiments are complimentary, a uniform, marginal increase in peer outlook about an asset will raise the future outlook of an investor about the asset.*

Proof. Follows from Proposition 1 and the assumption that $f(\cdot)$ is monotonic and increasing.

□

In addition to testing for peer effects, retail investor sentiments should consider an asset's historical performance. A rich literature, founded in prospect theory (Kahneman & Tversky 1972, 1973, 1979), discusses the effect of past observations on decision making in a setting with risk (Bordalo et al. 2012, 2019, 2020, Gennaioli & Shleifer 2010). Among others, Wang et al. (2006) demonstrate a Granger-causal relationship between returns, volatility and sentiment. Therefore, our framework should control for the Bayesian investor's observed historical returns as part of the signal set, $b_{i,t}$. We formalize this in the following proposition.

Proposition 2. *$g(\cdot)$ is increasing in stock returns. A uniform, marginal increase in an asset's returns will raise the future outlook of an investor about the asset. It will also indirectly increase outlook of an investor through increasing the outlook of his peers.*

Proof. Follows from the assumption that $g(\cdot)$ is monotonic and continuous, the fact that the investor is Bayesian and from derivations in Appendix B.1. Investors derive their utility from their sentiment and privately formed signal: $\phi_i g(b_{i,t})$. Given the investor is Bayesian, their estimate of future returns will include, and be positive in, the stock's historic returns.

□

Corollary 1.1 and Proposition 2 provide the main empirical implications for the proposed model that can be tested with our WSB data. In subsequent sections, we argue that the data are consistent with Corollary 1.1 and Proposition 2: observed sentiments are well explained by previous stock prices realisations, as well as peer sentiments. WSB, as a platform, is a venue for hype investors to realise their strategic information complementarities, explaining the exponential growth in its userbase.

3.2 Empirical strategy: consensus formation among investors

We empirically validate Corollary 1.1 and Proposition 2 using data from WSB and, thereby, substantiate the proposed framework for social dynamics in investor decision-making. We use two approaches with distinct specifications for 'peers', which we label i) the *Frequent Posters* approach, and ii) the *Commenter Network* approach. We subsequently introduce different instruments to isolate variation in peer sentiment, according to both definitions, orthogonal to current user sentiments.

The *Frequent Posters* approach leverages random, temporal variation in our data to identify peer influence. Given that 4,242 authors post at least twice on the same ticker, we quantify peer influence by identifying the impact of other authors who post about the same asset *between* an individual's two submissions. This allows us to control for the author's sentiment *prior* to exposure to his peers, in addition to market moves. The choice of peers is random due to the nature of Reddit: the time at which authors become active on the forum is random and independent of other authors, since, unlike other social media platforms, users on Reddit do not follow each other. Therefore, an author's 'peer group' is randomly assigned, and there is some naturally occurring variation in the sentiments of peers about specific assets.

One potential concern with the proposed approach is that an exogenous shock could affect peers and the individual who posts multiple times. We employ an instrumental variable (IV) to estimate the sentiments of peers. The use of an IV has become the gold standard to control for endogeneity in peer effects (Angrist 2014, Zenou 2016). The goal is to select an IV which is a good predictor of the independent variable, but unrelated to the dependent variable. The approximation technique involves two steps: first, predicting the independent variable of choice, and, second, using the *predicted* independent variable to estimate its relationship with the dependent variable.

The *Commenter Network* approach considers a submission-to-submission network, with an earlier submission exerting 'peer influence' on a future submission if the author of the later submission commented on the earlier one. In network science terminology, this network is termed a 'projection' of the bipartite graph displayed in Figure 3. The submission-to-submission network could help identify peer effects in user sentiments by offering a more precise filter for those peers an author actually interacts with. Here, we also control for market variables, and employ an IV to address endogeneity. The methods employed in constructing the independent and dependent variables are discussed in Appendix B.2.

3.2.1 Identifying peer influence – frequent posters

The random temporal variation in Reddit users allows us to test for peer effects. We identify the change in an author's outlook on an asset due to the expressed views of their peers. Author i initially expresses a view about an asset j, $\phi_{i,j,(t-1)}$, and, subsequently, creates a new post about the same asset at a later time, with an updated sentiment $\phi_{i,j,t}$. In the time between these posts, the author observes submissions by others on the same asset, in addition to market moves. This framework allows us to control for any public and private signal an author receives prior to exposure to his peers, by directly observing prior sentiment $\phi_{i,j,(t-1)}$.

The target variable is the log-odds of an author expressing bullish over bearish sentiment, $\Phi_{i,j,t}$, as detailed in Appendix B.2. A semi-supervised learning technique, explained in Appendix A.2, estimates the probability a given submission is bullish, $P(\phi_{i,j,t} = +1)$, or bearish, $P(\phi_{i,j,t} = -1)$. These probabilities are then transformed into our target variable, $\Phi_{i,j,t}$, as per Eq. 3.

We first estimate the effect of average peer sentiment between an author's two submissions with the following linear model:

$$\Phi_{i,j,t} = \kappa\bar{\Phi}_{-i,j,(t-1,t)} + X_{i,j,t}\beta + \epsilon_{i,j,t}, \tag{4}$$

where a vector of control variables, $X_{i,j,t}$, includes the author's previous sentiment about stock j, $\Phi_{i,j,(t-1)}$, as well as controls for market movements and ticker-specific fixed effects, and β is a vector of corresponding coefficients. Even though peers are randomly assigned in this formulation, an exogenous shock may affect the views of both peers and an author simultaneous in the period $(t-1, t)$. For this reason, the Ordinary Least-Squares (OLS) estimates do not enable us to make causal inference about peer influence.

To tackle this issue, we use the historical view of peers as an IV for their view expressed within $(t-1,t)$. Our choice of IV is founded in psychology: Ross et al. (1975) find that 'once formed, impressions are remarkably perseverant and unresponsive to new input, even when such input logically negates the original basis for the impressions', with many later studies, such as Anderson et al. (1980), supporting these findings. Our approach is similar to that employed in the peer effects literature on educational outcomes. For example, Duflo et al. (2011) use a student's admission test scores to predict their ability. We estimate investor k's sentiment (a peer of investor i) about asset j, $\Phi_{k,j,t}$, based on the sentiment they expressed previously, $\Phi_{k,j,t-1}$:

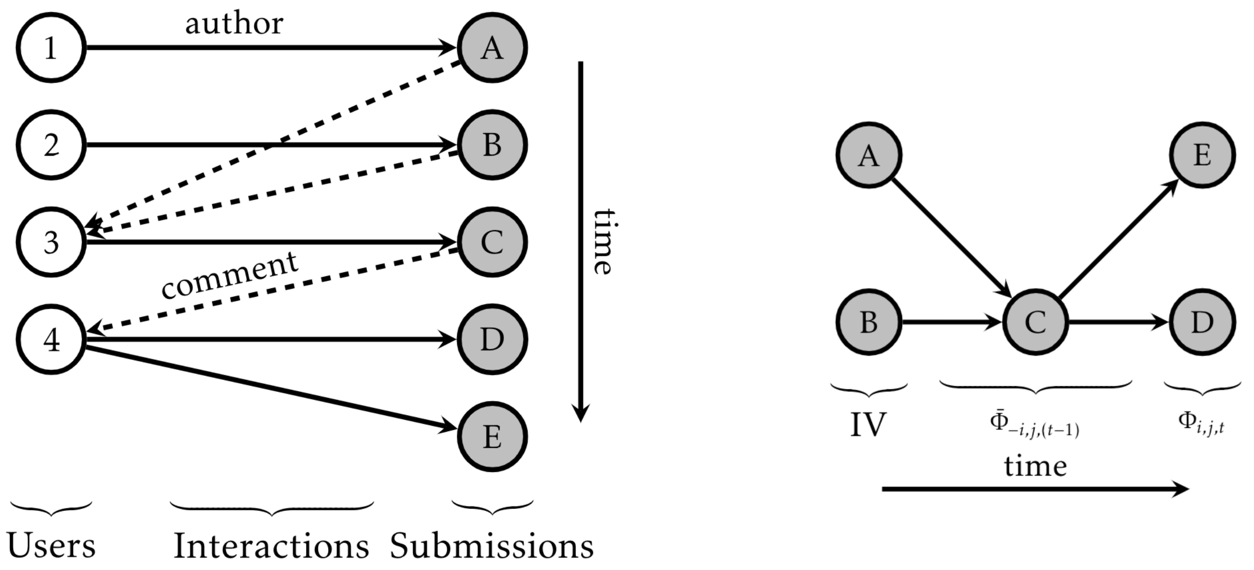
$$\Phi_{k,j,t} = \kappa_0\Phi_{k,j,t-1} + \epsilon^0_{k,j,t}, \tag{5}$$

where $\epsilon^0_{k,j,t}$ is an idiosyncratic error, and κ_0 a coefficient. Eq. 5 is estimated using a sample containing submissions by all authors who post multiple times. The F-statistic for this first stage estimate, presented in Table 1, supports that this is a strong instrument. The choice of this IV gives a good approximation for author sentiment, while controlling for common shocks in author k and i's sentiments.

We subsequently use the *predicted* outlook of peers between an author's posts to estimate peer effects. Equation 4 is estimated by the instrumental variable: sentiments expressed by peers prior to $(t-1)$ are used to estimate $\bar{\Phi}_{-i,j,(t-1,t)}$. The vector of control variables, $X_{i,j,t}$ is composed of stock-specific fixed effects, author i's past sentiment, and stock log returns, both on day t and the average of five days preceding t, and the variance in log returns in the five days prior to day t. In the Two-Stage-Least-Squares (2SLS) estimation, the coefficient on $\hat{\bar{\Phi}}_{-i,j,(t-1,t)}$, denoting predicted peer sentiment, reflects the causal effect of peer sentiment on an author's expressed sentiment about an asset.

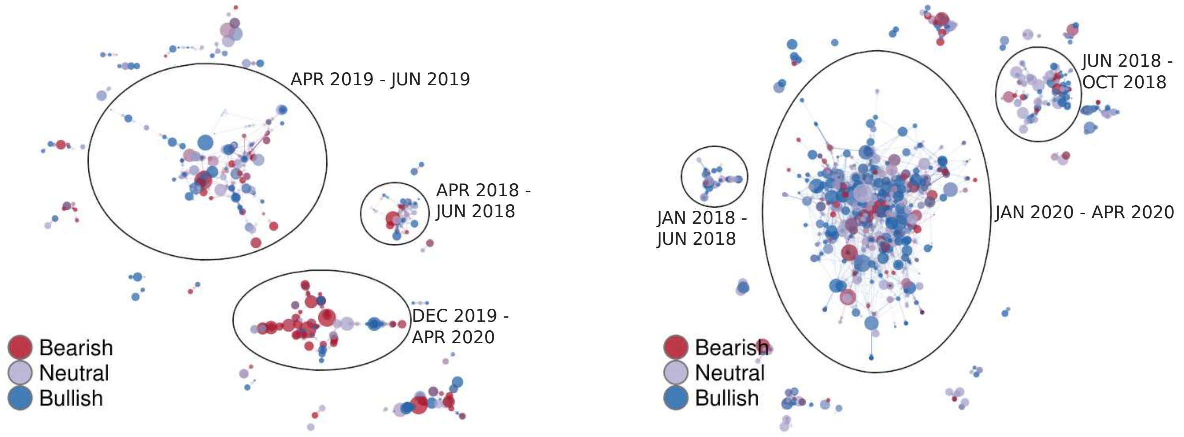
Credible estimation We check whether our estimation strategy is credible, with respect to the three problems proposed by Zenou (2016) in estimating peer effects. The first issue is in separating *peer effects* from *contextual effects* – the tendency of perspectives to vary with some observable characteristics of the group, rather than individuals influencing each other. Controls for asset price movements and ticker specific characteristics – the main source of exogenous variation – address these. Second, the random, anonymous nature of WSB, as well as controlling for ticker-specific fixed effects, address the possibility for *correlated effects*. The specification with the IV helps solve the *common shock* problem. A statistical analysis of our identification strategy is included with the results.

3.2.2 Identifying peer influence – commenter network



(a) Bipartite Interaction Graph between Authors and Submissions

(b) Submission-to-Submission Projection of Interaction Graph in Figure 4a



(c) Submission-to-Submission Network: DIS

(d) Submission-to-Submission Network: MSFT

Figure 4: **User networks in WSB conversations**; WSB data is summarised as a bipartite graph, illustrated in Figure (a), where users (left) are linked to submissions (right) when they author the submission (solid edge) or comment on the submission (dashed edge). The resulting projection of submissions, in Figure (b), tracts the propagation of sentiments, Φ. The data for two stocks, Figures (c) and (d), reveal that individuals post more submissions that are bullish(bearish) at times when the price of an asset increases(decreases) dramatically, with some visual evidence that similar sentiments tend to cluster.

WSB allows us to trace the interactions of users through a commenting networks, even though there are no user friendship ties. We exploit a submission-to-submission interaction network for each asset, tracking which submissions in the past influence future submissions based on authors' commenting histories. This method offers a more precise way to identify an individual's peers by observing which individuals, and submissions, an author explicitly interacts with.

We build an example network in Figure 4a for illustration. User **3** comments on submissions **A** and **B** prior to creating their own submission, **C**. Therefore, the sentiments in **C** could be influenced by sentiments in **A** and **B**. User **4** comments on submission **C** before creating their own two submissions, **D** and **E**. Figure 4b visualises the resulting projection of this bipartite user-to-submission network, onto a submission-to-submission network. We place a directed edge from an earlier submission to a later one if the author of the later submission commented on the earlier submission. Submissions with the same author are not linked – an author's own previous submissions about a ticker are considered as a separate, independent variable in evaluating peer effects.

Two examples of submission-to-submission networks in our data are displayed in Figures 4d and 4c. Distinct temporal clusters emerge on the forum as WSB participants turn their attention to the asset in question. They eventually loose interest, until a new cluster emerges. Some discussions appear fragmented: the DIS discussion in Figure 4c, for example, contains several smaller clusters, with distinct differences in overall sentiments. Others, such as the MSFT discussion in Figure 4d, contain a giant component where investors with different sentiments interact.

We use an IV approach to estimate peer influence. As the first stage, we estimate the sentiments of neighbours to estimate an author's view. As indicated in Figure 4b, the sentiments in submissions **A**, **B** can be used to predict that of submission **C**. The *predicted* sentiment of **C** can then, in turn, be used to predict the sentiments of **D** and **E**. This choice of IV is well-established in the networks literature, and discussed by Zenou (2016). It controls for exogenous events that might affect both the individual and his peers. For the network approach, the peer sentiment, $\bar{\Phi}_{-i,j,(t-1,t)}$, is the average *estimated* sentiment expressed within posts about the same ticker by a different author, that an individual has commented on prior to posting. We also modify our control for an author's past sentiment about the stock to account for authors who post for the first time: a dummy variable encodes whether the author's most recent previous post is bearish, neutral, bullish or missing.

The *Commenter Network* offers certain upsides, but also certain shortcomings, as compared to the *Frequent Posters* approach. The network method more precisely identifies the channels of influence between authors. However, the allocation of peers is no longer random, since the network structure is governed by a *choice* to comment on certain submissions over others. We use techniques outlined in Patacchini & Zenou (2016) and Bifulco et al. (2011) to address the three main challenges in correctly estimating peer effects through a network of interaction, as per Zenou (2016). The two approaches follow a similar strategy and allow us to paint a coherent picture for the influence of social interaction on retail investors' sentiment formation on WSB.

3.3 Results: Consensus Formation and Peer Effects

In this section, we present the Reduced Form, Second Stage, and First Stage regression estimates for both the *Frequent Posters* and *Commenter Network* approaches outlined above. The Reduced Form and Second Stage estimates, across both model specifications, show that peer sentiments directly impact an individual's sentiment about an asset, with the individual conforming to his peers.

Table 1 presents the results, with Panel A presenting OLS estimates for κ using observed variation in peer sentiments, and Panel B.1 using predicted variation in peer sentiments. We relegate estimated coefficients for control variables to Appendix B.2. Estimated peer effects can be summarised as follows: an average estimate for κ at 0.09, in the reduced-form case, means that an average doubling in the odds of peers expressing bullish over bearish sentiments increases the odds of a given submission to be bullish, over bearish, by 6.4%, on average (we raise the log-odds estimate for $\Phi_{i,j,t}$ to an exponent). In the IV setting, an average estimate of 0.26 translates to an increase in the corresponding

Table 1: Peer influence in WSB sentiments

	Frequent Posters (1)	Network (2)
Panel A: peer influence estimated using *observed* average sentiment of peers		
	Independent Variable Average peer sentiment, $\bar{\Phi}_{-i,j,(t-1)}$ (observed)	
Investor Sentiment ($\Phi_{i,j,t}$)	0.11 (0.02) ***	0.07 (0.01) ***
Author & asset controls ($X_{i,j,t}$)	Yes	Yes
Observations	11,366	19,639
Adjusted R^2	0.07	0.05
Panel B.1: peer influence estimated using *predicted* average sentiment of peers		
	Independent Variable Average peer sentiment, $\hat{\bar{\Phi}}_{-i,j,(t-1)}$ (predicted)	
Investor Sentiment ($\Phi_{i,j,t}$)	0.16 (0.06) ***	0.35 (0.07) ***
Author & asset controls ($X_{i,j,t}$)	Yes	Yes
Observations	8,558	19,639
Adjusted R^2	0.05	0.05
J-statistic	0.00	0.00
Panel B.2: estimating peers' sentiments		
	Independent Variable	
	Historical Sentiment of Peers	Sentiment of Neighbours' Neighbours
Sentiment of Peers	0.32 (0.01) ***	0.19 (0.01) ***
Author & asset controls ($X_{i,j,t}$)	No	No
Observations	15,737	27,630
Adjusted R^2	0.10	0.10
F-statistic	1,823	655

Notes: in column (1), the first stage is estimated using the initial sentiment expressed by an author about an asset to estimate his sentiment in the following post. In column (2), the first stage is estimated using the sentiment of the previous posts that the author of a post has commented on (about the same asset). The second stage is estimated using the average predicted sentiment of peers. Ticker-level dummies, asset return and volatility controls, and the intercept are included in the second stage and reduced form estimates, but not shown – the full estimates are presented in Appendix B.2. Robust standard errors, clustered at the ticker level for Panels A and B.1, are presented in parentheses. Observations with incomplete data are dropped.

*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

odds for bullish over bearish sentiments by 19.3%, instead. In all cases, the robust standard errors, clustered at the ticker level, produce estimates statistically significant at the 1% level.

The estimated coefficients in columns (1) and (2) of Panel B.1, specifically, suggest that an exogenous increase in average peer outlook appears to increase an investor's own future view about an asset. These findings demonstrate that the data are consistent with Corollary 1.1. As a result, we conclude that the data support a model where strategic complementarities govern the investment decisions of retail traders on WSB. Estimates for asset price movements also uphold Proposition 2. These results are discussed further, alongside the complete regression estimates, in Appendix B.2.

Support for identification One potential concern is that individuals who post multiple times about the same asset, or those who comment on others' submissions, may be different from investors who post once. If this were the case, our findings would not allow us to draw valid conclusions about the overall population of WSB investors. We provide evidence that sentiments expressed by our samples of posters are similarly distributed to those of the overall population in Appendix B.2.

A second concern is whether our proposed independent variables – asset price movements, ticker fixed effects and author historical sentiment – are effective controls for unobserved ticker characteristics. If our controls in the *Frequent Posters* formulation are valid, then a randomly selected cohort of peers, composed of individuals who post on the same ticker *before* the author's first post, should have no additional effect on the sentiment expressed in dependent submissions. Similarly, if our controls are useful in the *Commenter Network* formulation, a random rewiring of the network should yield no effect. The results are detailed in Appendix B.2: no statistically significant correlation of the randomly selected cohorts is produced. This provides further evidence that unobserved factors influencing within-ticker variation in both peer composition and author sentiment are not confounding our estimates.

A final concern with our approach is that historical peer sentiment or sentiments expressed by a neighbour's neighbours are systematically correlated with an author's opinion about an asset – in other words, that our instruments are not exogenous. We regress the residuals from our second stage on our instrumental variable and our endogenous regressors, a procedure known as the Sargan-Hansen test. A J-statistic of $1.06\times e^{-28}$ in the *Frequent Posters* case, and a J-statistic of $2.23\times e^{-28}$ in the *Commenter Network* case, imply that the residuals are not correlated with our instrumental variable: historical sentiments of peers and those of an author's neighbours' neighbours are exogenous.

3.4 Contagion dynamics and the origin of bull runs

Do investors influence each other's choice of assets? Previous research by Shiller (1984) and Banerjee (1993) emphasizes the role word-of-mouth information transmission plays in an investment context. In one empirical example, Banerjee et al. (2013) investigate the diffusion of microfinance by separating the uptake of policies by villagers in India, under peer influence, and the spread of awareness on those policies via influential leaders. In the WSB context, we would similarly expect awareness about specific assets to spread from one user to another. The emphasis of this section is not on identifying a causal relation, but rather understanding the dynamics which govern asset interest among investors. These insights, combined with our understanding of how investors adopt each other's sentiments, allows us to paint a more complete picture of retail investors' decision making and the resultant stock market dynamics.

Let us consider a pool of individuals searching for a lucrative investment. Before an individual chooses to express a sentiment about a stock, the stock must first capture his attention. If stock j captures his attention in period t, we consider that the individual will: i) decide his sentiment to maximize utility according to our Eq. 1, and ii) inform other investors about his interest in the stock – as expected in the presence of strategic complementarities. In this section, we are interested in the log-odds of an author posting about stock j over a baseline. Our baseline is the probability of posting about a stock that is not widely discussed within the forum, s_t, defined as an asset receiving fewer than 30 submissions within our sample.

Suppose that the log-odds of posting about a popular asset j over an unpopular stock are given by

$$l(a_{j,t}) = \log\left(\frac{a_{j,t}}{s_t}\right) = h_1(a_{j,t-1}(1-a_{j,t-1})) + h_2(a_{j,t-1}) + h_3(\bar{\phi}_{j,t-2}\bar{r}_{j,t-1}) + h_4(\bar{\phi}^2_{j,t-2}\sigma^2_{j,t-1}) + h_5(\bar{\phi}_{j,t-1}) + \zeta_{j,t}. \tag{6}$$

The function $h_1()$ captures the rate of independent mixing between investors aware of stock j, $a_{j,t-1}$, with unaware investors, $1-a_{j,t-1}$, creating newly informed investors. $h_2()$ captures the rate at which aware investors *remain* aware and engaged in discussions between $t-1$ and t. Intuitively, functions $h_1()$ and $h_2()$ capture how *popular* an asset has been in the recent past. The latter terms control for the asset's perceived *profitability*. $h_3()$ is a 'quality of signal' term – how correct previous investors were at predicting returns,' and $h_4()$ is a 'noise of signal' term. Finally, $h_5()$ captures the overall author propensity to prefer adopting bullish over bearish sentiment. We propose that the log-odds of posting about asset j in time period t are increasing in $h_1(), h_3(), h_5()$ and decreasing in $h_2(), h_4()$.

Empirical Strategy To measure the impact of *profitability* and *popularity* of past discussions about a stock on the log-odds of it being discussed in the future, we produce OLS estimates for the linear model

$$l(a_{j,t}) = ca_{j,t-1}(1-a_{j,t-1}) + da_{j,t-1} + \beta_1\bar{\phi}_{j,t-2}\bar{r}_{j,t-1} + \beta_2\bar{\phi}^2_{j,t-2}\sigma^2_{j,t-1} + \beta_3\bar{\phi}_{j,t-1} + X_j\beta_4 + \zeta_{j,t}, \tag{7}$$

where $a_{j,t-1}$ is the share of all active investors who post about ticker j at times $t-1$: $a_{j,t} \in [0,1]$ for all j and t, $\bar{r}_{j,t-1}$ is the average log return in $t-1$, and $\sigma^2_{j,t-1}$ is the variance of the same log returns (these variables are consistent with Section 3.2, and outlined in Appendix B.2). $\bar{\phi}_{j,t}$ is the average investor sentiment about asset j at time t – this is sum of the probabilities of all active investors in stock j at t to be bullish less the sum of the probabilities of them to be bearish, where the probabilities are the same as in Section 3.2. X_j is a vector stock-specific fixed effects. t denotes time, in weeks. The choice to aggregate over weeks is done to address sparsity of submissions, especially pre-2017. Two more choices are made to tackle this: i) we restrict ourselves to a sample that spans January 2016 to July 2020, and ii) we categorise stocks mentioned fewer that 30 times since January 2016 into an 'other stocks' group, which forms our benchmark s_t. For completeness, we also consider a different formulation where we test for the impacts of $\bar{r}_{j,t-1}$ and $\sigma^2_{j,t-1}$ directly.

Results Our results, presented in Table 2, demonstrate that WSB users follow each other in their choice of investment instruments. There is strong evidence that the homogeneous mixing property explains the uptake of new assets: using estimates in column (1), an increase in the share of authors discussing stock j from 0.1 to 0.2 increases the ratio of authors discussing j over 'other stocks' in the following week by a factor of two. This is opposed to an increase from 0.2 to 0.3, which prompts a collapse in the ratio of authors discussing j over 'other stocks' by 75%, on average – the difference is driven by the large negative coefficient on $a_{j,t-1}$. This is strongly reminiscent of epidemic contagion models, adapated to the spread of narratives (Banerjee 1993, Shiller 2017). Furthermore, users talk about opportunities with a proven track record: they are interested where returns have been predicted *correctly*, as seen in the interaction between the twice-lagged sentiments and once-lagged returns, and *consistently*, seen in the interactions in squared, twice-lagged sentiments and once-lagged variance in returns.

Table 2: Evolution of stocks discussed on WSB

	Dependent variable: $l(a_{j,t})$			
	(1)	(2)	(3)	(4)
$a_{j,t-1}(1-a_{j,t-1})$	112.74*** (11.33)	104.75*** (10.68)	68.55*** (6.61)	62.38*** (6.34)
$a_{j,t-1}$	−71.43*** (9.72)	−65.72*** (9.08)	−41.65*** (5.40)	−37.30*** (5.06)
$\bar{r}_{j,t-1}$	0.83* (0.45)		0.96** (0.48)	
$\sigma^2_{j,t-1}$	−2.48*** (0.80)		−1.42** (0.72)	
$\bar{\phi}_{j,t-2}\bar{r}_{j,t-1}$		1.53 (1.18)		2.24* (1.22)
$\bar{\phi}^2_{j,t-2}\sigma^2_{j,t-1}$		−18.27*** (3.58)		−8.89*** (2.99)
$\bar{\phi}_{j,t-1}$		0.08*** (0.03)		0.06 (0.04)
Constant	−3.98*** (0.02)	−3.89*** (0.03)		
Ticker FE	No	No	Yes	Yes
Observations	6,381	4,470	6,381	4,470
Adjusted R^2	0.37	0.39	0.14	0.14
F-Statistic	936.84	564.28	317.70	194.78

Notes: This table presents OLS estimates for the log-odds of users discussing stock j in week t, over a collection of stocks that have not been mentioned more than 30 times. Explanatory variables include: the lag in the share of authors discussing j, $a_{j,t-1}$, the interaction with the share of authors not discussing j, $a_{j,t-1}(1-a_{j,t-1})$, as well as the lag in stock j's weekly average log returns, $\bar{r}_{j,t-1}$, and variance, $\sigma^2_{j,t-1}$. In columns (3) and (4), the average log return is multiplied by the two period lag in the average sentiment expressed among WSB submissions on stock j, $\bar{\phi}_{j,t-2}$, and the variance in log returns by the same sentiment's square, $\bar{\phi}^2_{j,t-2}$. The one period lag in sentiments is also included in those specifications. Columns (3) and (4) include stock-specific fixed effects. Accompanying standard errors, displayed in brackets, are clustered at the stock level, and calculated in the manner outlined in MacKinnon & White (1985).

*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

When we consider the impacts of stock-specific variables in isolation, presented in columns (1),(3) in Table 2, volatility appears to be a leading factor for authors deciding what asset to discuss. Stock-specific heterogeneity matters for the average return itself; when pooling all observations, average returns are not a statistically significant predictor for the number of users posting about the related stock. These are statistically significant at the 5% level in column (3), indicating that discussion sizes are stimulated by large, notably positive, returns for tickers that may already be heavily discussed in the first place.

We focus on columns (2), (4) to discuss the correctness and consistency of past WSB predictions in an asset. While older sentiments alone do not appear to be useful in modelling the number of authors discussing a stock, as per columns (2) and (4), their interaction with the stock's returns are important. Looking at column (2) specifically, the impact of variance is larger if users held strong feelings on the stock, either in buy or sell directions, in the preceding week. This leads us to conclude that unreliable advice, promoting an ultimately risky strategy, dampens future conversation. This is compounded by the effect of large returns when explaining stock-specific variation, in column (4): positive(negative) returns following bullish(bearish) sentiment promote further discussions in the following week, and vice versa, although the coefficient of interest is only significant at the five percent level.

In summary, the composition of discussion between different stocks exhibits strong temporal persistence. This is not only because WSB serves to turbo-charge the spread of new stock opportunities within the hype investor community, but also because the forum reinforces associated narratives when existing opportunities exhibit large, consistent and, from the users' perspective, predictable returns.

4 Bull runs and instability: the market impact of social dynamics

Section 3 studies the social dynamics among retail investors. An outstanding question is whether the observed behaviours have implications for the financial markets. The debate regarding the importance of investor psychology in finance dates back to Shiller (1984) and Black (1986), labeling 'noise' traders as those who form expectations deviating from 'rational' rules. This discussion sparked a new body of research, separating market participants into two categories: i) a fully rational type, and ii) those who deviate from rational expectations (Shefrin & Statman 1994, Shleifer & Summers 1990, De Long et al. 1990). Authors argue for specific behaviours in each investor type, to ultimately derive predictions for market stability and price dynamics.

In subsequent years, researchers scrutinised the interactions of heterogeneous agents in stock markets, following many, sometimes adaptable, rules – Hommes (2021) offers a comprehensive review. In relation to social dynamics specifically, investor crowding and coordination is famously explored in Kirman's (1993) ant study. Avery & Zemsky (1998) discuss the asset price impact of herd behaviour, depending largely on the price charged by 'a competitive market maker' and the perceived informativeness of trades during periods of herding. Coordination among groups of investors in a type of network is discussed in Cont & Bouchaud (2000), albeit at much shorter time scales than explored within this paper.

A key feature scrutinised in this literature are the multiple equilibria that emerge in the presence of strategic complementarities. In Kirman's (1993) model, the rate at which ants conform in their exploitation of one food source over another determines whether a colony exploits two food sources simultaneously (one stable equilibrium), or they exploit one food source at a time, under the threat of a shock which causes a regime switch (two stable equilibria, and one unstable). Barlevy & Veronesi (2000) and Hellwig & Veldkamp (2009) discuss how strategic learning complementarities among investors can lead to multiple equilibria.

This is of particular relevance to WSB, especially in the aftermath of the infamous GameStop short-squeeze. Section 3 presents evidence that retail investors who are active online have strong tendencies to imitate each other's strategies. Over time, their strategies may exhibit abrupt shifts, either in the asset they trade, or the direction they trade it in. Conditional on these investors' purchasing power, does this instability in demand translate to instability in price?

In this section, we model the market impact of the social behaviours we observe, discerning between two types of agents – 'hype investors', who are susceptible to social forces, and 'other investors', who are not. Our model includes one endogenous variable, namely consensus in hype investors' strategies, and abstracts away from other sources of price variation. We outline where further modifications can be made, for example with respect to rules 'other investors' follow.

Our model provides an important motivation for us to test the impact that consensus, as well as contagion, exhibited by WSB, had on stock markets. Specifically, we exploit variation in *changes* of stock discussions that can be explained by the temporal dynamics of hype investors' behaviours, namely: i) interest depends on the information transmission mechanism, as in Section 3.4, and ii) the direction of the trade features strong temporal persistence due to strategic complementarities, as in

Section 3.2.

4.1 Modelling markets with social dynamics

Consider a market for one asset with two types of participants: 'hype investors', who buy quantity Y, and 'other investors', who buy quantity S. Time, denoted t, and the asset's price, p, are assumed continuous and differentiable. Market participants observe the behaviour of others, as well as the asset's price, and update their demand for the asset to maximize their expected utility. In this way, we set up a partial model to understand the impact of social dynamics in isolation: the endogenous variable is sentiment, which governs the propensity for the average hype investor to buy or sell the asset, thus shifting demand Y.

Total demand for shares at a given price and time, $Y(p,t)$ and $S(p,t)$, equals the number of shares outstanding, Q:

$$pQ = p(Y + S). \tag{8}$$

We model the emergent price dynamics due to social dynamics, namely consensus, rather than fundamental news. The 'other' investor is unaffected by social forces and chooses to keep the nominal value invested in the asset fixed, $d[pS]/dt = 0$. For completeness, Appendix C.1 considers alternative assumptions.

Our goal is to model the instantaneous change in an asset's price, $\dot{p}/p$, where a dot denotes the derivative with respect to time, determined by the change in hype investor demand, $d[Y(p,t)]/dt$. We leverage our findings from Section 3 and propose that hype investor demand is a function of two distinct components: i) an aggregated 'consensus' component, denoted by $\phi(p,t)$, which is the average buying/selling intensity across all active hype investors, with $\phi(p,t)$ in the interval $[-1,1]$, and ii) a 'contagion' component, denoted by $a(t)$, which captures the fraction of hype investors interested in the asset at a given point in time, with $a(t)$ in the interval $[0,1]$. Two other constants play a role: M, the purchasing power of one individual hype investor, and N, the total number of hype investors, which determine aggregate hype investor demand:

$$Y = MNa(t)\phi(p,t). \tag{9}$$

Intuitively, market impact depends on the interplay between consensus and contagion. As $a \to 1$, more and more hype investors are interested in the asset, increasing their market impact. As $\phi \to +1$ or $\phi \to -1$, the investment strategies of hype investors become more coordinated, increasing the market impact of the hype traders already invested in the asset. When $\phi \to 0$, on the other hand, new hype investors trade against each other and do not affect the price. As M and N increase, hype investors are also likely to have a higher market impact.

The derivative of Eq. 8 with respect to time yields:

$$\frac{\dot{p}}{p} = \underbrace{\frac{MN}{pQ}}_{\text{(1) Capacity}} \times \underbrace{\frac{1}{1+(1-s)\varepsilon_{\phi}}}_{\text{(2) Response}} \times \underbrace{\left(\phi\dot{a} + \dot{\phi}a\right)}_{\text{(3) Demand}}, \tag{10}$$

where $\dot{\phi} = \frac{\partial\phi}{\partial t}$ is the partial rate of change in buying intensity with respect to time, $\dot{a} = \frac{da}{dt}$ is the derivative of our contagion function with respect to time, $\varepsilon_{\phi} = -\frac{p}{\phi}\frac{\partial\phi}{\partial p}$ is the price elasticity of the buying intensity, and $s = S/Q$ is the fraction of shares owned by other investors. The precise steps taken in the derivation are outlined in Appendix C.1.

We assume that the share of informed investors, a, varies slowly relative to price p. Component (1), labelled 'capacity', is a market depth term, measuring how much market power hype investors have, as a share of the total market cap of the stock. Component (2), termed 'response', captures the price elasticity of hype investors, weighted by the fraction of shares owned by other investors s. Let us consider a simple exercise with components (1) and (2). As the share held by other investors approaches one ($s \to 1$), the price elasticity of the buying intensity of hype investors ceases to matter, since 'other' investors are setting the price. The response term approaches one at this limit, and the price change is simply the extra money hype investors pour into the asset, divided by the total nominal value of all shares. At the other extreme, as s approaches zero, all new shares are bought from existing hype investors: the response term is now a function of the inverted price elasticity, as any extra demand for shares must be met by supply from within hype investors' holdings.

An important point is that $a(t)$ is assumed independent of price. In addition to the behaviour of other investors, the reaction of hype investors to enter or exit the market should, in theory, depend on price movements, as we find in Section 3.4. However, for the purposes of this paper it suffices to note that at least one of the social components, consensus or contagion, must be endogenous to price changes. In testing the hypothesis that social dynamics matter for asset prices, the procedure must account for this reverse causality.

Term (3) captures the interplay of social dynamics, which are the focus of this paper: 'contagion', the rate at which new hype investors enter the market for the asset, and 'consensus' formation, the overall change in buying intensity among hype investors. Both serve to shift the asset's demand at any given price point. The first part of (3), $\phi\dot{a}$, quantifies how many active hype investors enter the market for the asset at a given buying intensity ϕ. The second part of (3), $\dot{\phi}a$, captures the change in buying intensity for a given number of hype investors.

Individual buying intensity, as in Section 3.3, is a function of aggregate buying intensity by other hype investors, $\phi(t,p)$, changes in the asset's price, as well as some prior, which we denote here as ϵ_i:

$$\phi_i(p,t) = \epsilon_i + \alpha\phi(p,t) + \beta\frac{\dot{p}}{p} - \gamma\left(\frac{\dot{p}}{p}\right)^2, \tag{11}$$

where γ captures aversion to large swings in price, which we proxy using return volatility in our empirical exercise. Under the assumption that ϵ_i are random and follow a type-I Extreme Value distribution,[7] we can aggregate all hype investors' sentiments:

$$\phi(p,t) = \tanh\left[\frac{\beta\frac{\dot{p}}{p} + \alpha\phi(p,t)}{2\gamma\left(\frac{\dot{p}}{p}\right)^2}\right]. \tag{12}$$

The properties of the hyperbolic tangent function have received much attention in the area of strategic decision-making (Brock & Durlauf 2001, Bouchaud 2013). An important feature is the potentially destabilising effect of noise in investor decision-making, which is in the denominator of the expression in brackets. In practice, it implies that higher stock price volatility dissuades investors from buying or selling ($\phi \to 0$), which we capture in our WSB data by observing neutral sentiments.

Key implications for price impact The model suggests that social dynamics destabilise financial markets, but the resulting change in price depends on several other factors. One important consideration is whether hype investors are sensitive to higher prices. If they are reactive to small changes in price, such that ε_ϕ is large, then the overall impact of a shift in demand may be small. Appendix C.2 explores the resultant price dynamics in more detail. Excitement about an asset is slow to take off, however, coordination among investors gradually moves the price in a self-reinforcing manner. When ϕ is positive, this manifests in a slow run-up in price, driven by the gradual increase in active hype investors due to contagion. Individuals eventually begin to doubt the value of an asset; the exit of investors from the market, after the number of aware investors peaks, creates panic selling among remaining investors, and volatility in price. The ability of hype investors to move the market is determined by how well they coordinate and whether they can form consensus to buy or sell persistently. The consensus parameter, α, thus plays a key role in determining market stability.

4.2 Market impact of social dynamics in WSB

Section 4.1 proposes that asset prices are affected by both coordination and consensus among hype investors. For an empirical evaluation, we consider the following relationship from Eq. 10:

$$\frac{\dot{p}}{p} = w_1 \underbrace{\left(\frac{\phi \dot{A}}{pQ}\right)}_{\dot{\Omega}} + w_2 \underbrace{\left(\frac{\dot{\phi} A}{pQ}\right)}_{\dot{\chi}}, \tag{13}$$

where we aggregate the number of active hype investors, $A = aN$, and where the remaining components of Eq. 10 are assumed fixed. This section establishes a causal relationship between social dynamics, proxied by WSB activity, and a set of stock market variables. To that end, we separately construct a measure for contagion, denoted by Ω, and consensus, by χ, using historical data from WSB.

4.2.1 Key variables

Dependent variables Section 4.1 proposes a relationship between asset price and hype investor activity. We expand our analysis to consider three variables that summarise stock market activity; i) mean returns, ii) variance in returns, and iii) trading volumes. Specifically, we consider the following target variables:

- $\Delta \bar{r}_{j,t}$, the first-difference in stock j's mean daily log return within calendar week t,
- $\Delta \sigma^2_{j,t}$, the first-difference of the variance in stock j's daily log returns during week t,
- $\Delta v_{j,t} = \Delta V_{j,t}/V_{j,t-1}$, the percent change in stock j's average daily nominal trading volume in week t.

Volumes are normalized to compensate for market capitalization and volume trading heterogeneity among stocks. The relationship between returns and social dynamics hypothetically follows Eq. 13. Volatility and volumes keep a similar relationship, but do not discriminate the direction in which hype investors trade – buying and selling generate the same amount of price volatility and volumes.

Independent variables We construct two explanatory variables to evaluate the stock market impact of social dynamics. The first proxies contagion, $\phi \dot{A}/pQ$:

$$\Delta \Omega_{j,t} = \frac{\phi_{j,t-1}}{q_{j,t-1}} (A_{j,t} - A_{j,t-1}), \tag{14}$$

where $A_{j,t}$ is the number of unique authors on WSB whose submissions solely mention stock j in week t, $\phi_{j,t-1}$ is the average sentiment to buy ($\phi = +1$) or sell ($\phi = -1$) expressed among submissions in week $t-1$, and $q_{j,t-1} = p_{j,t-1} \times Q_{j,t-1}$ denotes j's average market cap in week $t-1$. The second proxies consensus formation, $\dot{\phi} A/pQ$:

$$\Delta \chi_{j,t} = \frac{A_{j,t-1}}{q_{j,t-1}} (\phi_{j,t} - \phi_{j,t-1}). \tag{15}$$

They combine distinct components of hype investor demand. Contagion tracks the change in the number of investors interested in stock j as an investment opportunity, fixing the prevailing sentiment. On the other hand, consensus gauges the change in asset demand due to changes in the intensity of hype investor sentiments that determine their investment strategies, keeping their number fixed. In both cases, we divide by the market cap of the relevant stock to control for its market depth.

We adjust both independent variables when modelling volatility and volumes, specifically;

$$\Delta\Omega^*_{j,t} = \frac{|\phi_{j,t-1}|}{q_{j,t-1}}(A_{j,t} - A_{j,t-1}), \quad \Delta\chi^*_{j,t} = \frac{A_{j,t-1}}{q_{j,t-1}}(|\phi_{j,t}| - |\phi_{j,t-1}|). \tag{16}$$

Absolute values in average weekly sentiments are better suited, since the direction of the strategy, i.e. more bullish or more bearish, is not important, as opposed to the absolute size of associated sentiments, which generate new trading activity.

Distinguishing between these two types of asset demand sheds light on the origins of bull runs from WSB. On one hand, does asset demand stem from WSB due to its role as an information sharing platform, alerting growing numbers of amateur investors to profitable opportunities? On the other hand, does WSB serve as a coordination platform for existing users, who strategically reinforce each other's decision to enter a risky position?

4.2.2 Empirical Strategy

First, we formulate the linear relationship between market variables and social dynamic variables of choice. We then propose an estimation strategy, following the findings in Sections 3 and 4.1. The key challenge in measuring the impact of consensus and contagion on stock market variables is the reverse causality with respect to price changes, as established by Eq. 10.

Reduced form: We regress changes in weekly log returns, their variance, and percent change in average daily trading volumes on both measures for contagion and consensus:

$$\Delta\bar{r}_{j,t} = \beta_{\Omega,r}\Delta\Omega_{j,t} + \beta_{\chi,r}\Delta\chi_{j,t} + \eta_{r,t} + \varepsilon_{r,j,t}, \tag{17}$$

$$\Delta\sigma^2_{j,t} = \beta_{\Omega,\sigma}\Delta\Omega^*_{j,t} + \beta_{\chi,\sigma}\Delta\chi^*_{j,t} + \eta_{\sigma,t} + \varepsilon_{\sigma,j,t}, \tag{18}$$

$$\Delta v_{j,t} = \beta_{\Omega,v}\Delta\Omega^*_{j,t} + \beta_{\chi,v}\Delta\chi^*_{j,t} + \eta_{v,t} + \varepsilon_{v,j,t}, \tag{19}$$

where β_Ω and β_χ are coefficients of interest, η_t denote time fixed effects, $\varepsilon_{j,t}$ an idiosyncratic error. Subscripts r, σ, v serve to differentiate the coefficients, which we estimate separately for each dependent variables.

This reduced-form setup does not allow us to argue that a causal relationship exists between hype investor activity and stock market activity: the narratives discussed and sentiments expressed at a given point in time are often shaped by real-time news and events.

First stage: We use variation in our contagion and consensus measures that can be explained by past activity on WSB to identify our parameters of interest. In doing so, we assume that our target stock market variables are sufficiently uncorrelated between sequential trading weeks.

Predicting $\Delta\Omega$ – The number of new authors posting submissions is predicted using the model from Section 3.4:

$$l(a_{j,t}) = \log(a_{j,t}/m_t) = c a_{j,t-1}(1 - a_{j,t-1}) + d a_{j,t-1} + \beta_1\bar{\phi}_{j,t-2}\bar{r}_{j,t-1} + \beta_2\bar{\phi}^2_{j,t-2}\sigma^2_{j,t-1} + \eta_{a,t} + \varepsilon_{a,j,t}, \tag{20}$$

with the key difference that we replace ticker fixed effects, in Eq. 7, with week fixed effects $\eta_{a,t}$. The results are used to predict the author count in the subsequent period, by using the previous week's author count in infrequent tickers, $M_{j,t-1}$:

$$\widehat{A}_{j,t} = M_{j,t-1}\exp\left(\widehat{l(a_{j,t})}\right), \quad \Delta\widehat{\Omega}_{j,t} = \frac{\phi_{j,t-1}}{q_{j,t-1}}\left(\widehat{A}_{j,t} - A_{j,t-1}\right), \quad \Delta\widehat{\Omega}^{*}_{j,t} = \frac{|\phi_{t-1}|}{q_{j,t-1}}\left(\widehat{A}_{j,t} - A_{j,t-1}\right), \tag{21}$$

where a hat denotes fitted values.

Predicting $\Delta\chi$ – In the same vein as in Section 3.2, we predict future sentiment using past stock price behaviour, as well as past sentiments:

$$\Phi^{+}_{j,t} = \log\left(\frac{P(\phi_{j,t}=+1)}{P(\phi_{j,t}=0)}\right) = \lambda^{+}_{r}\bar{r}_{j,t-1} + \lambda^{+}_{\sigma}\sigma^{2}_{j,t-1} + \lambda^{+}_{\Phi}\Phi^{+}_{j,t-1} + \eta^{+}_{t} + \varepsilon^{+}_{j,t}, \tag{22}$$

$$\Phi^{-}_{j,t} = \log\left(\frac{P(\phi_{j,t}=+1)}{P(\phi_{j,t}=0)}\right) = \lambda^{-}_{r}\bar{r}_{j,t-1} + \lambda^{-}_{\sigma}\sigma^{2}_{j,t-1} + \lambda^{-}_{\Phi}\Phi^{-}_{j,t-1} + \eta^{-}_{t} + \varepsilon^{-}_{j,t}, \tag{23}$$

where superscripts differentiate between the average log-odds of a submission in week t expressing bullish (+) versus negative (−) sentiments, over neutral sentiments. Week fixed effects remain in the sentiment models, so that the full estimation strategy rests on within-week variation in all explaining, as well as explained, variables.

The approach outlined above relies on coarser aggregates for sentiments: the probabilities here are not estimated on data for individual submission sentiments, as is the case in section 3.2. Rather, the probabilities are calculated by averaging the probabilities for *all* submissions in week t, discussing ticker j, to be bullish ($P(\phi_{j,t}=+1)$), bearish ($P(\phi_{j,t}=-1)$), or neutral ($P(\phi_{j,t}=0)$). Predicted values for our consensus measure follow from the sentiment model predictions:

$$\widehat{\phi}_{j,t} = \frac{\exp\left(\widehat{\Phi}^{+}_{j,t}\right) - \exp\left(\widehat{\Phi}^{-}_{j,t}\right)}{1 + \exp\left(\widehat{\Phi}^{+}_{j,t}\right) + \exp\left(\widehat{\Phi}^{-}_{j,t}\right)}, \tag{24}$$

$$\Delta\widehat{\chi}_{j,t} = \frac{A_{j,t-1}}{q_{j,t-1}}(\widehat{\phi}_{j,t} - \phi_{j,t-1}), \quad \Delta\widehat{\chi}^{*}_{j,t} = \frac{A_{j,t-1}}{q_{j,t-1}}(|\widehat{\phi}_{j,t}| - |\phi_{j,t-1}|). \tag{25}$$

In all our estimates, we restrict ourselves to a sub-sample spanning January 2016 to July 2017. As discussed in Section 3.4, this choice serves to limit the amount of missing data in times when activity was relatively sparse.

4.2.3 Results

Table 4 presents our main results. Panel B in Table 4 presents causal evidence for the impact of consensus formation and contagion among hype investors on stock market variables. We do not argue that WSB alone affects the markets, but rather that WSB data offers a rich sample of retail investor behaviour. The observed behaviours and dynamics within WSB, in turn, allow us to estimate future stock market activity.

Our predicted contagion measure, $\Delta\widehat{\Omega}_{j,t}$, has a statistically significant impact on changes in weekly average returns. It's counterpart, with absolute values for sentiment, $\Delta\widehat{\Omega}^{*}_{j,t}$, is also a statistically significant predictor for volatility in weekly returns and associated trading volumes. Holding constant a stock's market cap and associated sentiments, one new user's choice to discuss the asset, given their exposure to previous conversations, increases the weekly average log return by an average

Table 3: First stage estimates for consensus and contagion in WSB

	Dependent variable:		
	$l(a_{j,t})$	$\Phi^{+}_{j,t}$	$\Phi^{-}_{j,t}$
$a_{j,t-1}(1-a_{j,t-1})$	89.75^{***} (10.33)		
$a_{j,t-1}$	-54.77^{***} (8.67)		
$\bar{\phi}_{j,t-2}\bar{r}_{j,t-1}$	2.09^{*} (1.26)		
$\bar{\phi}^{2}_{j,t-2}\sigma^{2}_{j,t-1}$	-14.24^{***} (2.94)		
$\bar{\phi}_{j,t-1}$	0.02 (0.03)		
$\bar{r}_{j,t-1}$		0.85 (0.62)	-2.05^{***} (0.72)
$\sigma^{2}_{j,t-1}$		-3.03^{***} (0.89)	-2.60^{***} (0.42)
$\Phi^{+}_{j,t-1}$		0.10^{***} (0.01)	
$\Phi^{-}_{j,t-1}$			0.13^{***} (0.01)
Week FE	Yes	Yes	Yes
Observations	4,470	6,668	6,668
Adjusted R^2	0.34	−0.01	0.00
F Statistic	477.39	24.23	44.31

Notes: column (1) of this table presents OLS estimates for the log ratio of the number of authors discussing stock j in week t, over the number discussing a stock mentioned fewer than 30 times. The dependent variable in column (2) is the average log-odds of a given submission in week t on stock j to express bullish over neutral sentiment, and in column (3) differs in using bearish over neutral sentiments, instead. Explanatory variables include: the lag in the share of authors discussing j, $a_{j,t-1}$, the interaction with the share of authors not discussing j, $a_{j,t-1}(1-a_{j,t-1})$, as well as the average log return multiplied by the two period lag in the average sentiment expressed among WSB submissions on stock j, $\bar{\phi}_{j,t-2}\bar{r}_{j,t-1}$, and the variance in log returns by the same sentiment's square, $\bar{\phi}^{2}_{j,t-2}\sigma^{2}_{j,t-1}$. The one period lag in sentiments is also included in this specification. The logit-transformed sentiments are regressed on the lag of weekly mean, and variance, in log returns, as well as the lag in logit-transformed sentiments. Each specification includes week-specific fixed effects. Accompanying standard errors, displayed in brackets, are clustered at the stock level, and calculated in the manner of MacKinnon & White (1985).

*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

of 1 percentage point (pp), and the variance by 0.5pp. The large coefficient on our contagion measure when predicting volumes is likely a result of the fact that user numbers are a good proxy for the attention a stock garners among the greater retail investor community. The contagion measure, like it's consensus counterpart, accounts for the asset's market cap, so the impact of contagion is attenuated by the stock's market depth.

The fitted consensus measure, $\Delta\widehat{\chi}_{j,t}$, displays a strong, yet noisy, correlation with changes in average weekly returns in column (1). Sentiments gravitating from 0.5 to 1 on average increase a stock's weekly return by 10 percentage points (pp), controlling for the previous week's discussion size and the stock's market cap. However, this estimate is only significant at the 10% level. This contagion measure using absolute measures for sentiment, $\Delta\widehat{\chi}^{*}_{j,t}$, is also a strong and more statistically significant predictor for increases in weekly levels of return variance, in column (2), and the percent change in trading volumes, in column (3).

We also consider the reduced-form estimates for the OLS coefficients from Eqs. 17-19, presented in Panel A. For these estimates, we use the observed, rather than predicted, measures of $\Delta\Omega_{j,t}$, $\Delta\chi_{j,t}$.

Table 4: Market impact of consensus and contagion in WSB

Panel A: reduced form relationship between WSB and market activity

	Dependent variable:		
	$\Delta\bar{r}_{j,t}$	$\Delta\sigma^2_{j,t}$	$\Delta v_{j,t}$
	(1)	(2)	(3)
$\Delta\chi_{j,t}$	0.07 (0.24)		
$\Delta\Omega_{j,t}$	−0.10 (0.10)		
$\Delta\chi^*_{j,t}$		0.12 (0.14)	7.95 (9.97)
$\Delta\Omega^*_{j,t}$		0.13^{***} (0.04)	11.69^{**} (5.79)
Week FE	Yes	Yes	Yes
Observations	6,667	6,641	6,666
Adjusted R^2	−0.03	−0.01	0.03
F-Statistic	34.46	96.79	239.34

Panel B: structural relationship between WSB and market activity

	(1)	(2)	(3)
$\Delta\widehat{\chi}_{j,t}$	0.21^{*} (0.11)		
$\Delta\widehat{\Omega}_{j,t}$	0.01^{***} (0.002)		
$\Delta\widehat{\chi}^*_{j,t}$		0.25^{***} (0.08)	18.41^{***} (6.71)
$\Delta\widehat{\Omega}^*_{j,t}$		0.005^{***} (0.002)	0.38^{***} (0.14)
Week FE	Yes	Yes	Yes
Observations	4,458	4,452	4,458
Adjusted R^2	0.04	−0.02	0.13
F-Statistic	210.48	79.13	460.10
J-statistic	7.547	3.246	4.792

Notes: this table presents OLS estimates for stock j's change in average log return, $\Delta\bar{r}_{j,t}$, change in variance of log returns, $\Delta\sigma_{j,t}$, and percent change in nominal trading volume, $\Delta v_{j,t}$, in week t. Stocks consider are mentioned more in at least 30 distinct submissions on WSB. Explanatory variables include a measure for consensus formation, $\Delta\chi_{j,t}$, which tracks the change in sentiments on WSB, fixing the number of users discussing stock j, and contagion, $\Delta\widehat{\Omega}_{j,t}$, which counts the number of new users discussing stock j, fixing their sentiment. Both variables are also divided by the market cap of stock j. In computing OLS coefficients for volatility and percent changes in trading volumes, both independent variables take the absolute value in sentiments, and are denoted by an asterisk. Each specification includes week-specific fixed effects. Accompanying standard errors, displayed in brackets, are clustered at the stock level, and calculated in the manner of MacKinnon & White (1985). Panel A computes the coefficients using values directly from WSB data, whereas Panel B uses consensus and contagion as predicted by past sentiments and stock discussions as discussed in the text. The relevant first-stage results are presented in Table 3, and the associated J-statistics are recorded at the bottom of Panel B.
*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

Panel A, therefore, gives us a sense for how consensus and contagion co-vary with our market variables. The results in column (1) are consistent with Appendix B.2, where we observe a relationship between daily log-returns and sentiment, but not weekly, average returns. Column (1) offers no indication that changes in sentiments, measured through $\Delta\chi_{j,t}$, or number of discussants, measured through $\Delta\Omega_{j,t}$, occur in weeks with increasing returns. Columns (2) and (3) demonstrate that the number of discussants is higher in weeks with elevated levels of volatility and trading volumes.

Table 3 helps assess the instruments' strength in predicting contagion and consensus on WSB. In column (1), lags in author shares discussing stock j, plus its lagged returns and volatility, are used to predict the share of authors discussing the same stock. This setup is identical to column (4) of Table 2, except in the use of week fixed effects instead of stock fixed effects. The high F-statistic, as well as adjusted R^2, confirm that these instruments are far from weak. In columns (2) and (3), we find that lagged weekly mean and variance in returns, combined with lagged sentiments, are significant predictors for the current log-odds in submissions expressing bullish and bearish, over neutral, sentiments. However, associated R^2 measures indicate that these account for a tiny amount in the variance of weekly average sentiments, which is likely the reason behind the noisy parameter estimates in Panel B of Table 4. The validity of all these lagged variables as instruments for consensus and contagion is checked by the J-statistic, which we report at the bottom of Table 4. Their low score demonstrate insufficient evidence to reject the hypothesis that the instruments are jointly insignificant in explaining the residuals for each target stock market variable. This compensates for the identifying assumption that stock market returns and trading volumes are largely uncorrelated from one week to the next.

Clearly, social dynamics among retail investors play an important role in the stock market today. We successfully pin down two channels by which they exert influence; consensus, by which retail investors coordinate a buying or selling strategy as a group, and contagion, by which rare information on opportunities can be communicate to growing crowds. To the disappointment of this paper's title, pinning down any instability onto WSB as a platform alone is impossible. At the very least, the forum offers a window into the minds of an expanding number of retail investors, and it is clear that social media play a key role in their growing clout on the virtual trading floor.

5 Conclusion

WSB data helps us document two empirical findings: i) 'consensus' and ii) 'contagion' among retail investors' trading strategies. We test for consensus formation in two different ways. Our results consistently imply that a causal link exists between the sentiment that an individual investor adopts about a stock and that of his peers. Individuals also tend to focus their attention on stocks that others are discussing, underscoring the existence of contagion in asset interest among investors.

Is it possible to draw some conclusion about how asset prices change with respect to these dynamics? To that end, this paper introduces a model that decomposes investor asset demand due to social dynamics into i) hype around an asset caused by *contagion* and ii) opinion formation leading to overall *consensus* to buy or sell. The model indicates that, when investors are not connected and consensus is difficult to reach, prices remains stable. However, as feedback among investors plays a greater role, the price impact becomes larger. This is verified through simulation. We empirically validate the impact of social dynamics, proxied by WSB data, and market activity using an IV approach. Our predicted 'consensus' measure has a statistically significant correlation with market returns, volatility and trading volumes. Sentiments gravitating from neutral ($\phi = 0$) to bullish ($\phi = +1$) among hype investors on average increases a stock's weekly returns by 20pp. We do not argue that WSB investors alone drive these returns, but rather that the WSB forum allows us to sample broader retail investor sentiment. An absolute change in sentiments by one, for example from neutral ($\phi = 0$) to bearish ($\phi = -1$), increases a stock's volatility by 0.25 and trading volumes by 18.4%. A change in sentiments appears to cause greater uncertainty in an asset's value, echoing the proposed framework of Barlevy & Veronesi (2000). Our prediction for contagion similarly has a strong correlation with all three market variables.

These findings are of particular importance to how we view the efficiency of financial markets and investor rationality. Even though the debate is not over and this work is far from conclusive, we believe our research present evidence in favor of narrative economics and irrational exuberance (Shiller 1984, 2005). The important role that peers and asset returns play in formulating investor opinion points to strategic complementarities in information acquisition among investors, underscoring the importance of work by Barlevy & Veronesi (2000) and Hellwig & Veldkamp (2009). Our findings complement the earlier experiments of Bursztyn et al. (2014), showing that peer influence plays a role in a broader investment context than that of the author's experimental setting.

Social media has changed the fabric of society. 4.2 billion people, or 53.6% of the world population, are active social media users, each just a few clicks away from the next popular phenomenon.[8] Now, a growing audience turns to social media for promising stock market gambles. Whether social media has been a boon or a bane to society is a highly contested topic, however, there is little debate over the fact that increased social media usage has generated new and exciting datasets for research. The rich text data from WSB can be used to verify economic behavioural theories: from narrative economics Shiller (1984), to models of 'noise trader' behaviour (Shefrin & Statman 1994, De Bondt & Thaler 1985, Black 1986) and information diffusion among investors (Banerjee et al. 2013, Hellwig & Veldkamp 2009, Barlevy & Veronesi 2000). It also presents new opportunities to leverage techniques in other areas of economics for understanding investor behaviour, such as the well-established peer effects literature (Duflo et al. 2011, Epple & Romano 2011, Sacerdote 2011, Angrist 2014, Bramoullé et al. 2020).

Evaluating economic theory is of great importance of its own right, however, we must also remember that the financial markets do not exist in isolation: investor decisions have broader implications on the cost of capital. A growing body of research considers how social media impacts society, with polarization, the spread of fake news and other societal challenges being some of its documented consequences (Tucker et al. 2018). In light of our findings, a relevant question may be how these consequences may impact capital allocation. Does the government have a right to track and penalize the spread of misinformation about asset prices on social media, despite its mandate to defend free speech? Do we need to think carefully about the power of social media personalities, given their potential destabilize the financial system?[9] Perhaps it is an important time to re-evaluate regulatory structure within the financial system, which closely monitors financial institutions and large players, but overlooks smaller investors.

As social media galvanizes a larger pool of retail investors with the potential for exciting stock market gambles, it is crucial to understand how social dynamics can impact asset prices. With the first publicly acclaimed victory of Main Street over Wall Street, in the form of the GameStop short squeeze, it is unlikely that socially-driven asset volatility will simply disappear. In fact we observe the opposite: WSB grew from approximately 1.8 million users at the start of January 2021 to over eleven million users in February 2022.[10] At this junction, it is perhaps more important than ever to consider our findings in the broader economic context, and ask what insights economic theory can offer to ensure financial stability and prosperity at a time when social media is rapidly changing the investment landscape.

Notes

[1] `https://www.alexa.com/topsites`

[2] `https://redditblog.com/2019/12/04/reddits-2019-year-in-review/`

[3] `https://subredditstats.com/r/wallstreetbets`

[4] `https://andriymulyar.com/blog/how-a-subreddit-made-millions-from-covid19`

[5] `https://www.reddit.com/r/wallstreetbets/`

[6] `https://pushshift.io/`

[7] This assumption is standard and justifiable in our present context, since investors make a choice governed by some maximisation process. The type-I Extreme Value distribution is one of three possible limiting distributions for the maximum of random variables.

[8] `https://datareportal.com/reports/digital-2021-global-overview-report`

[9] `https://www.theguardian.com/technology/2022/jan/14/dogecoin-value-soars-after-elon-musk-says-it-will-be-accep`

[10] `https://subredditstats.com/r/wallstreetbets`

References

Anderson, C. A., Lepper, M. R. & Ross, L. (1980), 'Perseverance of social theories: The role of explanation in the persistence of discredited information.', *Journal of personality and social psychology* **39**(6), 1037.

Angrist, J. D. (2014), 'The perils of peer effects', *Labour Economics* **30**, 98–108.

Araci, D. (2019), 'Finbert: Financial sentiment analysis with pre-trained language models', *arXiv preprint arXiv:1908.10063* .

Aral, S., Muchnik, L. & Sundararajan, A. (2009), 'Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks', *Proceedings of the National Academy of Sciences* **106**(51), 21544–21549.

Aral, S. & Nicolaides, C. (2017), 'Exercise contagion in a global social network', *Nature communications* **8**(1), 1–8.

Avery, C. & Zemsky, P. (1998), 'Multidimensional uncertainty and herd behavior in financial markets', *American economic review* pp. 724–748.

Bailey, M., Cao, R., Kuchler, T. & Stroebel, J. (2016), Social networks and housing markets, Technical report, National Bureau of Economic Research.

Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. (2013), 'The diffusion of microfinance', *Science* **341**(6144).

Banerjee, A. V. (1993), 'The economics of rumours', *The Review of Economic Studies* **60**(2), 309–327.

Barberis, N. C. (2013), 'Thirty years of prospect theory in economics: A review and assessment', *Journal of Economic Perspectives* **27**(1), 173–96.

Barlevy, G. & Veronesi, P. (2000), 'Information acquisition in financial markets', *The Review of Economic Studies* **67**(1), 79–90.

Bifulco, R., Fletcher, J. M. & Ross, S. L. (2011), 'The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health', *American Economic Journal: Economic Policy* **3**(1), 25–53.

Black, F. (1986), 'Noise', *The journal of finance* **41**(3), 528–543.

Blume, L. E., Brock, W. A., Durlauf, S. N. & Ioannides, Y. M. (2011), Identification of social interactions, *in* 'Handbook of social economics', Vol. 1, Elsevier, pp. 853–964.

Bollen, J., Mao, H. & Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science* **2**(1), 1 – 8.

Bordalo, P., Gennaioli, N. & Shleifer, A. (2012), 'Salience theory of choice under risk', *The Quarterly journal of economics* **127**(3), 1243–1285.

Bordalo, P., Gennaioli, N. & Shleifer, A. (2019), 'Diagnostic expectations and stock returns', *The Journal of Finance* **74**(6), 2839–2874.

Bordalo, P., Gennaioli, N. & Shleifer, A. (2020), 'Memory, Attention, and Choice*', *The Quarterly Journal of Economics* **135**(3), 1399–1442.

Bouchaud, J.-P. (2013), 'Crises and collective socio-economic phenomena: simple models and challenges', *Journal of Statistical Physics* **151**(3), 567–606.

Bramoullé, Y., Djebbari, H. & Fortin, B. (2020), 'Peer effects in networks: A survey', *Annual Review of Economics* **12**, 603–629.

Breza, E. & Chandrasekhar, A. G. (2019), 'Social networks, reputation, and commitment: evidence from a savings monitors experiment', *Econometrica* **87**(1), 175–216.

Brock, W. A. & Durlauf, S. N. (2001), 'Discrete choice with social interactions', *The Review of Economic Studies* **68**(2), 235–260.

Bulow, J. I., Geanakoplos, J. D. & Klemperer, P. D. (1985), 'Multimarket oligopoly: Strategic substitutes and complements', *Journal of Political economy* **93**(3), 488–511.

Bursztyn, L., Ederer, F., Ferman, B. & Yuchtman, N. (2014), 'Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions.', *Econometrica* **82**(4), 1273.

Chohan, U. W. (2021), 'Counter-hegemonic finance: The Gamestop short squeeze', *Available at SSRN* .

Cont, R. & Bouchaud, J.-P. (2000), 'Herd behavior and aggregate fluctuations in financial markets', *Macroeconomic Dynamics* **4**(2), 170–196.

De Bondt, W. F. & Thaler, R. (1985), 'Does the stock market overreact?', *The Journal of finance* **40**(3), 793–805.

De Long, J. B., Shleifer, A., Summers, L. H. & Waldmann, R. J. (1990), 'Noise trader risk in financial markets', *Journal of political Economy* **98**(4), 703–738.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Duflo, E., Dupas, P. & Kremer, M. (2011), 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *American Economic Review* **101**(5), 1739–74.

Epple, D. & Romano, R. E. (2011), Peer effects in education: A survey of the theory and evidence, *in* ‘Handbook of social economics’, Vol. 1, Elsevier, pp. 1053–1163.

Garcia, D. (2013), ‘Sentiment during recessions’, *The Journal of Finance* **68**(3), 1267–1300.

Gennaioli, N. & Shleifer, A. (2010), ‘What comes to mind’, *The Quarterly journal of economics* **125**(4), 1399–1433.

Gennaioli, N. & Shleifer, A. (2018), *A crisis of beliefs*, Princeton University Press.

Gentzkow, M., Kelly, B. & Taddy, M. (2019), ‘Text as data’, *Journal of Economic Literature* **57**(3), 535–74.

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

Hellwig, C. & Veldkamp, L. (2009), ‘Knowing what others know: Coordination motives in information acquisition’, *The Review of Economic Studies* **76**(1), 223–251.

Hirshleifer, D. (2001), ‘Investor psychology and asset pricing’, *The journal of Finance* **56**(4), 1533–1597.

Hommes, C. (2013), *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems*, Cambridge University Press.

Hommes, C. (2021), ‘Behavioral and experimental macroeconomics and policy analysis: A complex systems approach’, *Journal of Economic Literature* **59**(1), 149–219.

Kahneman, D. & Tversky, A. (1972), ‘Subjective probability: A judgment of representativeness’, *Cognitive psychology* **3**(3), 430–454.

Kahneman, D. & Tversky, A. (1973), ‘On the psychology of prediction.’, *Psychological review* **80**(4), 237.

Kahneman, D. & Tversky, A. (1979), ‘Prospect theory: An analysis of decision under risk’, *Econometrica* **47**(2), 263–291.

Kirman, A. (1993), ‘Ants, rationality, and recruitment’, *The Quarterly Journal of Economics* **108**(1), 137–156.

Kumar, A. & Lee, C. M. (2006), ‘Retail investor sentiment and return comovements’, *The Journal of Finance* **61**(5), 2451–2486.

Lahno, A. M. & Serra-Garcia, M. (2015), ‘Peer effects in risk taking: Envy or conformity?’, *Journal of Risk and Uncertainty* **50**(1), 73–95.

MacKinnon, J. G. & White, H. (1985), ‘Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties’, *Journal of Econometrics* **29**(3), 305–325.

Patacchini, E. & Zenou, Y. (2016), ‘Social networks and parental behavior in the intergenerational transmission of religion’, *Quantitative Economics* **7**(3), 969–995.

Ross, L., Lepper, M. R. & Hubbard, M. (1975), 'Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm.', *Journal of personality and social psychology* **32**(5), 880.

Sabherwal, S., Sarkar, S. K. & Zhang, Y. (2011), 'Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news', *Journal of Business Finance & Accounting* **38**(9-10), 1209–1237.

Sacerdote, B. (2011), Peer effects in education: How might they work, how big are they and how much do we know thus far?, *in* 'Handbook of the Economics of Education', Vol. 3, Elsevier, pp. 249–277.

Securities & Commission, E. (1963), 'Report of special study of securities markets of the securities and exchange commission', *House Document* **95**(4).

Shefrin, H. & Statman, M. (1994), 'Behavioral capital asset pricing theory', *Journal of financial and quantitative analysis* **29**(3), 323–349.

Shiller, R. J. (1984), Stock prices and social dynamics, Technical Report 2, The Brookings Institution.

Shiller, R. J. (2005), *Irrational Exuberance: (Second Edition)*, Princeton University Press.

Shiller, R. J. (2014), 'Speculative asset prices', *American Economic Review* **104**(6), 1486–1517.

Shiller, R. J. (2017), 'Narrative economics', *American Economic Review* **107**(4), 967–1004.

Shleifer, A. & Summers, L. H. (1990), 'The noise trader approach to finance', *Journal of Economic perspectives* **4**(2), 19–33.

Tetlock, P. C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of finance* **62**(3), 1139–1168.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. & Nyhan, B. (2018), 'Social media, political polarization, and political disinformation: A review of the scientific literature', *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* .

Vasileiou, E., Bartzou, E. & Tzanakis, P. (2021), 'Explaining Gamestop short squeeze using intraday data and Google searches.', *Available at SSRN 3805630* .

Wang, Y.-H., Keswani, A. & Taylor, S. J. (2006), 'The relationships between sentiment, returns and volatility', *International Journal of Forecasting* **22**(1), 109–123.

Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013), A biterm topic model for short texts, *in* 'Proceedings of the 22nd International Conference on World Wide Web', WWW '13, Association for Computing Machinery, New York, NY, USA, p. 1445–1456.

Zenou, Y. (2016), 'Key players', *The Oxford Handbook of the Economics of Networks* **11**.

A Data appendix

A.1 Tickers Mentioned on WSB

Table 5: Most Frequent Ticker Mentions

Ticker	Name	Comments	Submissions	Sum
SPY	S&P 500 Index	293,178	9,507	302,685
TSLA	Tesla, Inc.	126,974	6,063	133,037
AMD	Advanced Micro Devices, Inc.	124,722	5,727	130,449
MU	Micron Technology, Inc.	86,658	3,944	90,602
AAPL	Apple Inc.	48,768	1,888	50,656
AMZN	Amazon.com, Inc.	44,748	1,552	46,300
MSFT	Microsoft Corporation	41,421	1,810	43,231
SNAP	Snap Inc.	40,768	2,045	42,813
NVDA	NVIDIA Corporation	38,189	1,561	39,750
SPCE	Virgin Galactic Holdings, Inc.	30,777	1,648	32,425
FB	Facebook, Inc.	26,334	1,455	27,789
DIS	The Walt Disney Company	25,643	1,090	26,733
BYND	Beyond Meat, Inc.	23,306	908	24,214
NFLX	Netflix, Inc.	20,868	939	21,807
JNUG	Direxion Daily Jr Gld Mnrs Bull 3X ETF	15,761	1,095	16,856
GE	General Electric Company	15,733	931	16,664
RAD	Rite Aid Corporation	14,801	841	15,642
SQ	Square, Inc.	14,061	826	14,887
ATVI	Activision Blizzard, Inc.	13,094	675	13,769
USO	United States Oil	12,952	669	13,621

Notes: this table lists the 20 most mentioned assets on WSB, observed by submissions which uniquely mention the related ticker. 'Comments' is the number of comments posted on these submissions, and 'Submissions' counts submissions, and 'Total' is the sum of the two. The name of the asset corresponding to the identified ticker is retrieved from *Yahoo Finance*.

Conventionally, submissions or comments that mention a ticker will spell it using uppercase letters, or following a dollar sign. However, a challenge is that not all uppercase words are valid tickers.

We first match words in WSB submissions to assets by identifying any succession of two to five capital letters. Subsequently, we used a pre-determined list of tickers from CRSP to check whether a match is indeed present in the available financial data. Some abbreviations or capitalised words which are not valid tickers might still show up, such as 'USD' (*ProShares Ultra Semiconductors*), 'CEO' (*CNOOC Limited*), and 'ALL' (*The Allstate Corporation*). Single characters also appear, such as 'A' (*Agilent Technologies, Inc.*). We manually created a list of such tickers, and removed matches featured in WSB submissions, to build a preliminary list of candidate ticker mentions. We refined a second list of candidates by checking whether a collection of one to five letters, lower or uppercase, is preceded by a dollar sign. Any mentions of '$CEO' or '$a' count as the tickers 'CEO' and 'A', respectively. We checked these extracts again against the list of available tickers.

A small fraction of the 4,650 tickers we extract dominate the discourse on WSB. 90% of tickers

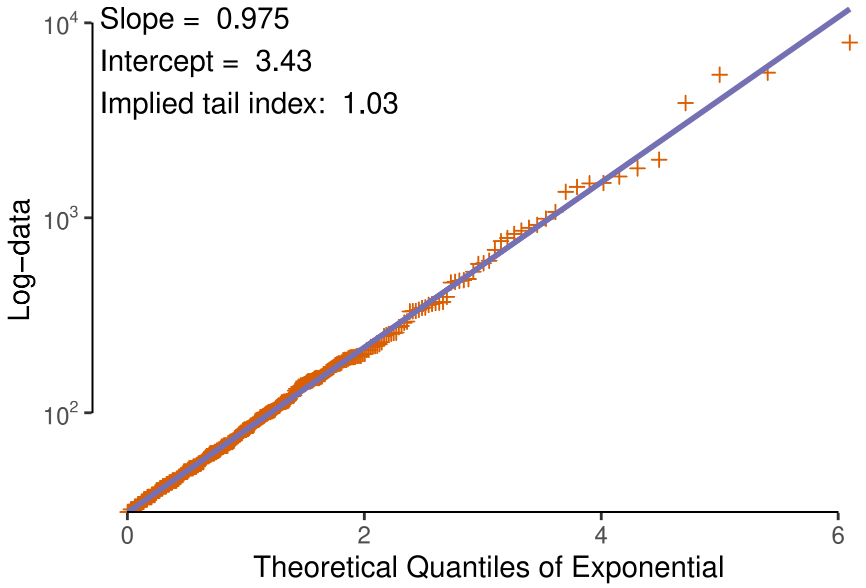


Figure 5: **QQ Plot of the Tail in Ticker Mentions on WSB**; the number of submissions for each ticker (on a log-scale) is plotted against the theoretical quantiles of an exponential distribution. Quantiles are calculated as $q(i) = -\log(1 - i/(N + 1)$, where N is the number of observations, and i the order of the statistic, from 1 to N. The linear fit suggests that the data follows a Pareto distribution, with the tail index equal to the inverse of the slope. The threshold for a ticker to be part of the 'tail' is 31 mentions; note the intercept, at $\exp(3.43) \approx 31$.

are mentioned fewer than 31 times, and more than 60% are mentioned fewer than five times. The frequency distribution of tail of ticker mentions demonstrates this point, for which Figure 5 displays a QQ-plot. We arbitrarily selected tickers with the number of mentions in the top 10^{th} percentile. Even though threshold of mentions for this top decile is 30 submissions, the most popular, SPY, features in almost 8,000 submissions. The orange crosses in Figure 5 locate the empirical densities, on a log scale, which are plotted against the theoretical quantiles of an exponential distribution on the x-axis. Under the assumption that ticker mentions are heavy-tailed (similarly to vocabulary distributions), the logarithm of the mentions follows an exponential distribution, with the intercept at the threshold, and the slope equal to the inverse of the tail index. Indeed, the linear fit in Figure 5 is close to perfect, supporting the assumption that the popularity of assets in WSB is heavy-tailed, with an estimated tail exponent of approximately 1.03. In what follows, we used submissions for which we identified a single ticker, unless otherwise specified, forming a dataset of 103,205 submissions with unique ticker mentions by our cutoff date.

A.2 Sentiment Modeling in WSB Posts

In order to thoroughly understand the social dynamics of asset discussions, it is not sufficient to simply identify what assets are being discussed; it is important to understand what is being said about them. Our goal, with regards to the text data in WSB, is to gauge whether discussions on certain assets express an expectation for their future price to rise, the 'bullish' case, to fall, the 'bearish' case, or to remain unpredictable, the 'neutral' case.

A series of studies link sentiment, measured through diverse approaches, to stock market performance (Garcia 2013, Tetlock 2007, Bollen et al. 2011). Gentzkow et al. (2019) offer a thorough

		Predicted Label		
		-	**0**	+
True Label	-	68%	17%	15%
	0	11%	63%	26%
	+	4%	11%	84%

Table 6: **Fine-tuned FinBERT Confusion Matrix:** We use 10% of our hand-labeled data to test the performance of FinBERT on out-of-sample sentiment prediction. The results highlight the model's ability to predict sentiment with reasonably high accuracy.

review. Many of these works use lexicon approaches, whereby documents are scored based on the prevalence of words associated with a certain sentiment. Recently, machine learning has offered alternative, powerful tools, such as Google's Bidirectional Encoder Representations from Transformers (BERT) algorithm (Devlin et al. 2018). The BERT algorithm trains a final layer of nodes in a neural network from a pre-trained classifier on labelled data. The classifier itself is a neural net, pre-trained by Google on a corpus of Wikipedia entries to i) predict the probability distribution of words appearing in a given sentence (Masked Language Modeling), and ii) predict the relationship between sentences (Next Sentence Prediction). BERT has further been modified through pre-training on a financial text corpus to produce FinBERT (Araci 2019).

Among other alternatives, we pursued a supervised-learning approach to identify the sentiment expressed about an asset within a WSB submission. This required a training dataset, for which we manually labelled 4,932 random submissions with unique ticker mentions as either 'bullish', 'bearish' or 'neutral' with respect to the authors' expressed expectations for the future price. We used the FinBERT algorithm for labeling. Work not shown here implements an alternative regression-based approach as a robustness check, but FinBERT performs better out-of-sample.

We trained FinBERT on 75% of the labelled data, and used the remaining 25% for validation and the test set. Table 6 plots the out-of-sample confusion matrix. For the out-of-sample test, we train FinBERT on 75% of the available data and use 15% for validation; we then compute what the algorithm predicts for the remaining 10% of data. We achieve 73% accuracy on the test set. It is worthwhile to note that, potentially due to class imbalance in the underlying posts (where WSB generally appears to have more bullish and neutral posts than bearish), our classifier is better at identifying bullish posts. Neutral appears to be the most difficult category for classification, potentially because of the presence of important keywords in the bearish and bullish cases, which are absent in the neutral case. This is better than a LASSO regression's accuracy, which was implemented separately and is not cover here.

B Details on testing for consensus and contagion

B.1 Target independent variable

We extend the utility framework to suit our empirical strategy. Under the assumption that v_i is drawn from a standard type-I Extreme Value Distribution, we model the log-odds of investor sentiments ϕ_i

by a standard multivariate logistic function,

$$\log\left(\frac{P(\phi_{i,t}=+1)}{P(\phi_{i,t}=0)}\right)=g(b_{i,t})+f(\bar{\phi}_{-i,(t-1,t)})-\theta\sigma^2_{i,t}+u^+_{i,t}, \tag{26}$$

$$\log\left(\frac{P(\phi_{i,t}=-1)}{P(\phi_{i,t}=0)}\right)=-g(b_{i,t})-f(\bar{\phi}_{-i,(t-1,t)})-\theta\sigma^2_{i,t}+u^-_{i,t}, \tag{27}$$

where t denotes time, and $(t-1,t)$ an interval preceding t. The goal of this paper, in light of Proposition 1, is to test empirically whether $f(\cdot)$ is increasing. To that end, we aggregate bullish and bearish sentiments into one continuous variable, $\Phi_{i,t}$, leaving out the neutral benchmark:

$$\Phi_{i,t}=\frac{1}{2}\log\left(\frac{P(\phi_{i,t}=+1)}{P(\phi_{i,t}=-1)}\right)=g(b_{i,t})+f(\bar{\phi}_{-i,(t-1,t)})+\frac{u^+_{i,t}-u^-_{i,t}}{2}. \tag{28}$$

In the main body, the error term is expressed as $\epsilon_{i,t}$. Under the assumption that $u^+_{i,t}$ and $u^-_{i,t}$ are independent identically distribution, $u^+_{i,t}-u^-_{i,t}$ will follow a logistic distribution with finite variance. The intensity of neutral sentiments are an interesting manifestation of uncertainty in WSB, but ultimately not insightful in measuring the degree of social contagion in $f(\cdot)$.

B.2 Consensus formation

This section details the variables used in Section 3, and provides additional results in support of the reported findings.

B.2.1 Variable creation

$\phi_{i,j,t}$ – We begin with the output of our sentiment classifier, detailed in Appendix A.2. It assigns three probability scores to each submission about a ticker: the probability of a submission being bullish, $P(\phi=+1)$, bearish, $P(\phi=-1)$, neutral, $P(\phi=0)$. The probabilities sum to one. At the time t when an author i posts about asset j, we use the probability scores above to calculate a continuous sentiment score between $(-\infty,\infty)$:

$$\Phi_{i,j,t}=\frac{1}{2}\log\left(\frac{P(\phi_{i,j,t}=+1)}{P(\phi_{i,j,t}=-1)}\right).$$

Submissions labeled as bullish ($P(\phi=+1)=1$), or bearish ($P(\phi=+1)=1$), are set to $P(\phi=+1)=0.98$, or $P(\phi=-1)=0.98$, to retrieve a finite value for the log-odds.

Control variables We include a set of market return and volatility control variables. The data source for these variables are the daily stock files issued by the Center for Research in Security Prices (CRSP), accessed through Wharton Research Data Services.

$r_{j,t}$: the log return for asset j at time t. From CRSP, we calculate it using their 'RET' variable: $r_{j,t}=\log(RET_{j,t}-1)$, which automatically corrects the percentage change in closing prices for share splits and dividend distributions.

$\bar{r}_{j,t}$: the average log returns for asset j in the five days prior to t (the log return on day t is not included). A minimum of three daily log-return observations is required, otherwise the observation is set as missing.

$\sigma^2_{j,t}$: the variance of log returns for asset j in the five days prior to t (the log return on day t is not included). A minimum of three daily log-return observations is required, otherwise the observation is set as missing.

Matching submission timings to trade days If a post occurs before 16:00:00 EST on day t, we match it with the log-return on the same day t. If a post occurs after 16:00:00 EST on a given day, we match it with the next trading day, $t+1$, for the price variables. This is done to capture the fact that many news announcements occur after hours and someone posting after the market close may be exposed to these after-hour moves. Instance in which submissions are made on weekends, or holidays, are matched to the next possible trading day. For example, a submission made at 5pm on Friday is paired to the observed log return for the following Monday.

B.2.2 Frequent Posters Approach – Further Results and Details on Estimation Strategy

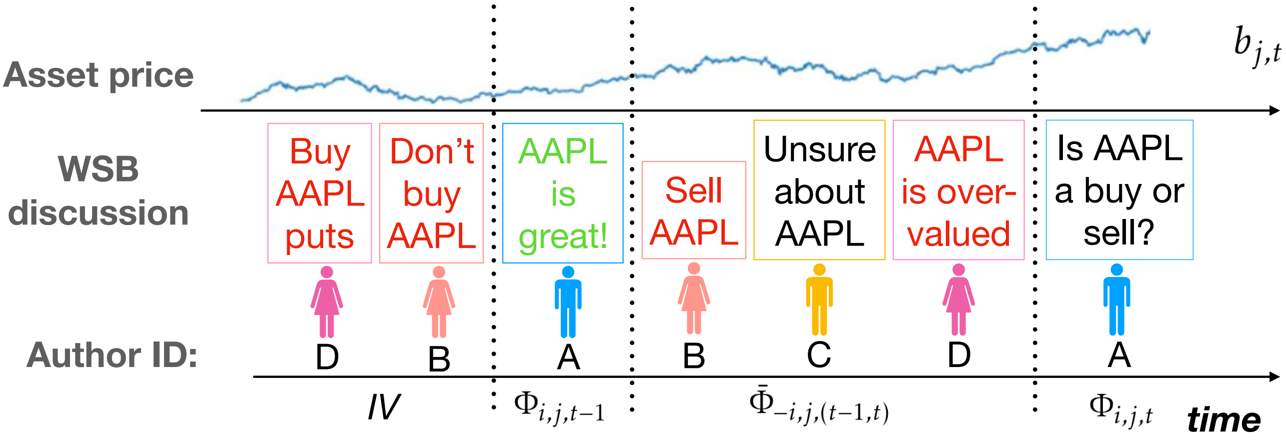


Figure 6: **Data on Investor Influence**; We consider how investor **A** is influenced by asset market movements and the sentiment of others. We use the past sentiments of author **A**'s peers (authors **B** and **D**) as an instrumental variable to predict the future sentiment of peers.

In the *Frequent Posters* approach, we attempt to find the influence that peers have on an individual who posts multiple times about the same asset. We estimate this through the coefficient on $\bar{\Phi}_{-i,j,(t-1,t)}$ in the following equation:

$$\Phi_{i,j,t} = \kappa\bar{\Phi}_{-i,j,(t-1,t)} + X_{i,j}\beta_{i,j} + \epsilon_{i,j}.$$

We set of controls, $X_{i,j}$, include the author's own previous sentiment, market movements and ticker fixed effects.

Table 7, column (1), presents the full result of estimating peer effects using a regression approach. We control for the asset's returns $(r_{j,t}, \bar{r}_{j,t})$, volatility $(\sigma^2_{j,t})$ and the author's own previous sentiment $(\Phi_{i,j,t-1})$.

One challenge to accurately estimating the coefficient on $\bar{\Phi}_{-i,j,(t-1,t)}$ is the potential for both the author and his peers to experience an exogenous shock, such as a news announcement, between an author's posts. This would incorrectly be attributed to peer effects in a naive estimation of the model. For this reason, we employ an IV approach: we approximate the sentiments of peers using the peer's previously expressed sentiment about the asset.

Figure 6 illustrates our approach. We attempt to estimate the influence of peers **B**, **C**, **D** on author **A**. We observe that peers **B**, **D** post about the same asset prior to **A**'s initial post. We, therefore, estimate **B** and **D**'s view about the asset using their historical view (highlighted as *IV* in Figure 6). Any exogenous shocks that occur between *(t-1,t)* would not affect the historic views of **B**, **D**. This method, therefore, allows us to eliminate confounding variables. Market moves, labeled $b_{j,t}$ and ticker fixed effects are also included as controls.

The full estimates from our second stage regression are presented in Table 7. Both the reduced form estimates and the complete second stage support Corollary 1.1 and Proposition 2.

Table 7: Peer Influence: Multiple Posters – Full Regression Estimates

		Dependent Variable – $\Phi_{i,j,t}$		
		Reduced Form	Full Second Stage	Random Peers
		(1)	(2)	(3)
Independent Variables	$\Phi_{i,j,t-1}$	0.16 (0.01) ***	0.13 (0.01) ***	0.16 (0.01) ***
	$\bar{\Phi}_{-i,j,(t-1,t)}$	0.11 (0.02) ***	0.16 (0.06) ***	-0.01 (0.02)
	$r_{j,t}$	1.01 (0.15) ***	1.17 (0.22) ***	1.05 (0.16) ***
	$\bar{r}_{j,t}$	0.62 (0.48)	0.80 (0.56)	0.74 (0.49)
	$\sigma^2_{j,t}$	-0.52 (0.37)	0.42 (1.47)	-0.01 (0.68)
	Ticker Fixed Effects	Yes	Yes	Yes
	No. Observations:	11,366	8,558	11,357
	R^2:	0.11	0.07	0.11
	R^2_{adj}:	0.07	0.05	0.07

Notes: The dependent variable is individual investor sentiment about an asset, scaled continuously between $(-\infty, \infty)$. We estimate it using the individual's previously expressed sentiment about the same asset ($\Phi_{i,j,t-1}$) and a set of market control variables ($r_{j,t}, \bar{r}_{j,t}, \sigma^2_{j,t}$). The sentiment of peers ($\bar{\Phi}_{-i,j,(t-1,t)}$) is estimated in several ways. In Column (1), we use observed, average sentiment of peers between an author's two posts. In Column (2), we estimate the sentiment of peers using an IV. In Column (3), we select a random cohort to estimate peer sentiment. Robust standard errors, clustered at the ticker level, are presented in parentheses. Observations with incomplete market data are dropped.

*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

B.2.3 Network Approach – Further Results

In this section, we present the results from our full second stage and reduced form regressions, from our network analysis.

Table 8 presents the reduced form for our network regression, as well as the full second stage estimates. Both the reduced form estimates and the complete second stage support Corollary 1.1 and Proposition 2.

B.2.4 Evidence of Identification Strategy

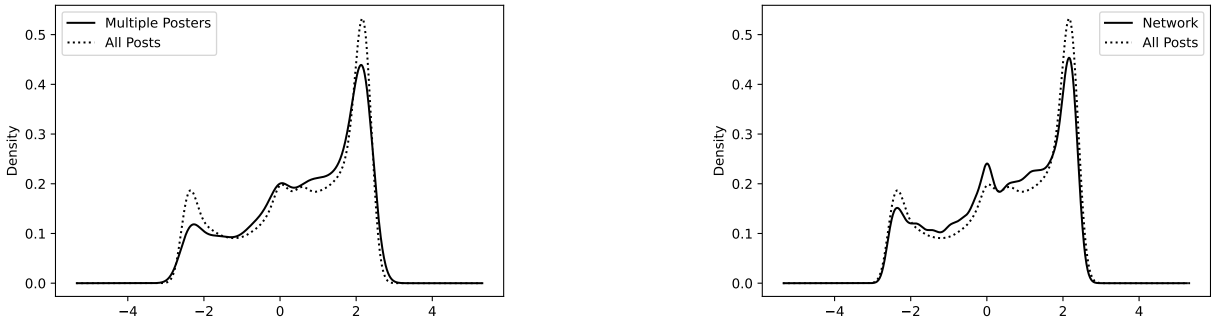
A potential concern with the approach in Section 3.2.1 is whether the sentiments expressed by individuals who post multiple times or are part of the commenters network follow the same distribution as those of that of all submissions on the forum. Figure 7a presents the distribution of sentiments for the second or later post of an author about a ticker, vs that of all submissions across all tickers. Figure 7b presents the distribution of sentiments for those who comment on other's posts, vs that of all submissions across all tickers. Figure 7 provides evidence that the distributions are very similar,

Table 8: Peer Influence: Quantifying Peer Influence – Network Regressions

		Dependent Variable – $\Phi_{i,j,t}$		
		Reduced Form	Full Second Stage	Random Network
		(1)	(2)	(3)
Independent Variables	$\phi^{-1}_{i,j,t-1}$	-0.35 (0.05) ***	-0.35 (0.05) ***	-0.35 (0.05) ***
	$\phi^{0}_{i,j,t-1}$	0.03 (0.03)	0.03 (0.03)	0.03 (0.03)
	$\phi^{+1}_{i,j,t-1}$	0.22 (0.03) ***	0.22 (0.03) ***	0.23 (0.03) ***
	$\bar{\Phi}_{-i,j,t-1}$	0.07 (0.01) ***	0.35 (0.07) ***	0.02 (0.01)
	$r_{j,t}$	0.94 (0.13) ***	0.95 (0.14) ***	0.95 (0.14) ***
	$\bar{r}_{j,t}$	0.64 (0.44)	0.64 (0.41)	0.71 (0.42)*
	$\sigma^2_{j,t}$	-0.63 (0.58)	-0.54 (0.58)	-0.56 (0.58)
	Ticker Fixed Effects	X	X	X
	No. Observations:	19,639	19,620	19,879
	R^2:	0.08	0.08	0.08
	R^2_{adj}:	0.05	0.05	0.05

Notes: The dependent variable is individual investor sentiment about an asset, scaled continuously between $(-\infty, \infty)$. We estimate it using the individual's previously expressed sentiment about the same asset ($\phi_{i,j,t-1}$) as a categorical variable, with the author not having posted previously ($\phi^{NA}_{i,j,t-1}$) as the baseline. We control for a set of market control variables ($r_{j,t}, \bar{r}_{j,t}, \sigma^2_{j,t}$). The sentiment of posts that the author commented on previously ($\bar{\Phi}_{-i,j,t-1}$) is estimated several ways. In column (1), we present the estimate using the sentiment of posts the author previously commented on. In column (2), we use an IV to predict the sentiment of posts the author comments on. In column (3), we randomly rewire the network, connecting the author to a random set of posts about the same ticker. Robust standard errors, clustered at the ticker level, are presented in parentheses. Observations with incomplete market data are dropped.

*** Significant at 1% level ** Significant at 5% level * Significant at 10% level



(a) Frequent Posters Sentiment PDF

(b) Commenters Sentiment PDF

Figure 7: **Density Plot of Sentiments Expressed on WSB**; We present the density plot of the sentiments expressed by users on WSB who post multiple, labeled as *Multiple Posters*, those who comment on others' posts, labeled as *Network*, and that of all submissions, labeled as *All Posts*.

which supports the hypothesis that our analysis offers insight into how all individuals on WSB form opinions.

A second concern is whether we effectively control for unobserved ticker characteristics. Similarly to Patacchini & Zenou (2016), we run 'placebo tests', where we replace the composition of an authors peers (who post between an author's posts about a ticker or posts that an author comments on) with a random cohort of people who post on WSB about the same ticker. The random cohort is chosen as follows. We observe how many peers an individual author has (how many other authors post about the same ticker mentioned in the author's posts between an author's two post or how many posts an author comments on). We then select a random sample of the same number of individuals, without replacement, who post *before* the author's original post (if fewer individuals post before, we select all of those individuals). The results are presented in Tables 7 and 8, column (3). We observe that all the coefficients remain close to their original values, except for the peer effect, which becomes insignificant. This lends credibility to our peer identification strategy and that unobserved factors that influence within ticker variation are not confounding our estimates.

C Details on testing for market impact

C.1 Model with Contagion and Consensus

Longer Time Scales and Elasticities of Demand On longer scales, larger bull runs form, depending on the responsiveness of hype and other investors to price changes. To see this, we write $Y(p)$, the demand function for hype investors, and $S(p)$ for other investors. This time, we assume

$$\frac{\partial S(p)}{\partial t} = 0,$$

so that the adjustment in shares held by other investors is purely endogenous with respect to price. The total derivative of Eq. 10 thus yields

$$\frac{\dot{p}}{p} = \frac{1-m}{(\varepsilon_Y + s[\varepsilon_S - \varepsilon_Y])} \frac{\dot{Y}}{Y}, \tag{29}$$

where

$$\varepsilon_Y = -\frac{p}{Y} \frac{\partial Y(p)}{\partial p}$$

is the price elasticity of demand for hype investors, and, likewise,

$$\varepsilon_S = -\frac{p}{S} \frac{\partial S(p)}{\partial p}$$

is the price elasticity of demand for non-hype investors. We denote $s3 = S/Q$, the ratio of shares owned by non-hype investors. Eq. 29 demonstrates that prices may shoot off indefinitely if, between all investors, increasing prices fail to trigger enough shareholders to sell. A crash would only take place once enough investors decide to increase their response to higher prices, causing a sign reversal in Eq. 29. This type of model, where the responsiveness of agents' demand to price varies, is studied extensively by Hommes (2013), and so we do not dwell on it here.

Contagion We leverage our findings from Section 3.4 to understand the dynamics which govern how hype investors enter the market for a certain asset through word-of-mouth information transmission. Ignoring the presence of alternative investment options, the mechanism for information transmission is alone responsible for the spread of awareness on the asset as an investment possibility (unlike the empirical framework, which assumed users to make a choice over many alternative investment options). We categorise the fixed number of hype investors N between the shares that are i) unaware (z), ii) aware ($a = A/N$), and iii) aware but disinvested (n):

$$\dot{z} = -caz, \quad \dot{a} = caz - da, \quad \dot{n} = da, \quad z + a + n = 1.$$

As in Section 3.4, parameter c is the average rate at which awareness of an investment opportunity transmits between investors per contact case, assuming homogeneous mixing, and d is the rate at which trading on a stock by an investor ceases permanently. While WSB only reflects awareness over time alone, and not the presence of unaware or disinvested users, we can still use time series data on the count of authors discussing specific stocks to estimate a logistic model for coefficients c and d, as in Section 3.4.

Derivation of Eq. 10 In this section, we show the steps for deriving Eq. 10. We begin with:

$$\frac{\dot{p}}{p} = \frac{MN}{pQ} \times \frac{d}{dt}[a(t)\phi(p,t)].$$

The derivative with respect to time is

$$\frac{\dot{p}}{p} = \frac{MN}{pQ}\left(\frac{da}{dt}\phi + \left[\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial p}\frac{dp}{dt}\right]a\right),$$

where we take the partial derivative of A, and break down the derivative of ϕ with respect to time and price, since both determine the overall change in ϕ. Rearranging yields

$$\frac{\dot{p}}{p} - \frac{MNa}{pQ}\left(\frac{\partial \phi}{\partial p}\frac{dp}{dt}\right) = \frac{MN}{pQ}\left(\phi\frac{da}{dt} + a\frac{\partial \phi}{\partial t}\right),$$

and the left-hand side is simplified as

$$\frac{\dot{p}}{p}\left(1 - \frac{MNa}{Q}\left(\frac{\partial \phi}{\partial p}\right)\right) = \frac{MN}{pQ}\left(\phi\frac{da}{dt} + a\frac{\partial \phi}{\partial t}\right), \tag{30}$$

given $\dot{p} = dp/dt$. which gives us the components (1) and (3) on the right hand side of our equation.

To derive component (2), recall that

$$pQ = pY + pS,$$

and denote $s = S/Q$, the proportion of shares owned by non-hype investors. Finally, the hype investor asset demand identity holds, $pY = \phi MNa$, such that the total value of shares owned by hype investors (pY) equals the number of active hype investors in the market, Na, the average amount of money they hold, M, and their buying intensity ϕ. Thus, we rearrange the identity

$$\begin{aligned} pQ &= pY + pS, \\ 1 &= \frac{\phi MA}{pQ} + s, \\ (1 - s) &= \frac{\phi MNa}{pQ}. \end{aligned}$$

We define the price elasticity of buying intensity as

$$\varepsilon_{\phi} = -\frac{p}{\phi}\frac{\partial \phi}{\partial p},$$

such that

$$
\begin{aligned}
(1-s)\varepsilon_{\phi} &= -\frac{\phi M N a}{pQ} \times \frac{p}{\phi}\frac{\partial \phi}{\partial p} \\
&= -\frac{MNa}{Q}\frac{\partial \phi}{\partial p}.
\end{aligned}
$$

Substituting this results into Eq. 30 yields

$$
\frac{\dot{p}}{p}\Big(1+(1-s)\varepsilon_{\phi}\Big) = \frac{MN}{pQ}\Big(\phi\dot{a}+\dot{\phi}a\Big).
$$

Eq. 10 is obtained by dividing by the elasticity term:

$$
\frac{\dot{p}}{p} = \frac{MN}{pQ} \times \frac{\phi\dot{a}+\dot{\phi}a}{1+(1-s)\varepsilon_{\phi}},
$$

where

$$
\dot{\phi} = \frac{\partial \phi}{\partial t} = \frac{\hat{\alpha}}{2\hat{\gamma}\left(\frac{\dot{p}}{p}\right)^2}\operatorname{sech}^2\left[\frac{\hat{\beta}\frac{\dot{p}}{p}+\hat{\alpha}\phi(p,t)}{2\hat{\gamma}\left(\frac{\dot{p}}{p}\right)^2}\right] \tag{31}
$$

is the partial rate of change in buying intensity with respect to time,

$$
\varepsilon_{\phi} = -\frac{p}{\phi}\frac{\partial \phi}{\partial p} \tag{32}
$$

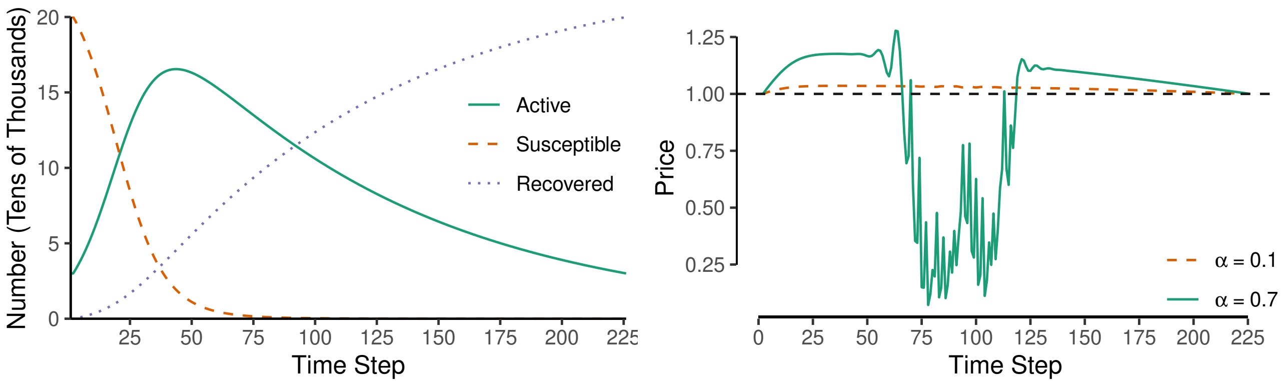
is the price elasticity of the buying intensity, and $s = S/Q$ is the fraction of shares owned by non-hype investors.

C.2 Asset Price Simulation

Discrete Time Simulation We perform a discrete time simulation of our final model, presented in Eq. 10.

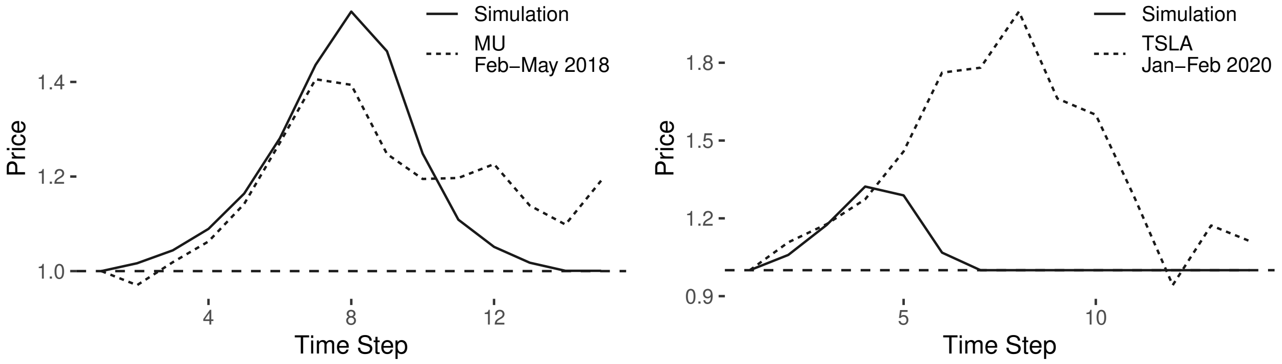
The choices made to simulate the model include: i) some fixed fraction of hype investors, a_0, are always present with a fixed buying intensity, and ii) non-hype investors fix the nominal value of their investment. In practice this implies that hype investors enter the market to buy the asset (since $\phi_0 > 0$), which other investors sell to them at a higher price. The price changes and coordination of buying strategies affect the intensity with which hype investors purchase the asset in subsequent periods. Gradually, they leave the market until only the starting amount a_0 remain, and the price returns to its 'true' value.

Since the focus is on asset demand by hype investors, we simulate various parameters for coordination, α. Given the complex dynamics, qualitatively different patterns could be generated depending on the choice of demand parameters. The result, in Figure 8b, displays three important characteristics for scenarios where consensus is strong ($\alpha = 0.7$) and weak ($\alpha = 0.1$); other parameters are chosen in line with empirical observations and discussed further below. Under strong consensus, the initial increase in demand, until time step 50, gradually drives up price. Once this growth is depleted, the price crashes dramatically, and extreme volatility persists until time step 150. At this stage, some active investors still remain in the market, but the price follows a steady downward trend until it reaches its initial level. In contrast, the effect under weak consensus is relatively mute, as new entrants fail to adopt the sentiments of existing investors. These findings support the conclusions in De Long et al. (1990), stating that if 'sophisticated investors' time horizons are long, compared to



(a) Contagion among Hype Investors

(b) Impact on Price



(c) MU Observed and Simulated Price Dynamics

(d) TSLA Observed and Simulated Price Dynamics

Figure 8: **Price Impact of Contagion and Opinion Dynamics**; In Figures 8a, 8b we illustrate the hypothetical dynamics that can emerge within our model. On the left, the number of active hype investors in an asset increases steeply until time step 50, after which the number slowly declines to its initial value (as 'active' investors 'recover'). On the right, at price first increases until time step 50, after which it experiences a volatile crash when consensus formation is strong ($\alpha = 0.7$), then returns to its initial level. In contrast, when consensus formation is weak ($\alpha = 0.1$), very little price impact is observed. Other parameter choices are discussed in Appendix C.2. Figure 8c, 8d show the dynamics simulated using parameters derived from WSB in the two stocks at a time when the tickers were increasing in popularity within the forum. The stock price displayed is the average weekly close price for each stock (ensuring consistent time steps between asset price and simulation).

'noise traders', they can 'buy low, confident that they will be able to sell high when prices revert to the mean'. Eventually, the asset's price reverses to its original level as hype investors exit the market and other investors trade to revert the asset to its fundamental value.

Figures 8c, 8d on the other hand consider the dynamics that our model implies for two popular stocks within the WSB forum. We initiate the model with coordination and contagion parameters from our data. We emphasize that our goal is not to show that our model can predict stock price movements, but rather that stock price fluctuations, indeed, appear to follow a bursty pattern, consistent with our model and contagion / consensus among hype investors. We observe that in our TSLA simulation, we significantly under-predict the price change. This may be due to our underestimation of of the number of hype-investors versus other investors, or the fact that following a significant price increase, other investors also change their dynamics.

D Topic model and narratives

D.1 Topic model

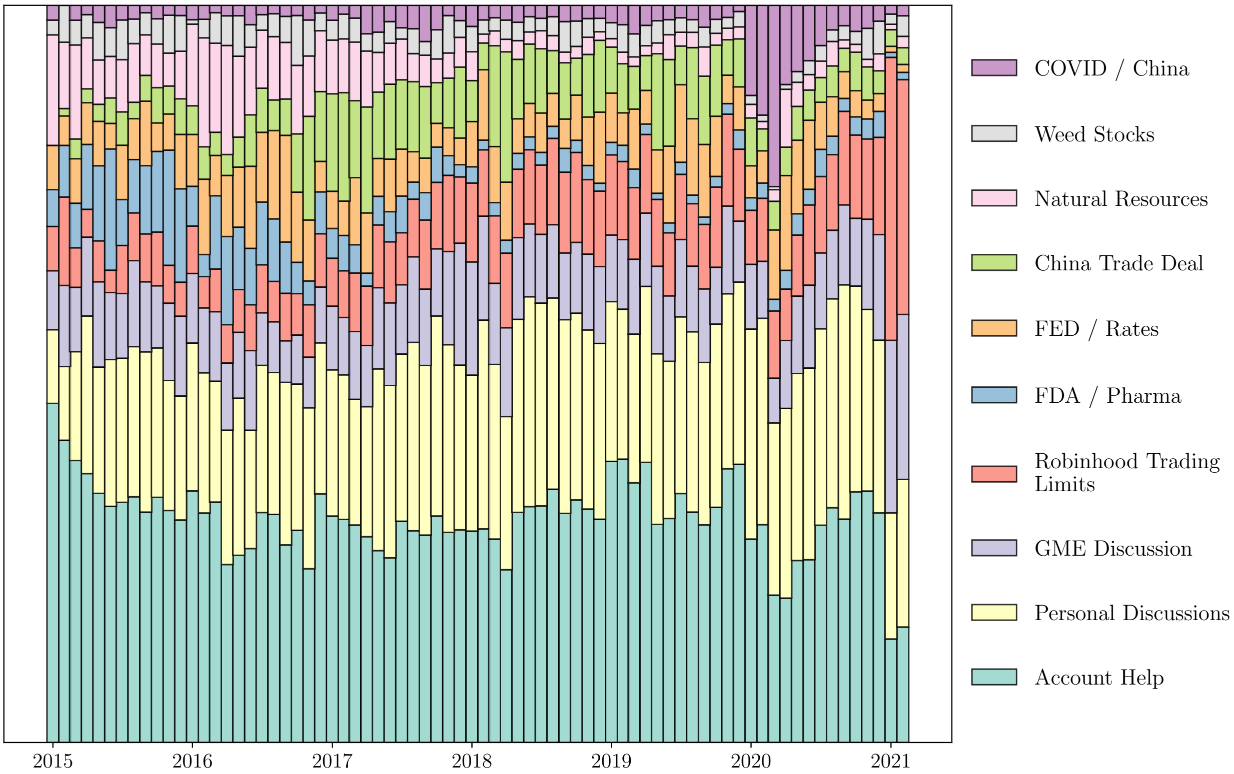


Figure 9: **Temporal trends in topics**; the stacked count, normalized to 100% at each time period, showing the prevalence of a select subset of topics discussed on WallStreetBets.

Does WSB reflect new information for the larger market to trade on, or social activity that drives perceived changes in value, regardless of fundamentals? A topic model offers a simple method to evaluate the content of WSB discussions. Figure 9 presents our preferred topic model, namely the Biterm Topic Model (BTM), which is optimal for smaller bodies of text (Yan et al. 2013). Submissions from April 2012 to February 2021 give a time series of almost 100 months. A random sample of submissions is drawn for the months of January and February 2021 in order to prevent these two months, with a high number of submissions, from skewing the topic model results.

Figure 9 presents a stacked plot of the monthly submission count of a selected subset of discussions, normalised by the total across the selected topics. It begins in 2015 when the forum gained a consistent user base. On one hand, some topics persist in the overall discussion: people consistently ask for advice about trading accounts and anonymously share details of how their trading is affecting their personal lives. On the other hand, topics concerned with larger economic trends wax and wane over the observation period. Two examples of this are the uptick in submissions discussing GME and Roinhood account trading limits, coinciding with the GME short squeeze, and the COVID-19 topic, which is negligible until January 2020, but gains prominence in the subsequent months. Pharmaceutical company and natural resource discussions, on the other hand, seem to loose popularity. A full list of topics with their respective keywords is presented below.

Topic Title	**Top Words**	**Topic Prevalence (%)**
Robinhood Trading Limits	robinhood, gme, account, nkla, margin, order, app, limit, broker, orders	1.8
International Trade	expected, yr, china, usd, europe, japan, pmi, manufacturing, korea, data	0.8
Retail Sales + Amazon	sales, amazon, home, stores, business, online, companies, food, store, retail	3.0
Top Stock Picks / Positions	tsla, news, sold, aapl, weeks, holding, hold, amd, months, dip	11.6
Other	comments, daily, best, moves, spy, weekend, fo, fn, fm, fp	1.1
Other	mentions, vote, log, wsbvotebot, submission, posts, check, reverse, mention, great	0.2
Electric Cars	tsla, energy, car, ev, cars, nio, electric, battery, elon, space	2.3
Revenues, Earnings, Ratings	revs, beats, tgt, line, eps, neutral, downgraded, initiated, fy, reports	0.9
FDA / Pharma	drug, fda, phase, patients, vaccine, trial, treatment, results, clinical, covid	2.9
Revenues, Earnings, Ratings	revenue, million, growth, quarter, share, sales, billion, net, expected, eps	4.0
China Trade Deal	trump, china, said, president, deal, bill, house, election, chinese, news	3.1
Social Media Stocks	fb, game, snap, aapl, disney, games, video, google, users, netflix	2.7
GME Discussion	companies, gme, investors, world, years, believe, hedge, actually, value, funds	7.3
Financial News	data, information, news, financial, report, based, find, sec, research, investors	4.3
Personal Discussions	life, wife, ass, little, said, spy, getting, old, red, went	7.6
Weed Stocks	million, share, capital, ceo, cannabis, ipo, public, management, merger, billion	2.1

Table 9: Topics Extracted from BTM Model

Topic Title	**Top Words**	**Topic Prevalence (%)**
Software Tech Stocks	amd, aapl, intc, companies, data, cloud, software, technology, services, tech	3.4
Earnings Release	release, estimates, consensus, share, revenue, move, open, beat, average, interest	1.0
Natural Resources	oil, gold, prices, silver, gas, futures, crude, production, demand, companies	1.8
FED / Rates	fed, rates, rate, economy, markets, economic, said, interest, growth, inflation	6.0
Other Tech Stocks	tsla, pltr, elon, musk, mods, ban, gme, retards, gains, autists	2.0
COVID / China	virus, covid, cases, coronavirus, china, world, weeks, corona, states, news	4.8
Other	spy, bear, gay, text, bears, bull, gang, msft, words, stonks	1.7
Debt / Loans	debt, cash, pay, credit, loans, loan, million, interest, billion, bank	3.4
Other	usd, bln, exp, revenue, newswires, eps, co, share, symbol, live	1.4
Other	spy, chart, close, index, month, performance, major, past, futures, sectors	1.3
Account Help	account, help, investing, best, start, robinhood, advice, work, years, please	7.4
Other	spy, volume, chart, support, bullish, low, trend, resistance, bearish, line	3.7
Other	amet, calendar, releases, wed, link, thurs, tues, fri, analyst, close	0.5
Options / Risk	option, spy, profit, strike, spread, risk, loss, value, selling, position	6.3

Table 10: Topics Extracted from BTM Model