Stock Tribes: Social Identity in online Stock Communities

Doris Zhou

January 16, 2025

Abstract

I fine-tune a large language model to construct a novel measure of a social characteristic: strong social identification with a stock. Using data from a popular investment-focused social media platform, I provide evidence that tribalism influences how investors form connections. Specifically, tribalist investors tend to receive fewer connections overall. Applying statistical social network models, I find that tribalist investors are less likely to connect with other tribalist investors. These findings suggest that tribalism may partially explain why retail investors often hold onto losing stocks.

JEL classification: G4, G11, G41, G53
Keywords: Social Finance, Sentiment Contagion, Social Identity, Social Networks

1. Introduction

Both social interactions and information acquisition have increasingly moved to online spaces in recent years. Despite researchers recognizing the importance of social interactions (Shiller, 2017), there still remains much to understand on how social activities influence outcomes in finance. Recent studies have used online social media data to investigate sources of investor disagreement (Cookson and Niessner, 2020; Cookson et al., 2024) and information dissemination (Hirshleifer et al., 2024). Theories tend to focus on information as sources for disagreement: 1) investors have difference in information sets or 2) investors have different models used for interpreting information. Interactions in financial markets may take on a more direct social nature that is dominated by emotions rather than information dissemination.

The GameStop event in 2021 highlights that investors may trade for non-financial or "rational" economic reasons and that such trading can have substantial economic effects. Shiller (2017) calls for the expansion of economics to "...include serious quantitative study of changing popular narratives". Political and social narratives may affect trading decisions. Both narratives and emotions can be contagiously spread and social media sites become an important environment to study the role of human emotion in trading decisions. Kakhbod et al. (2023) find that more unskilled users have higher followers than more skilled users, which suggests that users do not just go on social media sites to obtain information. In general, people often use social media for entertainment and to connect with like-minded communities. Similarly, there is no clear reason to believe that users of investment trading social platforms are any different – these users form social bonds and create "tribes".

In this paper, I study how tribalism, or the strong identification with a particular stock community affect the social networks that investors form and how such social networks impact social learning. Using data from a popular stock based social media site I fine-tune a state-of-the-art large language model to distinguish posts that exhibit identification with a particular stock community. I leverage statistical social network models to study how

community identifications impact how online social networks are formed.

My results provide a new perspective on investor disagreement that is not rooted in an information model. Prior studies in finance have extensively studied homophily (Stolper and Walter, 2019; Dagostino et al., 2023; Lu et al., 2024). Social identity has been studied previously in finance and has been shown to be able to explain various corporate finance outcomes (Fracassi and Tate, 2012; Lim and Nguyen, 2021; Jiang et al., 2019). To the best of my knowledge, this is the first study to explicitly measure community identity formed around a stock and study how this impacts retail investment behavior.

I use StockTwits, a popular social media site specifically for traders. The format of StockTwits is similar to X. Each post is subject to a 1000 character limit. Users mark the stock they are discussing using the $ symbol, for example $AAPL and similar to X, message posts are organized under threads.

I begin by identifying tribalism on StockTwits. I leverage machine learning tools by fine-tuning a RoBERTa model on a hand-labeled dataset. I first investigate the relationship between tribalism and the influence of the investor. I find that investors who exhibit strong tribalist traits have fewer followers compared to neutral investors. However, the number of investors tribablists follow are not significantly different from the those of neutral investors. Tribalism thus seems to social trait that is negatively perceived by the others.

Moreover, I examine the role of skill for investors forming social networks. If skilled investors are skeptical of the informativeness of the posts by other investors perceived as less skilled, then skilled investors should follow fewer people. On the other hand, if skilled investors think exposure to a variety of opinions is useful, then skilled investors should have large followees. I find that anti-tribalist investors who are skilled follow 62 fewer people than socially neutral investors.

Previous research has studied theoretical models of social networks (Pedersen, 2022). I use exponential random graph models (ERGMs) to analyze the formation and evolution of a social network and its implications for social learning.

This paper is related to several strands of literature. This paper contributes to the emerging but fast-growing literature of social finance (Hirshleifer, 2020). Dim (2020) studies Social Media Investment Analysts (SMAs), who are essentially influential users "finfluencers" using Seeking Alpha data. The author estimates a mixture model to separate the user's skill from luck and finds that over half of SMAs are skilled but 13% are able to generate substantial returns. Cookson et al. (2023) use StockTwits data and document that investors selectively expose themselves to information that already align with their views, resulting in echo chambers. Han et al. (2022) present a model of social transmission that focuses on one type of social transmission bias, the self-enhancing bias. They derive that investors in a social network are more likely to adopt the opposite trading strategy if there are more investors in their network with that strategy. Sui and Wang (2023) empirically study self-enhancing bias by using a Chinese social media site that is similar to StockTwits. Han et al. (2023) study social learning in bitcoin markets. They use data from a popular online forum for discussing bitcoin. They find a positive correlation between investors with positive sentiment and trading volume during bubble episodes. Kakhbod et al. (2023) study the role of finfluencers using StockTwits data. They find that users do not follow informed influencers. Rather, they follow other users with similar behavioral traits. An interesting implication of their study is that competition of information from social media does not drive out unskilled finfluencers.

Prior studies have used variables constructed from social networks to study financial outcomes. Xie et al. (2020) study how well network cohesion can predict prices, using data from a bitcoin social forum site. The authors find that less cohesive networks are better at predicting future returns compared to more cohesive networks. Chen et al. (2024) study the effect of social media network structure on short-term market reactions to buy recommendations from influential users. They document a negative relationship with social cohesion and short-term market reactions. Neither of these papers, however, have studied networks using social network statistical models. My study is the closest in methodology with the paper by Deng et al. (2023). The authors use a separable temporal EGRM to

conduct a detailed analysis of link formation and dissolution using data from a large social trading platform. My study is different from these papers because I focus on a specific behavioral phenomenon, tribalism, and I study how this social trait impacts the formation of the social network.

The rest of this paper is organized as follows. Section 2 describes the data and the construction of key variables. Section 3 presents the results of how tribalism impacts social influence. Section 4 discusses the relationship between tribalism and the structure of the social network. Finally, section 5 concludes.

2. Data and Variables

2.1. Data Sources

I obtain message and user level data from the popular trading-focused social media site, StockTwits. My sample period is from January 2, 2020 through September 2, 2024. Users of StockTwits use cashtags to signal the stock they are discussing. A post can have an unlimited number of cashtags and stocks. To reduce noise, I require that all posts contain exactly one cashtag so that it is clear that the content of the message refers to the specific stock. As Table 1 shows, there are 90,467,437 messages during this time period. In my sample, there are 266,251 unique users and 16,305 unique stocks that were discussed.

2.2. Variable Construction

I define social identity as the sentiment of belonging to the community of other users who trade the same stock. Often, this includes the position the investor has stated he/she has taken (long or short), or holding either bullish or bearish sentiment. The following are examples of actual posts taken from the data that I would define as exhibiting social identity:

"$VKTX hang tight Longs ... day traders see ya later!"

"$MNKD i blocked, muted so many shorts but still they are showing up. theyre just endless

no life craps."

In the first example, the poster refers to "Longs" as a team. In the second post, the investor is extremely anti-shorts, to the point of blocking others just for taking a different position. Thus, this user seems to identify strongly as part of the "Long team" of stock ticker $MNKD. The following are examples of posts that exhibit anti-social identity:

"$ROKU Thanks idiots. Sold @ 24.00. 10-G profit -"

"$HCLP @ryandelpew You're just about the most miserable person I've ever conversed with. Just go find a bridge..."

The first example shows the user uses negative words to describe the group of other users who presumably are also discussing about the stock $ROKU. In the second message, the poster is very mean to another individual user. The next two posts are examples of what I define to be neutral posts:

"$OTIC What's the scoop? Why the jump?"

"$SDRL sold at 0.37, probably should have held until tomorrow, oh well."

The first post contains two questions, with no obvious identification with either the long or short position or even the stock ticker. The statement in the second post does not suggest the user strongly identifies with the long position. In fact, the user states he/she has already sold, which is in contrast to the "diamond hands" sentiment. Because it is possible that a user may be very negative one day to a certain stock ticker community but may be more positively social to another stock ticker community, my final social identity value is aggregated across posts for each user.

In detail, to construct the social identity variable, I first apply a large language model to label each post in my dataset as exhibiting social identity, neutral, or exhibiting anti-social identity. I fine-tune a RoBERTa model (Liu, 2019), a state-of-the-art transformer model that is a refinement over the well-known BERT model (Kenton and Toutanova, 2019). To construct my training dataset, I hand-labeled 67,160 randomly selected posts as $\{-1, 0, 1\}$ displaying anti-social, neutral, or social identity, respectively. As Table 10 shows, my fine-

tuned model achieved greater that 94% precision for all three categories. From Table 1, the average social identity score *SocID_score* is positive but close to zero at 0.193, showing that most users tend to be neutral.

My final *SocID_score* is aggregated for each investor at the stock-day level. I compute *SocID_score* according to this definition of social sentiment from Kakhbod et al. (2023):

$$SocID\_score_{i,j,t} = \max\left\{-1, \min\left(1, \sum_{n=1}^{N_{i,j,t}} 1\{SocID_{i,j,t} = 1\} - \sum_{n=1}^{N_{i,j,t}} 1\{SocSent_{i,j,t} = -1\}\right)\right\} \tag{1}$$

where $n = 1, ..., N_{i,j,t}$ is the index of the StockTwits post. Thus, $SocID\_score_{i,j,t} \in \{-1, 0, 1\}$ is the social identity score for investor i about stock j on day t. For the rest of this paper, I call investors such that $SocID\_score_{i,j,t} = -1$ *skeptics*, $SocID\_score_{i,j,t} = 1$ *tribalists* and $SocID\_score_{i,j,t} = 0$ *social-neutral*.

To construct the sentiment measure, I apply the FinVADER model to classify posts, which is a fine-tuned version of the VADER sentiment model for the finance lexicon (Hutto and Gilbert, 2014; Korab and Contributors, 2025). The VADER model has been applied previously for sentiment analysis of StockTwits data by finance and economics researchers (see for instance Cookson and Niessner (2023)). The output of the FinVADER model is a continuous value bounded between -1 and 1. I follow the definition of the social sentiment score from Kakhbod et al. (2023) and construct an continous version of their $SocSent_{i,j,t}$ score. I aggregate all posts made by user i about stock j on day t:

$$SocSent\_score_{i,j,t} = \max\left\{-1, \min\left(1, \sum_{n=1}^{N_{i,j,t}} 1\{SocSent_{i,j,t} = 1\} - \sum_{n=1}^{N_{i,j,t}} 1\{SocSent_{i,j,t} = -1\}\right)\right\} \tag{2}$$

where $n = 1, ..., N_{i,j,t}$ enumerates the post. This definition normalizes $SocSent\_score_{i,j,t}$ to between -1 and 1. As Table 1 shows, the average $SocSent\_score_{i,j,t}$ is positive but close to zero, suggesting that most users write posts with neutral sentiment.

Additionally, I follow the methodology in Kakhbod et al. (2023) to compute α_i, which is

a measure of the skill of the users. Specifically, I estimate the univariate regression

$$SocSent\_score_{i,j,t} * AbnRet_{j,t+1,t+H} = \alpha_i + \epsilon_{i,j,t+1,t+H} \tag{3}$$

where H is the time horizon over which abnormal returns are computed for stock j. I set $H = 2$ days. For more detail, see Kakhbod et al. (2023).

3. Follower-Followee Decisions

I begin by investigating the relationship between the investor characteristic of exhibiting social identity and the decision to follow or be followed by others. Strong exhibition of tribalism may signal to others the poster is committed to the stock. This may invoke a positive reaction from others because they may think the poster has a rational reason to be so committed to the stock. On the other hand, strong tribalism may be interpreted as a negative signal by others. The poster may be posting repetitive messages with the same sentiment that are not informative. Tribalism may be seen as a negative behavioral trait associated with closed-mindedness and prejudice.

I construct annual snapshots of the social network of follower-followee ties. For each annual social network, I compute the in-degree and out-degree for each investor. The in-degree is the number of connections directed towards the investor and thus represents the number of followers the investor has. The out-degree is the number of connections directed away from the investor and so signifies the number of followees the investor has.

Table 2 presents the results on the relationship between the number of followers an investor has and the strength of their social identification to the stock. The reference variable is when $SocID = 0$ or when the investor is neutral. Across all model specifications, users with positive social identification is associated with less followers compared to investors who express neutrality regarding stock communities. The first column shows the baseline results. Skeptics have 50 less followers than neutral-social investors. Investors who exhibit positive

social identification have almost 19 less followers than social-neutral investors. I next control for sentiment. Previous research has documented evidence of echo chambers regarding sentiment on StockTwits (Cookson et al., 2023). Therefore, it is possible investors selectively follow others who share the same sentiment. Unsurprisingly, the coefficient for *sentiment* is positive. However, investors who show positive social identification still statistically have less followers than neutral-social investors. As column (3) shows, in the third model specification, I control for the number of ties that are mutual. Large online influencers will often have large follower counts but themselves follow a very low number of users. Investors who positively identity with the stock communities still have significantly less followers than neutral-social investors. In fourth model specification, I control for investor skill. Previous research has found that unskilled finfluencers tend to have larger followings compared to more skilled finfluencers (Kakhbod et al., 2023). The negative, though insignificant coefficient in column (4) is consistent with the notion of less skilled investors having less followers.

How does tribalism impact the investor's number of followees? Table 3 presents the results on the relationship between the direction of the investor's social identity with the stock community and the number of followees. Across all four model specifications, the coefficients for *Soc_neg* is positive. The results are significant in the third and fourth model specifications. This suggests that investors who do not identify strongly with particular stock communities may be more skeptical of stock communities and are more receptive to other investors who may have diverse opinions. Furthermore, this suggests that while these skeptic investors may be more broadly active in many stock communities. Interestingly, while the coefficients are also mostly positive (expect for model 2) for *SocID_pos*, these coefficients are much smaller in magnitude compared to the coefficients for *SocID_neg* and are insignificant.

Taken together, the results shown in Table 2 and Table 3 imply other investors have a negative impression of tribalist investors. In the next section, I use statistical social network analysis models to more rigorously test the network formation among users with different social identification scores. Interestingly, feelings of social identity do not have a statistically

significant effect on out-degree, or the number of followees. This suggests that investors with strong social identification feelings (positive or negative) are not themselves affected by this trait in determining who to follow. Rather, social identification is a trait that is perceived externally, entering the decision making by other investors.

Investors may go on StockTwits for entertainment or for information. In the latter case, the skill of the poster should be an important variable in determining the follower-followee social network. In the next set of analyses, I ask how does skill interact with tribalism? I construct an indicator variable *top_10* that is equal to 1 if investor i's measure of skill α_i is in the top ten percentile of the sample. Table 4 shows the results when the in-degree, or of number of followers is the dependent variable. Across all three model specifications, the coefficient for the interaction term *socID_neg*top_10* is negative but insignificant. However, the coefficient for the interaction term for positive social identity, *socID_pos*top_10* is negative and statistically significant. From column (3), investors who positively identity with their stock communities and are highly skilled, have on average 62 less followers compared so socially-neutral investors.

How does skill affect the investor's decision of followees? Skilled investors may be dubious or have the ability to discern the informativeness of other investors' posts. Thus, skilled investors may follow less people compared to less skilled investors. The negative coefficient from column (4) Table 3 suggests offers preliminary support for this conjecture. From Table 5, the coefficient for *top_10* is negative and statistically significant across all model specifications. This suggests that investors in the top ten percentile for skill tend to follow less people. The coefficients for the interaction term *socID_neg*top_10* are negative across all specifications and the coefficient is statistically significant in the full model (column (3)). This means that investors with negative social identification who are in the top 10 percentile of skill follow approximately 10 fewer people compared to neutral investors who are not highly skilled.

Overall, the results from Tables 2 - 5 suggest that tribalism is not perceived positively

by other investors though tribalists themselves do not follow less people than social-neutral investors. Additionally, skeptics tend to follow more users than social-neutral investors but highly skilled skeptics have less followees.

4. Social Network Formation

In this section, I apply statistical social network models to study the structure of the social networks that are formed by investors. I use Exponential Random Graph Models (ERGMs) to estimate the probability of ties forming between any two investor in the social network Wasserman and Pattison (1996); Hunter et al. (2008); Holland and Leinhardt (1981). ERGMs are a general class of models based on the exponential family theory. An ERGM is a generative statistical model, which means that the characteristics of the actors in the network and local structural properties can be used to predict properties for the entire network (Hunter et al., 2008; Luke, 2015). ERGMs are powerful tools for predicting the observed ties in a network. The model accepts a wide variety of predictors including investor-level characters (such as *SocID*) and local structural properties such as the observed degree distribution (Luke, 2015). From Luke (2015), the model that I fit is

$$P(y_{ij} = 1 \mid Y_{ij}^{C}) = \frac{1}{c} \exp\{\sum_{k=1}^{K} \theta_k z_k(y)\} \tag{4}$$

where $\frac{1}{c}$ is a constant, θ_k is the coefficient of the network statistics for each of the K included statistics $z_k(y)$. This model predicts the probability of a connection between investors i and j conditional on the rest of the network. I use Monte Carlo Markov Chain maximum-likelihood estimation to fit all subsequent ERGMs. All estimations of the ERGM model is performed using the STATNET package in R (Krivitsky et al., 2003–2024).

I begin by fitting a baseline model. Table 6 presents the results of a simple ERGM that only uses network variables. Edges represents the tendency for investors to form ties. Mutual measures reciprocity in the network between investors. Each model is ran separately for the

snapshot of the whole investor network for each year. Each of the five panels shows that the coefficient for *edges* is negative, suggesting that the network is sparse. From column (2), we see that the coefficient for *mutual* is positive, indicating that if one investor sends a friendship invite (thus initiating a tie) to another investor, the likelihood of the second investor reciprocating is higher.

Next, I directly model homophily. I examine whether tribalists are more likely to form ties with other tribalists, social-neutral investors are more likely to form connections with other social-neutral investors, or if skeptics are more likely to form ties with other skeptics. Table 7 presents the results. Panels A and B show that the model predicts that investors with the same social identification characteristic are less likely to form ties with each other. Thus, there is evidence of heterophily but from panels C, D, and E, the coefficient for *socID* is positive and insignificant.

Previous research has found evidence of investors forming echo chambers by selectively forming connections with other investors that share the same sentiment of other stocks (Cookson et al., 2023). Column (1) of Table 8 presents the results of the effect of sentiment on the likelihood of investors forming ties. The coefficient is positive for sentiment is positive and significant for 2020-2022 but becomes negative and significant for 2023-2024. Column (2) presents the results from the second model specification, where *sentiment_diff* is the absolute difference between sentiment of any two nodes. Interesting, there is a corresponding sign shift from 2022 to 2023.

Finally, I investigate the effect of both *socID* and sentiment on investors forming connections in the network. Table 9 displays the results. Overall, the results suggest that investors who are similar to each other in their social identity traits are less likely to form ties but investors who are share similar sentiment are more likely to form ties in 2020-2022, then become less likely to form ties in 2023-2024.

5. Conclusion

Tribalism is associated with strong emotions and may explain why retail investors hold on to their losing stocks. Such investors have a strong commitment to the company not necessarily due to economic fundamentals but an emotional attachment to the stock. My finding that tribalists receive fewer followers than socially-neutral investors lend support to the idea that other investors perceive tribalism as a negative social trait. Future research should further investigate if tribalists is a broad phenomenon or if there are certain types of firms that tribalists prefer to invest in.

References

Chen, C., Goh, K. Y., Ke, B., 2024. Social media network structure and stock market reactions to buy recommendations issued by social media analysts. Available at SSRN 4717663 .

Cookson, J. A., Engelberg, J. E., Mullins, W., 2023. Echo chambers. The Review of Financial Studies 36, 450–500.

Cookson, J. A., Lu, R., Mullins, W., Niessner, M., 2024. The social signal. Journal of Financial Economics 158, 103870.

Cookson, J. A., Niessner, M., 2020. Why don't we agree? evidence from a social network of investors. The Journal of Finance 75, 173–228.

Cookson, J. A., Niessner, M., 2023. Investor disagreement: Daily measures from social media. Available at SSRN 4529594 .

Dagostino, R., Gao, J., Ma, P., 2023. Partisanship in loan pricing. Journal of Financial Economics 150, 103717.

Deng, J., Yang, M., Pelster, M., Tan, Y., 2023. Social trading, communication, and networks. Information Systems Research .

Dim, C., 2020. Social media analysts' skills: Insights from text-implied beliefs. Available at SSRN 3813252 .

Fracassi, C., Tate, G., 2012. External networking and internal firm governance. The Journal of finance 67, 153–194.

Han, B., Hirshleifer, D., Walden, J., 2022. Social transmission bias and investor behavior. Journal of Financial and Quantitative Analysis 57, 390–412.

Han, B., Liu, H., Sui, P., 2023. Social learning and sentiment contagion in the bitcoin market. Available at SSRN 4543326 .

Hirshleifer, D., 2020. Presidential address: Social transmission bias in economics and finance. The Journal of Finance 75, 1779–1831.

Hirshleifer, D., Peng, L., Wang, Q., 2024. News diffusion in social networks and stock market reactions. The Review of Financial Studies p. hhae025.

Holland, P. W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. Journal of the american Statistical association 76, 33–50.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., Morris, M., 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of statistical software 24, nihpa54860.

Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol. 8, pp. 216–225.

Jiang, F., Qian, Y., Yonker, S. E., 2019. Hometown biased acquisitions. Journal of Financial and Quantitative Analysis 54, 2017–2051.

Kakhbod, A., Kazempour, S. M., Livdan, D., Schuerhoff, N., 2023. Finfluencers. Swiss Finance Institute Research Paper .

Kenton, J. D. M.-W. C., Toutanova, L. K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, Minneapolis, Minnesota, vol. 1, p. 2.

Korab, P., Contributors, 2025. Finvader: A financial sentiment analysis tool. https://github.com/PetrKorab/FinVADER, accessed: 2025-01-14.

Krivitsky, P. N., Handcock, M. S., Hunter, D. R., Butts, C. T., Bojanowski, M., Klumb, C., Goodreau, S. M., Morris, M., 2003–2024. Statnet: Tools for the statistical modeling of network data. Software.

Lim, I., Nguyen, D. D., 2021. Hometown lending. Journal of Financial and Quantitative Analysis 56, 2894–2933.

Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 364.

Lu, Y., Naik, N. Y., Teo, M., 2024. Diverse hedge funds. The Review of Financial Studies 37, 639–683.

Luke, D. A., 2015. A user's guide to network analysis in R, vol. 72. Springer.

Pedersen, L. H., 2022. Game on: Social networks and markets. Journal of Financial Economics 146, 1097–1119.

Shiller, R. J., 2017. Narrative economics. American economic review 107, 967–1004.

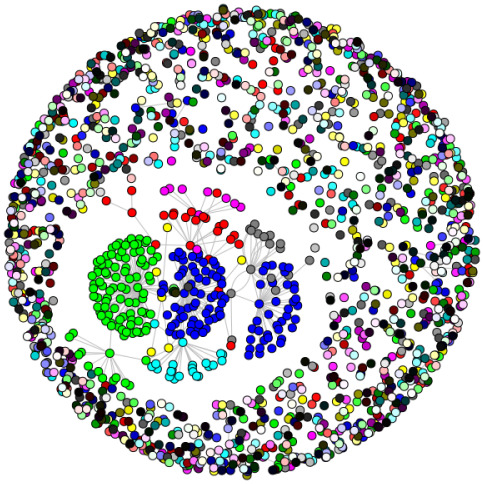
Stolper, O., Walter, A., 2019. Birds of a feather: The impact of homophily on the propensity to follow financial advice. The Review of Financial Studies 32, 524–563.

Sui, P., Wang, B., 2023. Social transmission bias: Evidence from an online investor platform. Available at SSRN 4081644 .

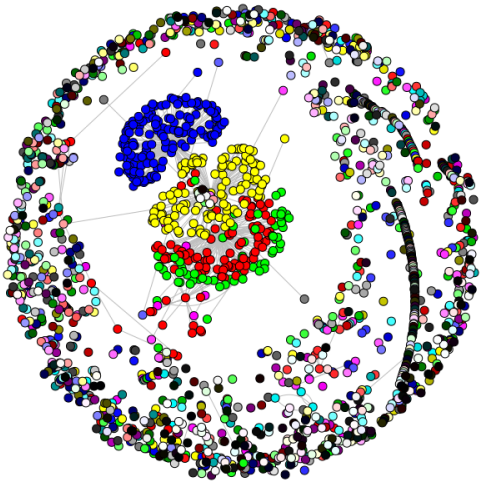
Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p. Psychometrika 61, 401–425.

Xie, P., Chen, H., Hu, Y. J., 2020. Signal or noise in social media discussions: the role of network cohesion in predicting the bitcoin market. Journal of Management Information Systems 37, 933–956.

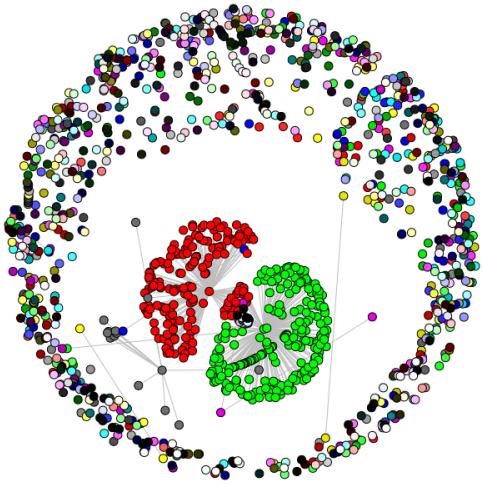
Fig. 1. Network Clusters of StockTwits Investors Through Time



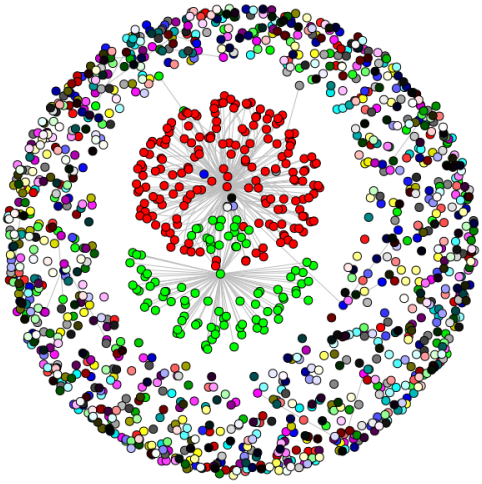
(a) Clusters of users in 2020



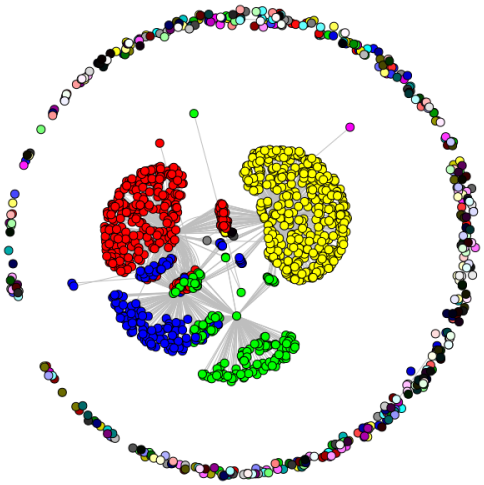
(b) Clusters of users in 2021



(c) Clusters of users in 2022



(d) Clusters of users in 2023



(e) Clusters of users in 2024

Table 1: Summary Statistics

This table presents the summary statistics for the sentiment score, *SocSent_score*, and the social identity score, *SocID_score*.

Statistic	count	mean	std	min	p50	max
Panel A: user-post-day level						
SocSent_score	90,467,437	0.081	0.398	-1.000	0.004	1.000
SocID_score	90,467,437	0.193	0.417	-1.000	0.000	1.000
Panel B: user-day level						
socSent_score_mean	40,014,393	0.071	0.354	-1.000	0.007	1.000
socID_score_mean	40,014,393	0.226	0.392	-1.000	0.000	1.000
num_stocks	40,014,393	1.970	15.636	1.000	1.000	6,001
Panel C: stock-day level						
socSent_score_mean	6,070,377	0.073	0.223	-1.000	0.033	1.000
socID_score_mean	6,070,377	0.081	0.186	-1.000	0.000	1.000
num_users	6,070,377	12.985	91.159	1.000	3.000	44,049
Panel D: user level						
socSent_score_mean	1,247,967	0.082	0.233	-1.000	0.046	1.000
socID_score_mean	1,247,967	0.184	0.263	-1.000	0.093	1.000
num_stocks	1,247,967	12.014	66.552	1.000	3.000	13,757
num_days	1,247,967	32.064	85.064	1.000	3.000	1,180
Panel E: stock level						
socSent_score_mean	28,935	0.091	0.128	-1.000	0.083	1.000
socID_score_mean	28,935	0.088	0.147	-1.000	0.038	1.000
num_users	28,935	518.180	3,176.597	1.000	20	180,527
num_days	28,935	209.794	301.298	1.000	51	1,180
Panel F: day level						
socSent_score_mean	1,180	0.078	0.022	0.023	0.076	0.143
socID_score_mean	1,180	0.186	0.028	0.114	0.183	0.319
num_stocks	1,180	5,144.387	1,152.042	2,565.000	5,288	9,026
num_users	1,180	33,910.503	14,376.744	8,547.000	28,520	101,831
Panel G: Other Variables						
Unique Users	266,521					
Unique Stocks	16,305					

Table 2: Number of Followers – Indegree

This table presents linear regression results where the dependent variable is in-degree, which is the number of followers the investor has. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	Indegree			
	(1)	(2)	(3)	(4)
socID_neg	−50.879*	−48.561	−46.736	−47.455
	(29.991)	(30.008)	(29.669)	(40.891)
socID_pos	−18.899***	−22.140***	−21.112***	−22.689**
	(6.574)	(6.718)	(6.642)	(9.153)
sentiment		9.925**	8.045*	7.537
		(4.230)	(4.183)	(5.788)
mutual_following			6.643***	5.767***
			(0.072)	(0.099)
alpha_no_intercept.y				−33.506
				(56.595)
Observations	374,898	374,898	374,898	185,502
Adjusted R^2	0.000	0.000	0.022	0.018
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: Number of Followees – Out-degree

This table presents linear regression results where the dependent variable is out-degree, which is the number of followees the investor has. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	Outdegree			
	(1)	(2)	(3)	(4)
socID_neg	2.123	2.242	2.609**	2.641*
	(1.435)	(1.436)	(1.116)	(1.586)
socID_pos	0.043	−0.124	0.083	0.054
	(0.315)	(0.322)	(0.250)	(0.355)
sentiment		0.512**	0.133	0.117
		(0.202)	(0.157)	(0.224)
mutual_following			1.337***	1.334***
			(0.003)	(0.004)
alpha_no_intercept.y				−1.571
				(2.195)
Observations	374,898	374,898	374,898	185,502
Adjusted R^2	0.000	0.000	0.396	0.395

Table 4: Number of Followers – Interaction with Skill

This table presents linear regression results where the dependent variable is in-degree, which is the number of followers the investor has. *top_10* is an indicator variable that equals 1 if the investors skill is in the top 10 percentile of the sample. The key independent variables are the interaction terms of the *socID* with *top_10*. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	Indegree		
	(1)	(2)	(3)
socID_neg*top_10	−38.962	−38.350	−41.300
	(152.077)	(152.077)	(150.701)
socID_pos*top_10	−60.499**	−61.320**	−62.404**
	(30.492)	(30.497)	(30.221)
socID_neg	−47.331	−45.218	−43.360
	(42.999)	(43.019)	(42.629)
socID_pos	−15.050	−18.030*	−16.917*
	(9.496)	(9.675)	(9.588)
top_10	31.868	31.335	35.290
	(25.560)	(25.562)	(25.331)
sentiment		9.391	7.739
		(5.841)	(5.788)
mutual_following			5.767***
			(0.099)
Observations	185,502	185,502	185,502
Adjusted R^2	0.000	0.000	0.018

Table 5: Number of Followees – Interaction with skill

This table presents linear regression results where the dependent variable is out-degree, which is the number of followees the investor has. *top_10* is an indicator variable that equals 1 if the investors skill is in the top 10 percentile of the sample. The key independent variables are the interaction terms of the *socID* with *top_10*. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	Outdegree		
	(1)	(2)	(3)
socID_neg*top_10	−9.492	−9.458	−10.140*
	(7.516)	(7.516)	(5.845)
socID_pos*top_10	0.101	0.055	−0.196
	(1.507)	(1.507)	(1.172)
socID_neg	2.880	2.999	3.429**
	(2.125)	(2.126)	(1.653)
socID_pos	−0.043	−0.212	0.046
	(0.469)	(0.478)	(0.372)
top_10	−2.596**	−2.626**	−1.711*
	(1.263)	(1.263)	(0.982)
sentiment		0.530*	0.148
		(0.289)	(0.224)
mutual_following			1.334***
			(0.004)
Observations	185,502	185,502	185,502
Adjusted R^2	0.000	0.000	0.395

x‘

Table 6: ERGM: Baseline

This table presents the results from the ERGMs. *edges* represent the propensity for any two investors to form a tie. *mutual* models the tendency for the ties to be reciprocated. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	(1)	(2)
Panel A: 2020	Year: 2020	Year: 2020
edges	-9.709***	-9.729***
	(0.003)	(0.003)
mutual		5.814***
		(0.04)
Panel B: 2021		
edges	-9.713***	-9.74***
	(0.003)	(0.003)
mutual		6.119***
		(0.029)
Panel C: 2022		
edges	-9.27***	-9.301***
	(0.005)	(0.005)
mutual		5.844***
		(0.037)
Panel D: 2023		
edges	-9.067***	-9.096***
	(0.005)	(0.005)
mutual		5.584***
		(0.039)
Panel E: 2024		
edges	-8.982***	-9.015***
	(0.007)	(0.007)
mutual		5.635***
		(0.056)

Table 7: ERGM: Role of Social Identity

This table presents the results from the ERGMs. *edges* represent the propensity for any two investors to form a tie. *mutual* models the tendency for the ties to be reciprocated. *socID* models homophily, or whether investors that have the same *socID* value will form ties with each other. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Panel A: 2020	
edges	-9.681***
	(0.005)
mutual	5.81***
	(0.034)
socID	-0.096***
	(0.007)
Panel B: 2021	
edges	-9.712***
	(0.005)
mutual	6.116***
	(0.03)
socID	-0.052***
	(0.007)
Panel C: 2022	
edges	-9.306***
	(0.007)
mutual	5.845***
	(0.038)
socID	0.008
	(0.009)
Panel D: 2023	
edges	-9.101***
	(0.007)
mutual	5.587***
	(0.04)
socID	0.01
	(0.01)
Panel E: 2024	
edges	-9.009***
	(0.01)
mutual	5.641***
	(0.057)
socID	-0.011
	(0.014)

Table 8: ERGM: Role of Sentiment.
This table presents the results from the ERGMs. *edges* represent the propensity for any two investors to form a tie. *mutual* models the tendency for the ties to be reciprocated. *sentiment* is a continuous variable. *sentiment_diff* is the absolute difference between the sentiment values of any two investors in the social network. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

	(1)	(2)
Panel A: 2020		
edges	-9.894***	-9.616***
	(0.005)	(0.005)
mutual	5.773***	5.809***
	(0.035)	(0.034)
sentiment	0.204***	
	(0.004)	
sentiment_diff		-0.161***
		(0.006)
Panel B: 2021		
edges	-9.829***	-9.664***
	(0.005)	(0.005)
mutual	6.111***	6.115***
	(0.03)	(0.031)
sentiment	0.112***	
	(0.004)	
sentiment_diff		-0.107***
		(0.005)
Panel C: 2022		
edges	-9.426***	-9.18***
	(0.006)	(0.007)
mutual	5.807***	5.84***
	(0.04)	(0.04)
sentiment	0.209***	
	(0.005)	
sentiment_diff		-0.161***
		(0.008)
Panel D: 2023		
edges	-9.035***	-9.204***
	(0.005)	(0.008)
mutual	5.566***	5.575***
	(0.045)	(0.043)
sentiment	-0.121***	
	(0.005)	
sentiment_diff		0.138***
		(0.008)

	(1)	(2)
Panel E: 2024		
edges	-8.939***	-9.155***
	(0.008)	(0.012)
mutual	5.615***	5.635***
	(0.057)	(0.058)
sentiment	-0.153***	
	(0.007)	
sentiment_diff		0.182***
		(0.012)

Table 9: ERGM: Full Model.
This table presents the results from the ERGMs. *edges* represent the propensity for any two investors to form a tie. *mutual* models the tendency for the ties to be reciprocated. *socID* models homophily, or whether investors that have the same *socID* value will form ties with each other. *sentiment* is a continuous variable. Standard errors are shown in parentheses. Statistical significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Panel A: 2020	
edges	-9.831***
	(0.006)
mutual	5.772***
	(0.037)
socID	-0.142***
	(0.007)
sentiment	0.215***
	(0.004)
Panel B: 2021	
edges	-9.79***
	(0.005)
mutual	6.107***
	(0.029)
socID	-0.082***
	(0.007)
sentiment	0.118***
	(0.004)
Panel C: 2022	
edges	-9.415***
	(0.007)
mutual	5.806***
	(0.039)
socID	-0.023*
	(0.009)
sentiment	0.21***
	(0.005)
Panel D: 2023	
edges	-9.046***
	(0.007)
mutual	5.574***
	(0.044)
socID	0.023*
	(0.01)
sentiment	-0.122***
	(0.005)

Panel E: 2024	
edges	-8.941***
	(0.011)
mutual	5.63***
	(0.057)
socID	0.005
	(0.014)
sentiment	-0.153***
	(0.007)

6. Appendix

Table 10: Confusion Matrix

This table presents the confusion matrix for the fine-tuned RoBERTa model. The variable of interest is the *SocID_score*.

Category	Precision	Recall	F1-Score
Anti-social identity	0.991	1.000	0.995
Neutral	0.998	0.934	0.965
Social identity	0.946	0.997	0.971