Reddit users unleashed - understanding user behaviour and their impact on meme stocks *

Simon Trimborn[1,2,3] and Inez Maria Zwetsloot†[4]

[1] *Amsterdam School of Economics, University of Amsterdam, Amsterdam*
[2] *Tinbergen Institute, Amsterdam*
[3] *Department of Management Sciences, City University of Hong Kong, Hong Kong*
[4] *Amsterdam Business School, University of Amsterdam, Amsterdam*

February 29, 2024

Abstract

This study develops a sparse network model and changepoint detection framework to identify drivers and changes in users' posting probability on social networks. With the model, we examine the impact of user behaviour on the Reddit forum Wallstreetbets on the stock price of GameStop (GME). Results show that changepoints in vital users' behaviour significantly predicted GME's integrated volatility with a time lag and non-vital users behaviour immediately, even when controlling for network activity and established metrics measuring influential user impact. This suggests that the change in users behaviour was more important for the GME market frenzy than the joint activity of the Wallstreetbets users. Likewise return dynamics and jump volatility can be explained with change in users behaviour. We conclude that the activity of a small group of vital nodes, impacting the behaviour of other nodes, on Wallstreetbets drove the activity of the GME traders.

Keywords: Influencer Market Impact, Social Network Modelling, Wallstreetbets, Changepoint detection

JEL classification: C55, C58, G12, G14

*The authors are thankful for helpful discussions and comments by the seminar participants at National University of Singapore, Tilburg University, Queens Management University, University of Amsterdam, City University of Hong Kong and Tinbergen Institute. The audiences comments during presentation of the paper at the Spring Research Conference of ASA and at Financial Econometrics meets Machine Learning Conference at Erasmus University Rotterdam are gratefully acknowledged.

†Corresponding author

1 Introduction

The emergence of the internet enabled people all over the globe to connect with each other in ways unseen before. Via social networks, information exchange, discussions, and opinion expression became easier, wider spread, and increasingly instantaneous. With social networks playing an ever increasing role in peoples life, all kinds of discussions take place on forums such as Reddit or Twitter. The impact of Twitter on society and also financial markets was intensively studied from various angles (Gu and Kurov, 2020; Behrendt and Schmidt, 2018; Yang et al., 2015) . However, the impact of Reddit on society and markets is less frequently studied even though the social network is the 6th most visited social media network globally as of August 2023 according to similarweb.com. In the USA, home to the largest financial market, Reddit.com is even the 4th most visited social media network, being more popular than LinkedIn, Tiktok and WhatsApp. Latest since the GameStop (GME) market frenzy in January 2021, which was driven by users of the subreddit 'Wallstreetbets' (WSB), the platform received notable attention in society and academia (Pedersen, 2022; Klein, 2022; Umar et al., 2021).

It is well acknowledged that the actions and discussions on the subreddit WSB on Reddit led to the GME market frenzy, see Schroeder et al. (2021). The effect of the market frenzy was so detrimental, that even a US senate hearing took place, during which they questioned the main actors involved in the market frenzy, including one of the Reddit users hyping GME. In the aftermath of the GME frenzy, it became obvious that the 'internet crowd' represents a new force driving financial markets, see e.g. the study by Pedersen (2022). Indeed attention within a community is well known to impact stock prices (Andrei and Hasler, 2014). Lately the effect of users on stock prices of various kind is so strong and widespread, that it led the U.S. Securities and Exchange Commission (SEC) to introduce a new term for stocks being on the radar of the internet crowd: the 'meme stock phenomenon' (SEC, 2021). The hype surrounding 'meme stocks' leads regularly to herding behaviour and market excesses, impacting asset prices (Aloosh et al., 2021; Costola et al., 2021). The persistent effect from Reddit on financial markets via meme stocks drive the need to analyse the behaviour of users of Reddit forums and quantify their impact on the market.

In this paper we present a network model to study Reddit users posting behaviour as well as changes in their behaviour within a subreddit. The model rests on post and comment level data collected from Reddit as well as studies investigating the developments on Reddit. Research covers topics such as subreddit developments (Thukral et al., 2018), factors for influencers on subreddits (Gianstefani et al., 2022), network metrics (Datta et al., 2017), sentiment (Umar et al., 2021), financial literacy (Klein, 2022), and price impact (Pedersen, 2022). Many of the studies on Reddit, WSB and networks in general rely on aggregated network metrics which provide an overview over the general state of the network, such as network activity, mean distance, mean degree, and number of active nodes, among

others. These aggregated metrics mask interactions on the individual node level, whereas these interactions may contain relevant information to explain the users actions. Yu et al. (2022) show that aggregated network metrics indeed mask changes and interactions happening at the nodal level. Node level analysis allows to study user posting behaviour on Reddit, how they influence each other, and what drives changes in their behaviour over time. Understanding individual users behaviour enables one to identify users vital for the network, study their impact on other users, study and understand what drives the overall changes in the network, which, as it is the case for meme stocks, influences the stock markets.

Analysing individual users influence and changes in the influence, is important for understanding how a network develops (Aral and Walker, 2014; Li and Du, 2011). The model developed in this study accounts for two types of users, vital and non-vital (one may think of this as influencers and non-influencers), to take into account that users posting behaviour is naturally driven by different factors subject to their role in the network. The model predicts the probability for a user to post in the next period subject to their own past posts, in- and out-degree, overall network activity, responses (activity) to their past posts as well as the hour of the day. In the case of non-vital users we add parameters to take into account that they may pay considerable attention to certain vital users (follow them) which may impact the probability of them posting. The model is flexible and allows easily for including other metrics if relevant. When the model is applied to a subreddit focused upon asset discussions such as 'Wallstreetbets', the model is extended with a parameter accounting for the opening hours of the market. Users are not always online hence a time lag between the occurrence of a determinant for posting behaviour can occur with a time lag as well, hence we incorporate time lagged parameters into the model.

Focusing on individual users implies that a tailored model per user is needed to understand the drivers behind their behaviour. Not all parameters included in the model may be relevant for every user behaviour, hence the system may experience sparsity. We observe that the posting frequency on Reddit is sparse, see section 2, which motivates the use of a sparse modelling framework for the user network. We observe that the posting frequency on Reddit is sparse, see section 2, which underscores the stated reasoning and further motivates the use of a sparse modelling framework for the user network. Indeed many networks are characterised by huge dimensionality but have sparse links, de Paula (2017). The high dimensional nature of networks challenges the joint analysis of nodes and identification of important links. Methods from sparse modelling, originally developed for linear models (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006), were increasingly used for network analysis to handle the sparse nature of the underlying models, (Barigozzi and Brownlees, 2019; Trimborn et al., 2022a; Trimborn et al., 2022b; Zhang and Trimborn, 2023). Likewise the network model developed in this study incorporates a LASSO type estimator (Tibshirani, 1996) to account for the observed sparse structure.

It is well known that social networks and user behaviour are highly dynamic over time. The Reddit network model, developed in this paper, has to account for this network property as well. Focusing on the subreddit 'Wallstreetbets', we observe that it went through various stages of activity over time, see section 2. The discussions about GME increased in frequency from summer 2020 and peaked during the market frenzy in January 2021 after which the discussions about GME became less frequent. This dynamic network behavior motivates the existence of changepoints on Reddit. Consequently we present a changepoint detection framework for the model we are introducing, based upon a likelihood based search algorithm as used in Bai et al. (2021).

The most famous and likely best documented case of Reddit users impacting meme stocks, is the GME market frenzy with the Wallstreetbets subreddit at its core. We will motivate and evaluate the developed network model on this subreddit to improve the understanding of the network microstructure, its change over time and the effects on the market. We obtain all the submits (posts and comments) on 'Wallstreetbets' which mention 'GameStop' or 'GME' (the GameStop market ticker) in the title of the corresponding post. We obtain the data for the time period 01.06.2020 until 31.05.2021, whereas the market frenzy took place in January 2021. We study the relationship between the changepoints in the behaviour of vital and non-vital users and the market behaviour of GME. We analyse the effect of users change in behaviour on the integrated and jump volatility which we derive from hourly and 5 minute realized volatility, hourly and 5 minute trading volume as well as the corresponding returns. The results show that the integrated variance is significantly affected by vital users behavioural change with a time lag of 6 hours. Non-vital users significantly impact the integrated variance at various lags as well. As we found that the change in non-vital users behaviour is linked to vital users behavioural change, this translates into network effects impacting the markets. These results are robust when controlling for past integrated variance, aggregated network activity and network statistics such as mean distance, mean degree, number of active nodes and betweenness. It is also robust when accounting for hour of the day effects. Hence we follow that the regular volatility of the GME return (integrated variance) was impacted by the change in behaviour of the network users. Performing a similar analysis on the jump volatility, we find that the network statistics and network activity are a better predictor for the jump volatility than the change in users behaviour. Consequently the users of 'Wallstreetbets' had a persistent impact on the volatility via the integrated variance whereas changes in the overall form of the network (network statistics, activity) impacted the spurious jump volatility instead. Likewise we find the return of the GME series and the change in trading volume being affected by the change in users behaviour.

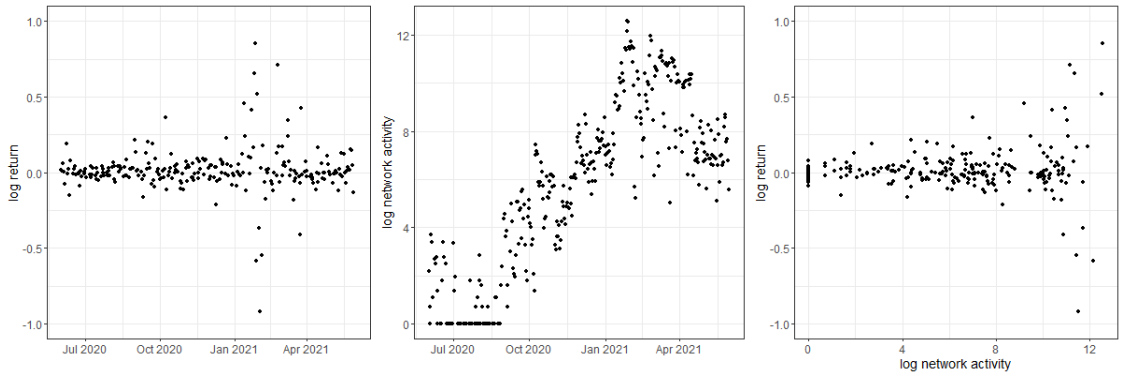
This paper is organised as follows. In Section 2 we introduce the structure of a subreddit on Reddit and present the data for the 'Wallstreetbets' subreddit. In Section 3 we introduce the model which is determined based upon the structure of the Reddit subreddits. We

further introduce the changepoint detection method as well as the Vuong type test (Vuong, 1989) to identify the significant changepoints and the LASSO (Tibshirani, 1996) framework to identify relevant parameters in the model. In Section 4 we present the results from applying the model and changepoint detection framework to the 'Wallstreetbets' subreddit. In Section 5 we link the users change in behaviour to changes in the integrated variance, jump volatility, trading volume and return series of GME, illustrating that the change in behaviour of users in the subreddit predicted changes in the market. In Section 6 we analyse how users change in behaviour impacted the market during the frenzy period, hence we compare users market impact during calm and frenzy periods. Section 7 we test the robustness of the studies settings and section 8 concludes.

2 Wallstreetbets data description

The market frenzy of the GME stock price was driven by users uniting via the forum Wallstreetbets on Reddit. It started with Keith Gill, who is known on Wallstreetbets and Youtube via his accounts 'DeepFuckingValue' and 'RoaringKitty', posting screenshots of his long positions in GME calls on Wallstreetbets. In summer 2020 he also started to analyse the stock on his Youtube channel. Users of Wallstreetbets invested increasingly in GME via trading platforms such as Robinhood while hedge funds were betting on a falling price of GME via short positions. Too many funds shorted GME while Wallstreetbets investors continued to buy the stock and go long in call options, which led to a short squeeze. This situation led to a market frenzy with huge daily volatility and trading volume. The David-against-Goliath hype of WSB users on Reddit led to the bankruptcy of a hedge fund, the suspension of GME trading on the popular Robinhood trading app, and an investigation on by the US COngress into alleged market manipulation. To analyse the relationship between the activity on Wallstreetbets and the GME stock price, we collected the discussions about GME on Wallstreetbets from Reddit (www.reddit.com), an online social network. Reddit is divided into subreddit's, which are specific online communities and the content associated with it is dedicated to a particular topic. For this study we focus on the subreddit 'Wallstreetbets', ie. /r/wallstreetbets, focused on discussions about stock and option trading, which was the central platform for communication related to the GME market frenzy. Individuals submit posts on the subreddit and other members of the network can submit comments to the posts. Then others (including the original post author) can comment on the comments as well. The central point of each subreddit are the posts given they are required to start a conversation. We collect the posts from the Reddit forum 'Wallstreetbets' which have either the keyword 'GME' or 'Gamestop' in the title. In a second step we download the comments from the identified posts. Since Keith Gill started his Youtube channel in summer 2020 and the market frenzy took place in January 2021, we collect one year of data from Wallstreetbets which span the frenzy. Since we model users posting behaviour

Figure 1: Daily GME return (left Figure), daily log number of posts about GME on Wallstreetbets (middle Figure) and scatterplot between the two variables (right Figure)



over time, users which deleted their accounts cannot be analysed as they appear on Reddit as 'anonymous'. We collected 111,331 posts and 3,712,017 comments from 502,650 users over the time period 01.06.2020 until 31.05.2021. We visualize the relationship between the daily GME return and the daily log number of posts and comments about GME on Wallstreetbets in Figure 1. We see that activity started to slowly increase in September 2020 and reach a peak at the end of January 2021, which coincides with the spikes in return in both directions for GME.

2.1 From data to network

The total number of users who post about GME on the Wallstreetbets forum amounts to 502,650 users which are our nodes in the network. To analyse the change in the network, we are focusing upon the communication patterns between the users, hence the edges connecting the nodes. If a user i commented on another user j's post or comment at time point t, this is modeled as a directed communication towards that user. Hence,

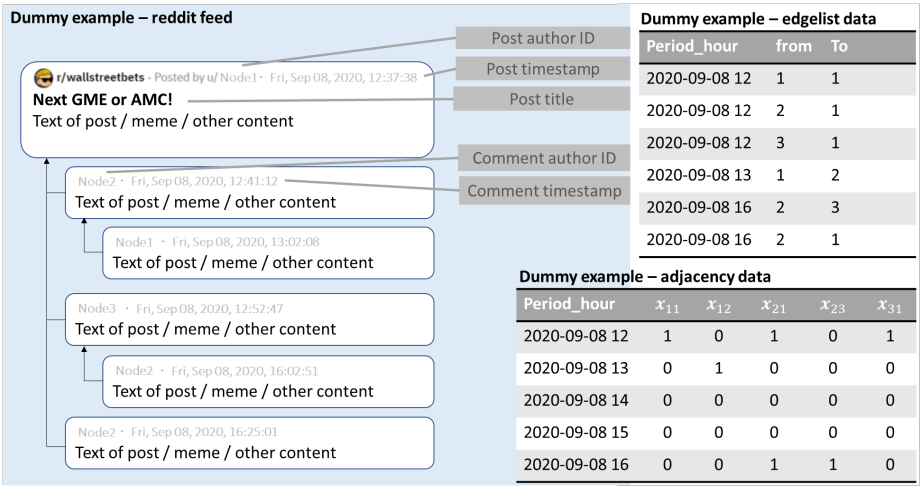
$$x_{ij,t} = \begin{cases} 1 & \text{if } i \text{ sends a comment to } j \text{ at time } t \\ 0 & \text{if } i \text{ does not send a comment to } j \text{ at time } t \end{cases}$$

with $x_{ij,t}$ a binary variable indicating the directed communication between i and j. If a user i comments on their own thread or comment, then this is going to be recorded as a self-communication,

$$x_{ii,t} = \begin{cases} 1 & \text{if } i \text{ posts a thread or a comment to their own thread/comment} \\ 0 & \text{otherwise} \end{cases}$$

Out of the total of 3,823,348 total posts and comments submitted, 3.4% (129,704) are

Figure 2: Dummy example of posts, comments and network data

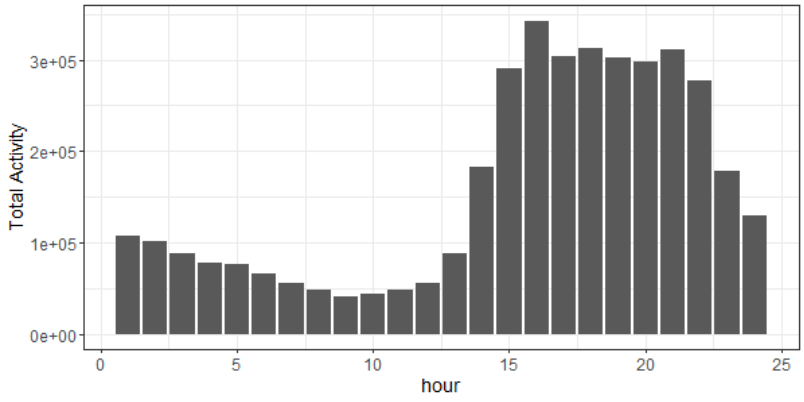


posts and self-comments (ie $x_{ii,t}$), the others 96.6% are comments to other users (ie $x_{ij,t}$ for $i \neq j$). Together $X_t = (x_{ij,t})$ form the $n \times n$ adjacency matrix. We assume that X_t is a binary random matrix subject to conditional probabilities of communication emergence, whereas the matrix features directed communication (ie. non-symmetric) and has the possibility of self-communication (ie. a non-zero diagonal).

In Figure 2 we show an example of a Wallstreetbets conversation stream and how to transform the conversation into network data. On the left side we see that a posts starts the conversation and users can comment on the post. This post has received 5 comments. As by its design, a subreddit requires posts before a conversation can be started based upon the post. We assign unique identifiers to each user and setup the communication network. In this example, the user who submitted the post is denoted as node1. The start of the post is a self-communication of node1 with itself. Node2 and node3 respond directly to the post, hence this is a directed communication from 2 and 3 to 1, where node 2 posts two comments in different hours. Similarly the other posts display a communication from node 1 to 2 and from node 2 to 3.

Since the goal of this study is to identify drivers of posting activity and changes in the network to explain market behaviour, we will split the data into hour-by-hour activity and build network models to predict the probability of edge occurrence between two nodes as well as the probability that a node submits a post. In a second step we will analyse the data 5minute-by-5minute during the main phase of the market frenzy to identify the leading drivers of users posting behaviour as well as the impact of users on each other. By this we will identify differences between users impact on each other and onto the market during calmer and 'frenzy' market situations.

Figure 3: Network activity by hour of the day



3 Network model for microstructure

We are interested in modeling the microstructure of the network to analyse users impact on each other, identify changes in their behaviour and to link these patterns with stock market activity. The most straightforward way of studying microstructure is to consider link prediction. Users posting probability to one another can be triggered by ones own past communication with another node or activity of influential nodes, to name a few, as shown by Guille and Hacid (2012). In addition, strong differences in user activity over time suggest that changes happen in the network, see Figure 3. Therefore, we proposed a model for link prediction conditional on variables such as past node activity, network activity, etc., in the presence of changes of users behaviour:

$$P[x_{ij,t} = 1] = \sum_{s=1}^{S+1} \text{logit}(A_s X + \epsilon_{ij,t}) I(\tau_{s-1} < t \leq \tau_s), \tag{1}$$

with S the number of changepoints that occur at time points τ_s, whereas $1 < \tau_1 < \tau_2 < ... < \tau_S \leq T$ and $I(\cdot)$ is defined as the indicator function. We define $\tau_0 = 1$ and $\tau_{S+1} = T$. Furthermore, $\text{logit}^{-1}(z) = \frac{1}{1+\exp(-z)}$ is the logistic link function applied to a linear combination of features X extracted from the network and A_s is the parameter matrix for time period s. Equipped with the model in equation (1), we study the microstructure of the network. A number of challenges arise when fitting this model to large scale networks like WSB and Reddit in general:

1. Typically subreddits have many nodes, i.e. WSB consists of $n = 502,650$. This means that we have to fit n^2 models.

2. Which variables are reasonable to include in the feature matrix X in order to capture the dynamics of the network microstructure? Additionally it is relevant to account for time dependent relationships between users posting behaviour and the variables

impacting them.

3. How to estimate the number of change points S and their respective locations? How do we prevent overfitting on the changepoints?

4. How do we ensure interpretable modelling results, in order to facilitate an understanding of how the microstructure changes and influences the markets?

In the sequel we will introduce various parts of our proposed methodology, each dealing with one of the mentioned challenges above.

3.1 Dimension reduction: fitting for relevant dyads only

Typically subreddits have thousands or hundred of thousands of nodes, in our WSB dataset there are $502,650$ nodes. For our model, (Eq. (1)), we fit a model for each possible dyad. Most social media networks are sparse and nodes tend to communicate within only a very small subgroup of other nodes. In addition a unique feature of Reddit is that nodes can only send a message in a post. This implies that communication can only happen *after* somebody has submitted a post. In addition, it is well established that networks haven certain users who take stronger roles in the network and are central to its evolvement.

We capitalize on the concepts of model importance to reduce the number of models that we will fit. This is based on the rational that links between less important nodes are less relevant for the networks development. We do this by identifying vital nodes in the network in a data-driven manner and only estimating $P[x_{ij=1}]$ (Eq. (1)) for j who are vital nodes. We reduce the number of models to be fitted further by restricting i to be either a vital node or an admissible non-vital node (ie nodes who communicate regularly within posts of vital nodes). Next, we first define and discuss vital nodes followed by the definition and discussion of admissible non-vital nodes.

3.1.1 Vital nodes (VN)

We identify vital nodes in the network in a data-driven manner. Naturally, they have to be active post submitters as post are required to start communications in subreddits. Of the total number of nodes in the WSB network 16.9% have submitted a post, all other nodes merrily comment on other peoples posts. By the nature of this, one cannot be an influential node in a subreddit if they have never created a post.

Table 1 gives an overview of the top ten nodes with the highest number of posts in the WSBs network. Some nodes submit many posts but receive little to no response on them or did so in a very short period of time. For example, node 17261 submitted a total of 29 posts

Table 1: Top 10 nodes with highest number of posts in the network

Node	No. posts	Median no. comments	Total no. comments	No. of weeks
3827	270	0	0	12
15725	64	0	1167	6
1583	52	30761	1811622	14
16382	47	28	4037	4
4300	45	653	137520	30
1075	38	0	13	2
46870	31	0	11	1
3194	30	0	330	6
13089	30	0	0	2
17261	29	0	1128	2

in two different weeks only and received a total of 1128 comments in the posts with each post receiving a median of 0 comments, measured over the cross section of all posts. A high number of reactions to posts is naturally important for someone who may count as a vital node for a network. From the 10 nodes displayed in Table 1, potentially vital nodes might be 1583 and 4300 provided they are active in 14 and 30 weeks respectively and receive a huge amount of comments to their posts, the others however are unlikely to be influential in the network. Therefore, we set the following definition based upon activity and reactions by the network:

Definition 1. *Vital nodes are those nodes in the network that:*

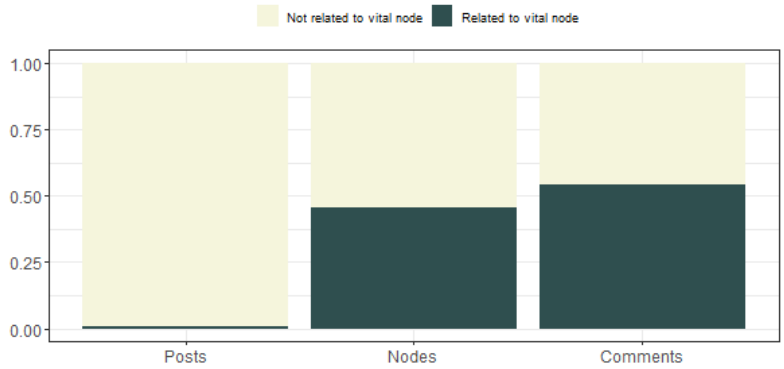
- *Submit posts in at least 10% of the weeks*
- *Receive the highest 10% median comments on their posts*

Notably the definition is relaxed in its nature as it only requires posting activity in 5 weeks. The requirement for the highest 10% of median comments ensures that the users attract high attention relative to the rest of the users in the network. Using definition 1, we identify 64 vital nodes in the network. Among them is the well known Reddit user Keith Gill with Reddit username 'DeepFuckingValue'. In terms of the overall network activity, these 64 vital nodes post only a small part of all posts (0.7%), however nearly half of all nodes in the network have commented in at least one of their posts (45.7%) and over half of all comments is posted in a posts of a vital node (54.4%) as displayed in Figure 4.

3.1.2 Non-vital nodes and admissible non-vital nodes (NVN and ANVN)

The nodes which are not vital, we will classify as non-vital. These nodes can be seen as followers of the influencers (vital nodes). It is well known that many nodes in social networks are very infrequent communicators, as it is the case for the WSB dataset, see Table 2. These

Figure 4: Number of posts, comments and nodes that are related to the 64 identified vital nodes



nodes do not add to the network. For example in our WSB network 71.9% of the nodes have posted less than four comments. During the model estimation, we are interested in an interpretable model as indicated in the challenges for the model construction. The model will feature a high number of parameters which may experience estimation bias given the infrequent communication of NVNs, leading to sparsity of the network. We tackle this issue by estimating model (1) with LASSO, utilizing cross-validation to select the tuning parameters of the LASSO estimation. Cross-validation requires that the dataset can be split into at least 3 subsets, hence at least 3 communications are required to ensure the communication stream of two nodes can be split into the necessary number of subsets for cross-validation. Hence, we distinguish between non-vital and admissible non-vital nodes for estimation purposes, the latter being defined as follows:

Definition 2. *A node will be classified as an* admissible non-vital *node (ANVN) of vital node i when that node has posted sufficient comments in the posts of vital node i to allow for model* (1) *estimation under LASSO and cross-validation.*

Note that this defines ANVNs in relationship to specific vital nodes. We observe that the definition commonly holds when 4 communications took place. This definition allows for a single node to be a follower of multiple vital nodes and it also allows for posting probability estimation between vital nodes.

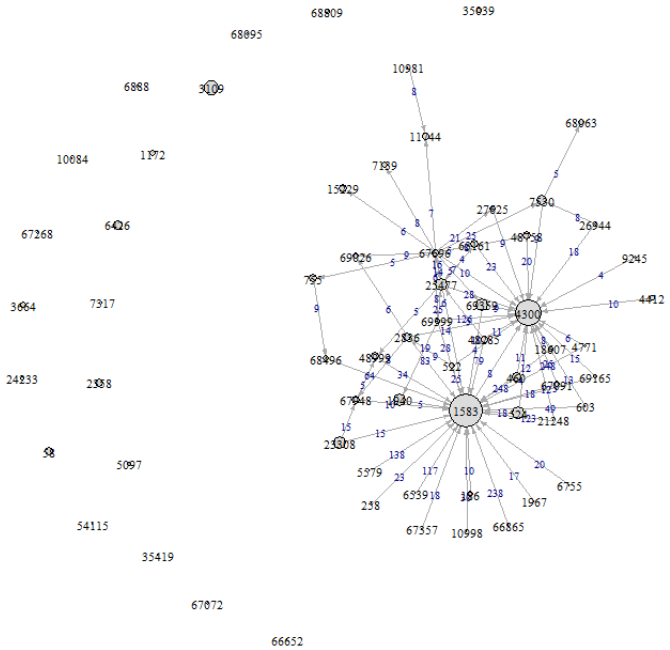
Table 2 gives an overview of the number of admissible non-vital nodes in the network using the above definition. Figure 5 shows the network structure of the 64 vital nodes. The node size represents the number of ANVN for that vital node. The edges indicate if a vital node is an ANVN of the other vital node and the edge weights represent number of communications between the two. It is clear that a large number of vital nodes also follow each other and that there is a large variety in the number of ANVNs per vital node.

Using these definitions for vital nodes (definition 1) and admissible non-vital nodes

Table 2: Nodes and number of ANVN in network

No. nodes	502,650
No. nodes sending at least 1 message	498,979
No. nodes sending at least 4 messages	141,381
No. ANVN	48,213
No. ANVN who follow multiple vital nodes	2778
No. ANVN who are also a vital node	39

Figure 5: Network structure of vital nodes



(definition 2), we define a set of vital nodes S_{VN} and a set of ANVNs, S_{ANVN}. Then we estimate our model (Eq. (1)) for $j \in S_{VN}$ and $i \in S_{VN} \cup S_{ANVN}$. This results in a total of $51,245$ models to estimate, which solves our first challenge.

3.2 Construction of feature matrix X

The second challenge regards the construction of the feature matrix X. Network data have very specific structures that are well studied in the network science literature. For example, topological structures like the in-degree and out-degree are well-known factors used for modeling network links. In addition, it has been shown that overall network activity is an important factor in modeling network influence on markets. Also behavior of people is observed over time and their reaction to posts may have a time lag. To integrate these various factors (microstructure, macrostructure and time lags), we construct our feature matrix by imposing an autoregressive model on a set of network features. In addition, we include a number of control variables.

3.2.1 Micro-structure features

First of all we include the in-degree for node i, this is defined as:

$$D_{i,t}^{in} = \log(\sum_{j,j\neq i} x_{ji,t} + 1), \tag{2}$$

representing the log accumulated posts send to node i by all the users in the network at time point t. The 1 is added to avoid $\log(0)$.

Similarly we include the out-degree for node i obtained as:

$$D_{i,t}^{out} = \log(\sum_{j,j\neq i} x_{ij,t} + 1), \tag{3}$$

representing the log accumulated submits made by node i to all the users in the network at time point t. The 1 is added to avoid $\log(0)$.

As seen from Figure 5, the vital nodes are also connected with each other and each others activity can be the driver behind posting activity. As such, we include a variable measuring the activity of all other vital nodes:

$$AC_{i,t}^{vn} = \log(\sum_{j} \sum_{h\in S_{VN}, h\neq i} x_{hj,t} + 1).$$

We also include the number of comments send to the posts of the vital node i. We

include this in addition to the in-degree as communication within a vital nodes posts but sent as a response to a comment is not recorded as a message to i. This variable is denoted as:

$$AC_{i,t}^{post} = \log(\sum_{j} \sum_{h \neq i} x_{jh,t} I(x_{jh,t} \in \text{post of } i)),$$

representing the log accumulated submits made by all users in the network at time point t to any post of node i, but not directly to i. The 1 is added to avoid log(0).

3.2.2 Macro-structure features

We include the following overall network (macro-structure) features:

$$AC_t = \log(\sum_{i,j} x_{ij,t} + 1), \tag{4}$$

representing the log accumulated communications posted by all the users in the network at time point t. The 1 is added to avoid log(0). Equivalently,

$$AC_k^{24} = \log(\sum_{i,j} x_{ij,t-24*k} + 1), \tag{5}$$

representing the log accumulated communications posted k days prior to time point t. So if we are modeling time point t = 2PM on a Tuesday, than AC_{t-1}^{24} is the activity at 2PM on Monday and AC_{t-2}^{24} is the activity at 2PM on Sunday. The 1 is added to avoid log(0).

We include these overall activities in the model to control for the strong changes in activity, see Figure 1.

3.2.3 Control variables

Further we include a dummy variable for each hour of the day, Dum_{hour}. Also, as we are interested in analysing the network in relation to the financial markets and WSB is traded in New York, we include a dummy variable indicating if the post took place during trading hours of New York Stock Exchange (NYSE), Dum_{wls}.

3.2.4 Constructed feature matrix X

Using the defined features we construct the feature matrices X. We make a distinction between models for vital nodes, i.e. $P[x_{ij} = 1]$, where $i = j$ and $i \in S_{VN}$, and models for admissible non-vital nodes $P[x_{ij} = 1]$, where $i \neq j$ and $j \in S_{VN}$ and $i \in S_{ANVN} \cup S_{VN}$.

Table 3: Feature included in X for the vital node models

Type	Variable	Description	scale
Dependent	$x_{ii,t}$	Posts and self-comments by VN	binary
Micro-structure	$D^{in}_{i,t}$	In-degree of node i: number of submits to node i at time t	log(a+1)
	$D^{out}_{i,t}$	Out-degree of node i: number of submits by node i at time t	log(a+1)
	$AC^{vn}_{i,t}$	Number of submits by all other vital nodes at time t	log(a+1)
	$AC^{post}_{i,t}$	Number of comments within any post of VN i at time t	log(a+1)
Macro-structure	AC_t	Overall network activity at time t	log(a+1)
	AC^{24}_k	Overall network activity at time $t - 24 * k$	log(a+1)
Controls	Dum_{wls}	Dummy indicating trading hours	binary
	Dum_{hour}	24 Dummies for each hour	binary

An overview of the included features in the vital nodes models is given in Table 3. We set up the following 64 logit models, one for each VN. The logit part of Eq. (1) becomes:

$$\begin{aligned}\text{logit}(AX + \epsilon_{ii,t}) = \text{logit}(\theta_0 + \sum_{k=1}^{K}[&\theta_{1,k}x_{ii,t-k} + \theta_{2,k}D^{in}_{i,t-k} + \theta_{3,k}D^{out}_{i,t-k} \\ &+ \theta_{4,k}AC^{vn}_{i,t-k} + \theta_{5,k}AC^{post}_{i,t-k} \\ &+ \theta_{6,k}AC_{t-k} + \theta_{7,k}AC^{24}_k \\ &+ \theta_8 Dum_{wls} + \sum_{h=1}^{24}\theta_{9,h}Dum_{hour,h}] \\ &+ \epsilon_{ii,t}),\end{aligned} \tag{6}$$

where K is the number of lags included in the model.

For the admissible non-vital nodes and their communication with a vital node, the model includes characteristics of the ANVN as well as of the VN. An overview of the included features is given in Table 4. We set up the following 51,245 logit models, one for

Table 4: Feature included in X for the admissible non-vital node models

type	Variable	Description	scale
Dependent	$x_{ij,t}$	Comments send from ANVN to VN	binary
Micro-structure	$D^{in}_{i,t}$	In-degree of ANVN: number of comments send to ANVN at time t	log(a+1)
	$D^{out}_{i,t}$	Out-degree of ANVN: number of comments send by ANVN at time t	log(a+1)
	$D^{in}_{j,t}$	In-degree of VN: number of comments send to VN at time t	log(a+1)
	$D^{out}_{j,t}$	Out-degree of VN: number of submits by VN at time t	log(a+1)
	$AC^{vn}_{j,t}$	Number of submits by all other vital nodes (excluding j) at time t	log(a+1)
	$AC^{post}_{j,t}$	Number of comments within any post of VN j at time t	log(a+1)
Marco-structure	AC_t	Overall network activity at time t	log(a+1)
	AC^{24}_k	Overall network activity at time $t - 24 * k$	log(a+1)
Control	Dum_{wls}	Dummy indicating trading hours	binary
	Dum_{hour}	24 dummies for each hour	binary

each relevant ANVN-VN dyad. The logit part of Eq. (1) becomes:

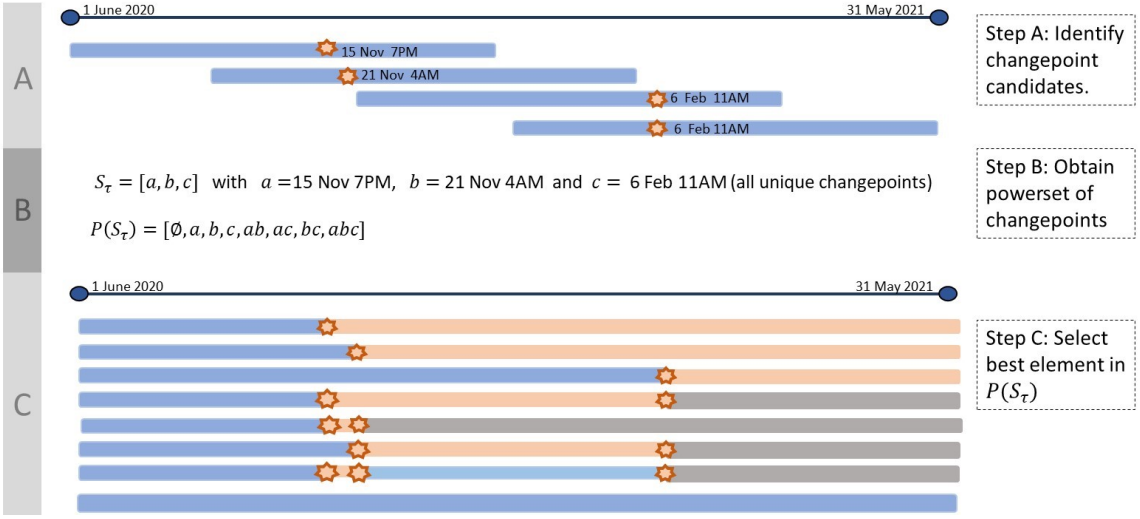
$$\begin{aligned}\text{logit}(AX + \epsilon_{ij,t}) = \text{logit}(\theta_0 + \sum_{k=1}^{K}[&\theta_{1,k}x_{ij,t-k} + \theta_{2,k}D^{in}_{i,t-k} + \theta_{3,k}D^{out}_{i,t-k} \\ &+ \theta_{4,k}D^{in}_{j,t-k} + \theta_{5,k}D^{out}_{j,t-k} \\ &+ \theta_{6,k}AC^{vn}_{j,t-k} + \theta_{7,k}AC^{post}_{j,t-k}] \\ &+ \theta_{8,k}AC_{t-k} + \theta_{9,k}AC^{24}_k \\ &+ \theta_{10}Dum_{wls} + \sum_{h=1}^{24}\theta_{11,h}Dum_{hour,h} \\ &+ \epsilon_{ij,t})\end{aligned} \tag{7}$$

This solves our second challenge.

3.3 Change point identification

The third challenge regards to determination of change points. To identify if and when users change their behavior, we specify a changepoint detection framework for changes within models (6) and (7). The method is based upon a greedy algorithm for searching for multiple changepoints in timeseries data. The central idea is to first identify candidate changepoints by searching for a single change point in a rolling window and then combine these into possible changepoint sequences, as done in e.g. Bai et al. (2021). Our implementation of this general framework is visualized in Figure 6. The algorithm we propose consists of three steps (A-C) and is entirely data driven and is implemented as an exhaustive search algorithm:

Figure 6: A three step algorithm for change points identification.



Step A: Identify changepoint candidates by conducting a rolling window search and identifying single (possible) changepoint in each window.

Step B: Obtain all possible changepoint sequences by constructing a set op unique changepoints and building the powerset.

Step C: Select the best changepoint sequence by estimating a model for each combination of changepoints and selecting the best model and related changepoint sequence.

Next we provide details and the mathematical framework for steps A-C.

3.3.1 Step A: Identify changepoint candidates

In Step A, we conduct a rolling window search over windows of length h. Between the start points of 2 windows we define a step size $l \geq 1$. For a given dataset of length T, G windows of length h with step size l are given (Figure 6 shows four such windows). For $g \in \{1, ..., G\}$, define $T_{1,g}$, $T_{2,g}$ as the start and endpoint of window g respectively. Next we wish to derive the most likely changepoint within each window g. For each possible changepoint τ in window g we estimate models (6) or (7) by minimizing the function $L()$ which is the difference between response and predicted values:

$$L(T_{1,g}, \tau) = \frac{1}{\tau - T_{1,g} + 1} \arg\min_{A} \sum_{t=T_{1,g}}^{\tau} (y_t - I(\text{logit}(AX_t) > 0.5)) \tag{8}$$

$$L(\tau + 1, T_{2,g}) = \frac{1}{T_{2,g} - \tau} \arg\min_{A} \sum_{t=\tau+1}^{T_{2,g}} (y_t - I(\text{logit}(AX_t) > 0.5)), \tag{9}$$

where $L(a, b)$ is related to the model loss based on the data from timepoint a to b. Then, the final changepoint candidate in window g is the minimizer of

$$\hat{\tau}_g = \arg\min_{\tau} L(T_{1,g}, \tau) + L(\tau + 1, T_{2,g}).$$

3.3.2 Step B: Identify possible changepoint sequences

We derive the changepoint candidates for all windows G, which provides us with a set of $\{\hat{\tau}_g\}_1^G$. Define the set S_τ as the set with all unique elements in $\{\hat{\tau}_g\}_1^G$. We next obtain $P(S_\tau)$, the power set of S_τ containing all subsets of S_τ, including the empty set and S_τ itself. In Figure 6, S_τ consists of 3 elements and $P(S_\tau)$ consists of 8 elements.

3.3.3 Step C: Select the best changepoint sequence

Take an element $CPs \in P(S_\tau)$, where CPs is a sequence of changepoints: ie. $CPs = \{cp_1, \ldots, cp_R\}$, where $R = |CPs|$ is the number of changepoints, a non-negative integer. Note that CPs cut the total dataset into $R + 1$ consecutive windows. We denote by T_{1,cp_r} and T_{2,cp_r} the start and endpoint of window r with $r \in \{1, \ldots R\}$. Note that by definition $T_{1,r+1} = T_{2,r} + 1$ as the windows are consecutive and $T_{1,1} = 1$, $T_{2,R+1} = T$. We estimate the model for the full dataset of length T with changepoint set CPs and compute $L(CPs)$, which is obtained as

$$L(CPs) = \sum_{r=1}^{R} L(T_{1,r}, T_{2,r}) + \omega n_{CPs} \tag{10}$$

We repeat this for each element of CPs in $P(S_\tau)$. Note that for the empty element CPs, i.e. no changepoints in the model, equation (10) simplifies to $L(CPs = \emptyset) = L(1, T) + \omega n_{CPs}$. To penalize the $L(\cdot)$ function, we incorporate ωn_{CPs} where ω is the penalty parameter and n_{CPs} is the number of non-zero parameters estimated in our model (ie. the number of non-zero elements in A in models (6) or (7)). As we are interested in a model with good forecasting accuracy, we choose to work with the AIC criterion, hence

$\omega = 2$ (Akaike, 1974). Then, the final changepoint set for this model is the minimizer of

$$\widehat{CPs} = \underset{CPs \in P(S_\tau)}{\arg\min} \; (-2\, L(CPs)) \tag{11}$$

This results in a sequence of changepoints $\widehat{CPs}$ that minimizes the penalized loss over the whole dataset and this solves the third challenge.

3.4 Model selection and interpretability

Our final challenge is related to model interpretability. Currently we have models with many parameters: assume that we have found R changepoints candidates, than we have $R + 1$ submodels and each submodel has either 32 or 34 features as listed in Tables 3 or 4 and these features are included for L lags, hence each model has $(R + 1) * 32 * L$ or $(R+1) * 34 * L$ features resulting in an uninterpretable model with hundreds of parameters. To solve this issue we perform LASSO regularization, Tibshirani (1996), to detect which parameters are relevant for the users posting behaviour. After we identified the relevant parameters, we estimate the models 6 and 7 with the changepoints $\widehat{CPs}$ as estimated in Eq (11) again but only include the parameters selected in the previous step. This step follows the finding of Belloni and Chernozhukov (2013) that a reestimation of the model removes the bias introduced by the LASSO estimation procedure. This solve the fourth challenge.

3.5 Speeding up computational time

Overall the methodology presented in this section is very greedy and hence time consuming to run. In addition to running it only for a subset of nodes, ie the vital nodes and the active-non-vital nodel, we also incorporate the following strategies to speed up computation:

- In step A the algorithm computes the likelihood over the window for each possible changepoint in the window. To speed up this step we first compute the likelihood for each 12th changepoint in the window using the changepoint which yields maximum likelihood we than compute the likelihood for the change point in all 11 steps before and after the identified changepoint. And use the changepoint that give the maximum value as the final result.

- The set S_τ often consists of many possible changepoints resulting in $P(S_\tau)$ having a large number of elements. To prevent overfitting and to speed up computation we exclude all elements of 5 or more changepoints from $P(S_\tau)$. Our final results (see Section 4) show that the maximum number of identified changepoints is 2 so this does not seem a restrictive solution.

- The set S_τ often consists of many possible changepoints that are close together, we therefor restrict the set $P(S_\tau)$ to have only elements for which consecutive changepoint are at k time units apart. From an interpretation point of view it seems reasonable to assume that a Reddit user does not changes his/her behavior to frequently. In addition our data is super sparse so we are also not able to accurately estimate models in such short time spans.

4 Modelling results

To understand the microstructure in the WallStreetBets network we take the network data as described in Section 2 and apply the algorithm as described in Section 3. We split the whole year of data into hour-by-hour networks and report on the modelling results in Section 4.1. In addition, to get a detailed view of the network around the frenzy we also study the network at a 5minute-by-5minute frequency for the period of one month around the frenzy (11 Jan 2021 until 13 Feb 2021) in Section 4.2. Finally in Section 4.3 we study the performance of our model by reporting deviance and accuracy.

4.1 Hour-by-hour modelling results

We apply the models in Equations (6) and (7) to the vital and active non-vital nodes. We combine the models with the changepoint detection method described in Section 3.3. We perform the changepoint selection under the following settings:

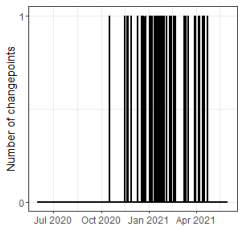
- Window size: $h = 1440$ hours (2 months)
- Stepsize: $l = 168$ hours (1 week)
- 5 Lags
- Time between change points: $k = 168$ (1 week)

We first look at the identified changepoints in our models. Table 5 displays the results, we see that most VN and NVNs change their posting behavior once in the year (84% and 81%) whereas 14% of the VN and 18% of the NVN do not have any changepoints in their models. This also confirms that restricting our changepoint search to maximum 5 changepoints is a reasonable restriction to speed up the computations (see Section 3.5 for detail) as the maximum identified number of changepoints is 3. Figure 7 shows the changepoints of all VN (Fig 7a) and all NVN (Fig 7b) versus time. We see that some VN change their behavior as early as November and December 2020 which is well before the frenzy. On the contrary, most NVNs change their behavior during and after the frenzy.

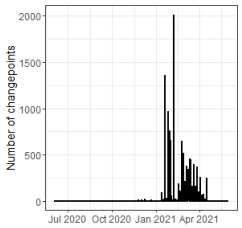
Table 5: Identified changepoints

No. cps	VN	%	NVN	%
0	9	14.1	9376	18.32
1	54	84.4	41551	81.19
2	1	1.6	240	0.47
3	0	0.0	11	0.02
≥ 4	0	0.0	0	0.00

Figure 7: Changepoints over time



(a) Vital Nodes

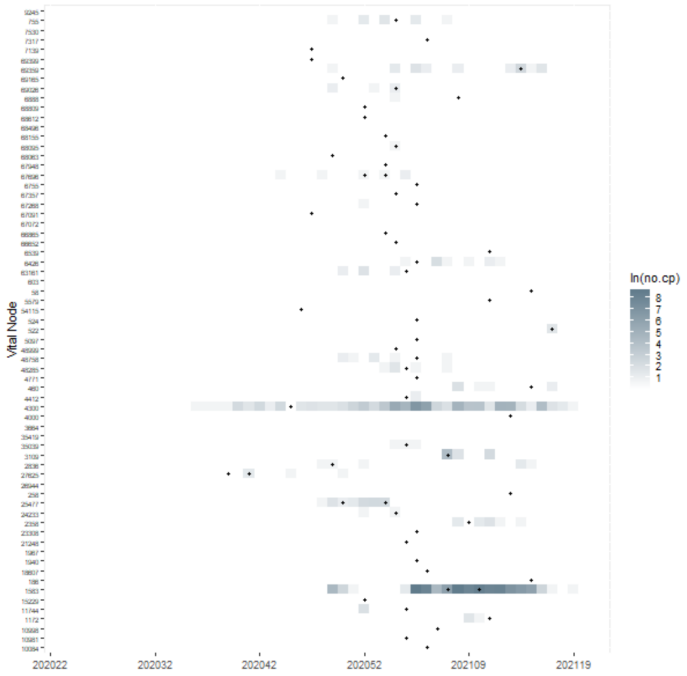


(b) Non Vital Nodes

To study the identified changepoints in more detail we look at each Vital Node in Figure 8. The dots indicate a changepoint of a vital node and we place it in relation to the changepoints of active non-vital nodes which comment in the post of the respective vital nodes. We observe that most of the changepoints which define the behaviour of the nodes takes place right when the market frenzy starts. A few of them even change their behaviour up to 2 months before that. Further changepoints appear a few weeks after the frenzy which would be in line with the rational that after the frenzy ebbed off, users would change their posting behaviour. We observe a particularly large amount of changepoints of ANVNs who comment in the post of a particular VN, 4300. This user is Keith Gill, which shows that his posts indeed triggered changes of users behaviour. The change of users behaviour is the most intense right before and during the market frenzy, which raised the question what exactly triggered the users change in behaviour. Another user also triggered the change of many users behaviour however mostly during and after the market frenzy. This is the founder of Wallstreetbets who regularly submits posts asking users opinions about particular stocks and their plans for investments in the week following. To study in more detail the reaction of NVNs to VNs changepoints we look at the time between the VNs (first) changepoint and the NVNs (first) changepoint. Figure 9 shows a density plot of these times for the vital nodes 4300 and 1583 as well as all the other 62 VNs. We see that especially for VN 4300 (Keith Gill) the change in posting behavior seems to result in NVN changing their posting behavior a while later, with the peak of the distribution at 400 hours (16 days) after VN 4300 changed its behavior. For the other VNs this effects is not apparent in the obtained results.

Next we are going to analyse the parameters of the models. As discussed in Section 3.4 we use LASSO to enforce sparsity in the models we estimate. For the VN and NVN models we show the value of the non-zero parameters in Figure 10. We see that during trading hours the VNs have an increased probability of posting (hr 14 untill hr 20). For the NVNs the effect of hours is more dispersed which is expected as there are many more individuals. Next we look at the dummy for trading hours (Dum_{wls}) both VNs and NVNs have predominantly positive values for this parameter indicating that posting is more likely during opening hours of WallStreet. There is a group of VNs and NVNs that have high values (around 20) for the trading hour dummy, indicating a strong likelihood of posting during trading hours. For the VNs the overall activity in the network seems to have no specific affect on the likelihood of posting, parameter values are centered around zero. This implies that our VN are not letting their posting behavior depend on the overall network activity. On the contrary, for the NVNs the overall network activity has parameter distributions skewed to the right of the zero line indicating a slight increase in posting likelihood of nodes when the overall network is active. As expected D_{out} for the VNs and $D_{out,i}$ for the NVNs is positive for the majority of the estimated parameters, indicating that as users posts they are likely to post again. This jsut indicated that if people are using reddit they are using it to post multiple

Figure 8: Changepoints of vital nodes and active non-vital nodes



messages. However there is also a group of users for which this parameter has a negative value around -20 to -25 indicating that these people tend to post and then probably leave the platform to come back at a later time. Finally we consider for the VNs the effect of X_{ii} which is mostly negative indicating that after submitting a post the influencers are less likely to submit another post. On the contrary, when we consider the NVNs and the parameters of the variable X_{ij} we see the opposite effect, as NVNs submit comments to posts they are more likeli to comment again in the post.

4.2 5-minute-by-5-minute modeling results

For a detailed look of the microstructure of the network around the frenzy we use networks constructed each 5 minutes. We apply the models 6 and 7 to the vital and active non-vital nodes. We combine the models with the changepoint detection method described in section 3.3. We perform the changepoint selection under the following settings:

- Window size: $h = 2016$ 5-minute intervals (1 week)
- Stepsize: $l = 288$ 5-minute intervals (1 day)
- 5 Lags
- Time between change points: $k = 288$ (1 day)

Figure 9: Time between changepoints of vital nodes and change points of active non-vital nodes

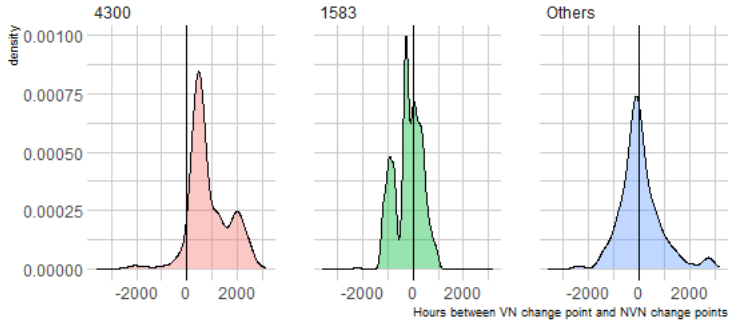


Figure 10: Parameter Estimates hourly model.

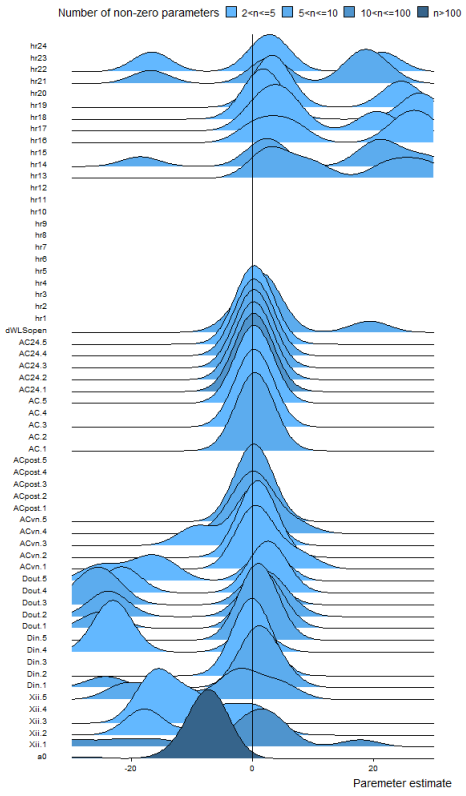
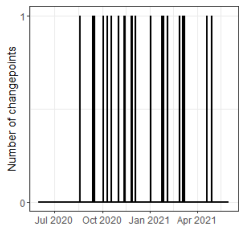


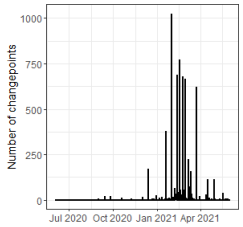
Table 6: Identified changepoints

No. cps	VN	%	NVN	%
0	38	59.4	2895	13.5
1	26	40.6	18566	86.4
2	0	0.0	16	0.1
3	0	0.0	1	0.0
≥ 4	0	0.0	0	0.0

Figure 11: Changepoints over time



(a) Vital Nodes



(b) Non Vital Nodes

We first look at the identified changepoints in our models. Table 7 displays the results, we see that most VNs, 59.4%, do not change their behavior in the period around the frenzy but that many of the NVNs, 86.4% do show a change in posting behavior. Figure 11 shows the changepoints of all VNs (Fig 11a) and all NVNs (Fig 11b) versus time. We see that the VN changepoints are spread out of the month January whereas most NVNs change their behavior towards the end of January.

Next we are going to analyse the parameters of the models. As discussed in Section 3.4 we use LASSO to enforce sparsity in the models we estimate. For the VNs only two vital nodes have non-zero parameters in the estimated models (excluding a non-zero constant), the first VN is 4300 (Keith Gill) and the other is 24233. We see that past own activity reduces likelihood of posting in addition when Wallstreet is open and at hour 21 (just after closing of Wallstreet) the likelihood of posting is increased. All other parameters are rather close to zero.

We also study the parameter estimates for the NVN models and the pattern is very similar to the hourly model as shown in Figure 10 so we don't include the results here.

Table 7: Parameters for VN 4300 in 5 minute model

Variable	Parameter estimate
a0	-46.9
Xii.1	-36.8
Xii.2	-54.0
Xii.3	-51.7
Xii.4	-46.0
Xii.5	-29.5
Din.1	1.9
Din.2	3.6
Din.4	-2.6
Din.5	-3.5
ACvn.1	1.4
ACvn.3	0.2
ACvn.5	-0.8
AC.1	-0.5
AC.5	0.9
AC24.1	-0.002
AC24.2	-0.6
AC24.3	0.5
dWLSopen	18.0
hr21	23.7

4.3 Model deviance and accuracy

We evaluate model deviance and accuracy to understand how well our model fits and predicts our data. Figure 12 shows the percentage decrease in deviance where we compare our model to an intercept model without change points. We only include the models which have at least one significant change-point (ie. 55 VN models out of the 64 and 41802 NVN models out of the 51178). We see that our model decreases the deviance substantially. Figure 13 shows the additional percentage points 1s predicted compared to intercept model without change points. We only consider the predictions of communications (ie. the 1's in our data) as the data is highly unbalanced. For the VNs we see a small group of models for which we predict all communications accurately and overall the accuracy is better for the models with change points than the base line constant model.

5 Impact of Wallstreetbets on GME stock price

We analyse if the users behaviour on Wallstreetbets influence the realized intraday volatility of the GME stock on the markets. From 5-minute price data for GME, we construct the realized volatility for every hour. Provided GME experienced large price swings with

Figure 12: Relative decrease in deviance for our model compared to an intercept model without change points. Left: Vital nodes, Right: Non-vital Nodes

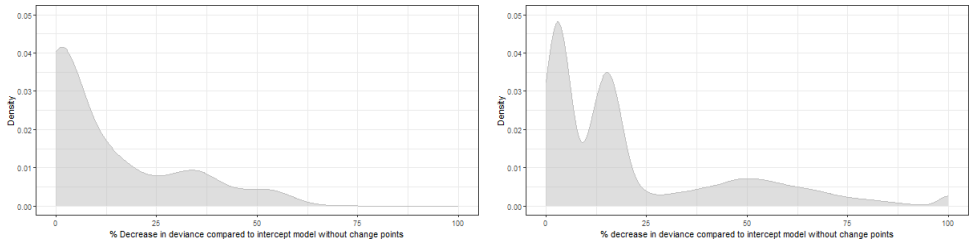
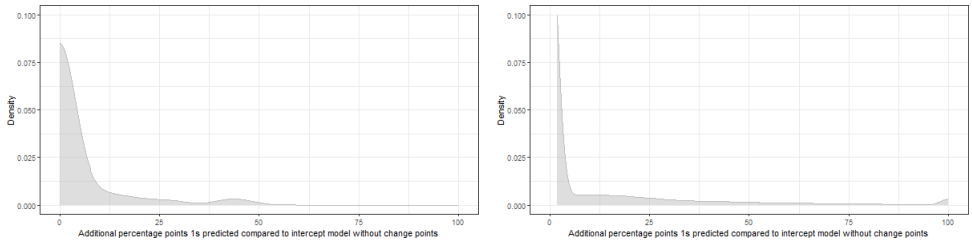


Figure 13: Percent points increase in prediction accuracy for our model compared to an intercept model without change points. Left: Vital nodes, Right: Non-vital Nodes



intraday jumps, we split the volatility into the integrated volatility (IV) and the jump component (JV) with the Andersen et al. (2007) estimation procedure.

We specify the a model for the analysis of IV which included contemporaneous and time-lagged variables related to the network, the changepoints and the IV's own past:

$$IV_t = \alpha_0 + \sum_{k=1}^{K} \alpha_k IV_{t-k} + \sum_{k=0}^{K_1} \beta_k^{VN} VN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{NVN} NVN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{AC} AC_{t-k} + \sum_{k=0}^{K_1} \sum_{i=1}^{M} \beta_k^i x_{t-k}^i + \epsilon_t. \tag{12}$$

where IV_t is the integrated variance of GME, VN_t the accumulated changepoints of vital nodes at time point t, NVN_t the accumulated changepoints of non-vital nodes at time point t, AC_t the network activity at time point t and X is an $T \times M$ matrix consisting of M network related variables which are included as control variables. As variables in X we consider various variables which measure the network activity and importance of nodes. Specifically we include the log mean distance, log mean degree and the log number of active nodes.

Alike to the model (12), we specify the following model for the jump component JV_t:

$$JV_t = \alpha_0 + \sum_{k=0}^{K_1} \beta_k^{VN} VN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{NVN} NVN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{AC} AC_{t-k} + \sum_{k=0}^{K_1} \sum_{i=1}^{M} \beta_k^i x_{t-k}^i + \epsilon_t. \tag{13}$$

The same definitions as before hold. Provided the jumps are assumed to appear independent of each other, we do not include time-lagged JV variables into the model.

We are further interested in the effects of the behaviour of the users of WSB on the return and trading volume of GME. We construct similar models to (12) and (13) for the return (r_t) and trading volume (TV_t):

$$r_t = \alpha_0 + \sum_{k=1}^{K} \alpha_k r_{t-k} + \sum_{k=0}^{K_1} \beta_k^{VN} VN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{NVN} NVN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{AC} AC_{t-k} + \sum_{k=0}^{K_1} \sum_{i=1}^{M} \beta_k^i x_{t-k}^i + \epsilon_t. \tag{14}$$

and

$$TV_t = \alpha_0 + \sum_{k=1}^{K} \alpha_k TV_{t-k} + \sum_{k=0}^{K_1} \beta_k^{VN} VN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{NVN} NVN_{t-k} + \sum_{k=0}^{K_1} \beta_k^{AC} AC_{t-k} + \sum_{k=0}^{K_1} \sum_{i=1}^{M} \beta_k^i x_{t-k}^i + \epsilon_t. \tag{15}$$

The parameters and variables are defined alike to model (12).

We investigate if a change in the users behaviour has an impact on the returns of the GME stock. We consider two different specifications for the impact of the changepoints on the stock performance. We consider the number of changepoints occurring at a time point t of the vital nodes N_{VN} and non-vital nodes N_{NVN}:

$$VN_t = \sum_{i=1}^{N_{VN}} VN_{it} \tag{16}$$

$$NVN_t = \sum_{i=1}^{N_{NVN}} VN_{it}. \tag{17}$$

We observe for the Vital Nodes (VN) and Non-Vital Nodes (NVN) a significant effect on the integrated variance of GME, compare Table 9. We observe that the baseline model, which only contains 7 Lags of the IV variable, provides already a R^2 of 0.175. After adding the network activity (AC), the R^2 rises to 0.189. But only when adding the changepoints of the vital-nodes VN and non-vital nodes NVN, the R^2 increases drastically. Adding VN increases the R^2 to 0.239 with a significant parameter in lag 6. This infers that it takes a few hours until the change in behaviour of vital-nodes is reflected in the integrated variance of the GME stock. This is reasonable given the change in behaviour will be accompanied by a post on Wallstreetbets and it would take a while until sufficient users have seen the post, reacted to it and (potentially) acted based on it on the stock market. For the NVN we see an even more drastic increase in the R^2, it rises to 0.392, and various variables over several lags are significant. Given there are drastically more users in the group of the NVN as the VN, a lot of changepoints occurred around the same time and reflected a change of various users behaviour. Hence it is reasonable that this effect is faster reflected in the integrated variance. It is interesting though that the network activity AC does not proxy the individual users change in behaviour and also that the change in behaviour from vital nodes remains important for the explainability of the integrated variance.

We also investigate if the detection of the changepoints and their explainability for the dynamics of the integrated variance could be proxied by the network measures mean distance, mean degree and number of active nodes. These measures proxy network activity and importance of nodes. However, including them instead of the VN and NVN changepoints, yields a tremendously lower R^2 of 0.255 compared to 0.392 when VN and NVN are included but the network control statistics are not.

As we established, the changepoints have a strong and significant impact on the integrated variance. On the jump component, JV, which is by its nature a short-lived part of the variance, the change in behaviour of users has less of an impact, compare Table 10. The R^2 increases strongly with the exponential weighting specifications for the changes behaviour, however the impact is not as strong as for the network statistics. Given this is the jump component of the variance, this infers that the change in users behaviour is not strongly related to the immediate excesses in the GME variance. However it is strongly related to the integrated part of the variance, hence the regular variance.

We further investigate if not only the variance of GME is explained by the behaviour of users on WSB, but also if the return of GME is determined by it. We observe a persistent increase in terms of R^2 and also adjusted R^2 when adding network activity and the vital nodes changepoints, VN, to the model comprised of the hourly lagged return of GME. The largest increase takes place when the changepoints of the Non-Vital Nodes, NVN, is being added to the model. We observe a R^2 and adjusted R^2 of 0.132 and 0.115 respectively. We observe that all NVN variables, the contemporaneous and lagged ones, are significant and in combination with the drastically increasing model fit, we conclude that the changepoints of the Non-Vital Nodes has high explainability for the GME return evolution. The model containing only the network statistics does not reach achieve an equally high R^2, consequently the VN and NVN indeed explain the return movements of GME better than established network statistics. Adding said network statistics to the model containing also the VN and NVN, achieves an even higher R^2 and adjusted R^2, which suggests that the VN and NVN explain part of the return movements which are not captured by the network statistics, network activity and also not by the past return movements.

Likewise we investigate if the dynamics of the trading volume (TV) of GME is explained by the changepoints of the Vital Nodes (VN) and Non-Vital Nodes (NVN). We observe that the variables embedded in model (15) explain parts of the dynamics of the trading volume. The R^2 and adjusted R^2 reach the best values for the model containing the past trading volume, network activity and VN. Note that the VN parameter is significant whereas none of the AC parameters is significant. We observe that adding the NVN and network statistics causes a decrease in the adjusted R^2, conseqeuntly including them only marginally increases the explainability of the system. As in the previous examinations, we conclude that the changepoints of the vital nodes, VN, explain the dynamics of the trading volume.

6 Network during market frenzy

A particular period of interest for GME is the time spanning the market frenzy. During this time we also observed that a lot more activity took place on WSB, consequently we zoom in on the market frenzy and analyse on a minute-by-minute setting the market movements and their relations to the user behaviour on WSB.

We compute the 5-minute RV from the minute-by-minute returns of GME and construct the IV and JV by the Andersen et al. (2007) method, similar as for the hourly data. We estimate the models with 12 lags, representing the lagged observations of the past hour, and summarise the results in Table 13. We observe that adding the Vital and Non-Vital Nodes (VN, NVN) to the model increases the R^2 and adjusted R^2 strongly. Considering both VN and NVN increases the R^2 to 0.242 whereas a model consisting only of lagged variables spanning the past IV's, provides an R^2 of 0.185. For the VN we observe that the 5 and 20 minute lagged ones are significant, hence the impact from a change in behaviour of the vital nodes during the market frenzy affects the integrated volatility of GME immediately. For the NVNs a change in their behaviour transmits slower to the integrated variance as the 35, 55 and 60 minute lagged ones are significant. Next we analysed if conventional network statistics proxy the explained movements from the VN and NVN. We observe that the respective model achieves a comparable R^2 to the one containing the VNs (but not the NVNs), however a lot more network statistics are needed to achieve this R^2. Consequently the adjusted R^2 is much lower and renders it worse than the model containing only the IVs and ACs. Considering all variables jointly results in an adjusted R^2 lower than the one for the model not containing the network statistics. Overall this suggests that the VN and NVN better explain the dynamics of the integrated variance than the conventional network statistics.

Switching focus towards the jump part of the realised variance, JV, we observe that the models hardly explain the jump dynamics compare Table 14. In fact, the adjusted R^2 is very low, almost zero, or negative for all models, suggesting that the JV are not determined by the WSB dynamics. Similarly to the hourly modelling results, this implies that the long run part of the realised variance, IV, can be explained by the WSB dynamics, whereas the short run part, JV, is not explainable by the network dynamics.

Moving focus towards the 5-minute return dynamics during the market frenzy, we model the dynamics by the past returns, AC, VN, NVN and network statistics, see Table 15. We observe that the return dynamics are best explained by their own past dynamics with a small improvement of the adjusted R^2 when adding the network activity to the regression. However, VN, NVN and the established network statistics do not improve the adjusted R^2. We follow from this that the return dynamics during the frenzy are partly explainable by the WSB network activity, however the behavioural change of the Vital and Non-Vital

Nodes had no significant affect on the returns.

The modelling of the 5-minute trading volume shows further insightful results, compare Table 16. Adding the network activity variable, AC, to the lagged TVs does not improve the adjusted R^2, instead keeps its value similar compared to the model utilising only the lagged TVs. Note that also the lagged TVs provided only a low R^2 of only 0.01, however with several significant parameters. The models containing the VN and NVN result into negative adjusted R^2, however the model consisting of the lagged TVs, ACs and network statistics results in a much better R^2 as well as adjusted R^2. This implies that the WSB network activity does impact the trading volume dynamics however the behavioural changes in Vital and Non-Vital Nodes do not during the market frenzy.

Overall we conclude that the integrated variance is explained by the behavioural changes of the Vital and Non-Vital Nodes whereas the JV is largely unaffected by the network variables and measures. For the return and trading volume dynamics we observe that the WSB network activity and various network measures improve their explainability however VN and NVN do not. This implies that metrics derived from the WSB network do explain the market movements of GMEs stock price and its trading volume, whereas the behavioural changes of the Vital and Non-Vital Nodes explain the integrated variance of GME.

7 Robustness

We perform 2 kind of robustness studies to validate our results. First we perform for the definition of vital nodes a robustness analysis to validate definition 1 where we define vital nodes. The definition states that vital nodes *are those nodes in the network that submit posts in at least 10% of the weeks and receive the highest 10% median comments on their posts.* This is a relaxed definition as it only requires posting activity in 5 weeks. To validate this definition we also rerun the full analysis while requiring the VNs to post in at least 15% and 20% of the weeks. In addition the 10% median comments is also validated by requiring the highest 5% median comments, hence we make this condition stricter. We perform the robustness analysis for each combination of the stated values. Table 8 shows for each of these values how many VNs are identified in the network. The top row shows the values as used Definition 1 which resulted in 64 VNs and 48211 NVNs. As we become stricter in our definition we see that the number of identified VNs declines, the number of NVNs only slightly declines. Note that we do not consider more relaxed definitions (ie % of weeks ¡10 and % median comments ¿10) as the current setting is already very relaxed.

We validate the robustness of the results in Tables 21, 22, 23 and 24. These tables show the R^2 and adj. R^2 for the models 12, 13, 14 and 15 when a different definition for the VN identification was chosen, which impacts also the modelling of links with the

Table 8: Robustness VN definition

% of the weeks	% median comments	No. of VN	No. of NVN
10	10	64	48211
10	5	47	48149
15	10	16	48021
15	5	14	48018
20	10	9	47993
20	5	9	47993

respective ANVNs, which consequently impacts the amount of the changepoints of the VNs and NVNs. Hence we validate if the change in the number of changepoints affects the results of this study. Each table contains the setting 'Robustness: 10-10' which is the base setting used in this study. Each table and analysis conveys the observation that the results are robust. Regardless of the definition for the vital nodes, we observe that the identified changepoints strongly improve the modelling of the return, integrated variance, jump volatility and trading volume. Notably we observe that for several instances, the base setting of this study, '10-10', leads to a lower R^2 and adj. R^2 than other settings. Hence we conclude that the results of the study are robust towards the VN definition and that the chosen definition is a conservative choice, which is supported by the fit of the regressions displayed in Tables 21, 22, 23 and 24.

In a second robustness analysis we investigate if the results of our study hold up when we focus the analysis upon the trading ours of NYSE and compute one model per trading hour. As there are 7 trading hours, this leads to 7 models for each of the time series to be analysed, namely return, integrated volatility, jump volatility and trading volume time series. The results of these analysis are displayed in Tables 17, 18, 19, and 20. Overall the results indicate that the change in posting behaviour of the VNs and ANVNs plays an even stronger role when focusing upon the individual trading hours. Across the hours and data series, the R^2 and adj. R^2 are much higher. For the return series, Table 17, we observe that for all hours but the first and last one, a higher proportion of the variability is explained by the changepoints of VN and NVN compared to the network statistics. In all cases considering the changepoints and network statistics improves the R^2 and adj. R^2 the most. Notably the models strongly outperform a standard autoregressive model. For hour 5, an AR(7) model only yields 0.1 adj R^2 but the model considering AC and the changepoints of VN and NVN yields 0.48 adj R^2.

In terms of modelling the IV and JV by the trading hours, similar results as for the entire time series are observed but with much higher R^2 and adj. R^2, Table 18. For various hours, the changepoints of the VN and ANVN add so much explainability of the variability to the base model, that the network statistics only explain very little, consider hours 4, 5, 6, or 7. For hour 4 the changepoints of the VN and ANVN explain such a large proportion of the

variability, $R^2 = 0.9$, whereas the model with the network statistics only explains $R^2 = 0.67$, that it can be concluded the change in behaviour of the users on WSB explain almost the entire variability of the IV. For the JV, Table 19, we observed that the changepoints of the VN and ANVN add a lot more explainability to the variability of the time series in the per trading hour analysis compared to the entire time series. For the latter, modelling the JV yielded $R^2 = 0.129$ for the model consisting of changepoints, network statistics and activity. For the per trading hour analysis, all most explain at least 41% of the variability of the system, for hour 4 even 86% of the variability are explained by the changepoints, network statistics and activity. The model containing only network activity and changepoints of VN and ANVN explain in most cases, but hour 1, 6, 7, more of the JVs variability than the network statistics. For the trading volume, Table 20, we observe similarly to the modelling of the entire time series, that the network statistics and changepoints do not explain the variability of it well. For hours 1-4, the adj. R^2 is even frequently negative. For the hours 5-7 the results are better and we observe that the adj. R^2 is better for the model containing only network activity and the changepoints of the VN and ANVN compared to the model with the network statistics.

Overall these results show that the change in behaviour from the users on WSB, measured by the changepoints of the VN and ANVN, explain a large proportion of the variability of the return series, integrated variance, jump volatility and to some extend also the trading volume series. The results also show that during market trading hours, an even larger proportion of the affected stocks variability can be explained. The latter result is in line with our observation that the hour of the day and trading hours have an effect upon the WSB users posting behaviour, compare section 4.

8 Conclusion

In this study we investigate the posting behaviour of the users on Wallstreetbets and study the impact of their behavioural change on the price of GameStop (GME). We develop a sparse network model and propose a changepoint detection framework to analyse the change in behaviour of users dependent on general network metrics, the community behaviour and the posting behavior of other users. We observe that vital users (nodes) posting behaviour was driven by the network activity before the market frenzy but they changed to be driven by their own posting activity during and after the frenzy. Remaining network participants show little exposure to the networks activity but are impacted by the vital nodes activities, in particular from the onset of the market frenzy. We observe network impact originating from several users (nodes) who we term vital-nodes.

After identifying the drivers behind a users activities and the changepoints, we study the impact of the users change in behaviour on the GME stock price. We find that the

changes in behaviour of vital users (nodes) in the network explain the changes in the realized volatility with a strong impact on the integrated volatility of GME. The vital users behavioural changes significantly predicts realized volatility even when controlling for aggregated network activity and established network measures proxying influential user impact such as mean degree and distance, hence focusing on individual users behaviour outperforms measures aggregating network activity. We observe that it takes a few hours until the change in behaviour of vital-nodes is reflected in the integrated variance of the GME stock. This is reasonable given the change in behaviour will be accompanied by a post on Wallstreetbets and it would take a while until sufficient users have seen the post, reacted to it and (potentially) acted based on it on the stock market. Also the changes in behaviour from the non-vital users significantly impacts the changes in integrated variance. Given there are drastically more users in the group of the NVN as the VN, a lot of changepoints occurred around the same time and reflected a change of various users behaviour. Hence it is reasonable that this effect is faster reflected in the integrated variance.

Given we observe that the developed network model outperforms established metrics for measuring influential user impact, we conclude that the proposed methodology provides a viable framework to investigate a nodes behaviour over time. From the modelling results we conclude that changes in the activity on Wallstreetbets originates from vital users behavioural change which then drove the integrated volatility of GME.

References

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Aloosh, A., H.-E. Choi, and S. Ouzan (2021). "Meme stocks and herd behavior". In: *Available at SSRN 3909945*.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007). "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility". In: *The review of economics and statistics* 89.4, pp. 701–720.

Andrei, D. and M. Hasler (2014). "Investor Attention and Stock Market Volatility". In: *The Review of Financial Studies* 28.1, pp. 33–72.

Aral, S. and D. Walker (2014). "Tie strength, embeddedness, and social influence: A large-scale networked experiment". In: *Management Science* 60.6, pp. 1352–1370.

Bai, P., A. Safikhani, and G. Michailidis (2021). "Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models". In: *Working paper*.

Barigozzi, M. and C. Brownlees (2019). "NETS: Network estimation for time series". In: *Journal of Applied Econometrics* 34.3, pp. 347–364.

Behrendt, S. and A. Schmidt (2018). "The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility". In: *Journal of Banking & Finance* 96, pp. 355–367.

Belloni, A. and V. Chernozhukov (2013). "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2, pp. 521–547.

Costola, M., M. Iacopini, and C. R. Santagiustina (2021). "On the 'mementum' of meme stocks". In: *Economics Letters* 207, p. 110021.

Datta, S., C. Phelan, and E. Adar (2017). "Identifying misaligned inter-group links and communities". In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW, pp. 1–23.

de Paula, Á. (2017). "Econometrics of Network Models". In: *Advances in Economics and Econometrics: Eleventh World Congress*. Ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson. Cambridge University Press, pp. 268–323.

Fan, J. and R. Li (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360.

Gianstefani, I., L. Longo, and M. Riccaboni (2022). "The echo chamber effect resounds on financial markets: A social media alert system for meme stocks". In: *arXiv preprint arXiv:2203.13790*.

Gu, C. and A. Kurov (2020). "Informational role of social media: Evidence from Twitter sentiment". In: *Journal of Banking & Finance* 121, p. 105969.

Guille, A. and H. Hacid (2012). "A predictive model for the temporal dynamics of information diffusion in online social networks". In: *Proceedings of the 21st international conference on World Wide Web*, pp. 1145–1152.

Klein, T. (2022). "A note on GameStop, short squeezes, and autodidactic herding: An evolution in financial literacy?" In: *Finance Research Letters* 46, p. 102229.

Li, F. and T. C. Du (2011). "Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs". In: *Decision support systems* 51.1, pp. 190–197.

Pedersen, L. H. (2022). "Game on: Social networks and markets". In: *Journal of Financial Economics* 146.3, pp. 1097–1119.

Schroeder, P., S. Herbst-Bayliss, and J. McCrank (2021). "Long, tense with cat photo for relief; how the GameStop hearing unfolded". In: *Reuters*.

SEC (2021). *Staff report on equity and options market structure conditions in early 2021*. Tech. rep. US Securities and Exchange Commission.

Thukral, S., H. Meisheri, T. Kataria, A. Agarwal, I. Verma, A. Chatterjee, and L. Dey (2018). "Analyzing behavioral trends in community driven discussion platforms like reddit". In: pp. 662–669.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Trimborn, S., Y. Chen, and R.-B. Chen (2022a). "Influencers, inefficiency and fraud-The Bitcoin price discovery network under the microscope". In: *Available at SSRN*.

Trimborn, S., H. Peng, and Y. Chen (2022b). "Influencer Detection Meets Network AutoRegression–Influential Regions in the Bitcoin Blockchain". In: *Available at SSRN*.

Umar, Z., M. Gubareva, I. Yousaf, and S. Ali (2021). "A tale of company fundamentals vs sentiment driven pricing: The case of GameStop". In: *Journal of Behavioral and Experimental Finance* 30, p. 100501.

Vuong, Q. H. (1989). "Likelihood ratio tests for model selection and non-nested hypotheses". In: *Econometrica: journal of the Econometric Society*, pp. 307–333.

Yang, S. Y., S. Y. K. Mo, and A. Liu (2015). "Twitter financial community sentiment and its predictive relationship to stock market movement". In: *Quantitative Finance* 15.10, pp. 1637–1656.

Yu, L., I. M. Zwetsloot, N. T. Stevens, J. D. Wilson, and K. L. Tsui (2022). "Monitoring dynamic networks: A simulation-based strategy for comparing monitoring methods and a comparative study". In: *Quality and Reliability Engineering International* 38.3, pp. 1226–1250.

Zhang, K. and S. Trimborn (2023). "Influential assets in Large-Scale Vector AutoRegressive Models". In: *SSRN Working paper*.

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429.

Table 9: Significant variables from models 13 with differing variables include and Lag length 7. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:					
	IV					
	(1)	(2)	(3)	(4)	(5)	(6)
IV1	0.322***	0.308***	0.301***	0.255***	0.266***	0.238***
	(0.025)	(0.025)	(0.025)	(0.026)	(0.025)	(0.026)
IV2	0.120***	0.107***	0.102***	0.205***	0.100***	0.182***
	(0.026)	(0.026)	(0.025)	(0.026)	(0.026)	(0.026)
IV3	−0.043*	−0.052**	−0.046*	0.024	−0.054**	0.004
	(0.026)	(0.026)	(0.025)	(0.029)	(0.026)	(0.029)
IV4	−0.005	−0.013	−0.011	−0.137***	−0.040	−0.165***
	(0.026)	(0.026)	(0.025)	(0.031)	(0.026)	(0.031)
IV5	0.043	0.034	0.028	0.036	0.039	0.040*
	(0.026)	(0.026)	(0.025)	(0.023)	(0.026)	(0.024)
IV6	0.113***	0.104***	0.107***	0.125***	0.050*	0.088***
	(0.026)	(0.026)	(0.025)	(0.023)	(0.026)	(0.023)
VN6			0.044***	0.033***		0.032***
			(0.004)	(0.004)		(0.004)
VN7			0.006	0.009**		0.009**
			(0.005)	(0.004)		(0.004)
NVN				0.0002***		0.0002***
				(0.00004)		(0.00004)
NVN1				0.0002***		0.0002***
				(0.0001)		(0.0001)
NVN2				0.001***		0.001***
				(0.0001)		(0.0001)
NVN3				−0.001***		−0.001***
				(0.0001)		(0.0001)
NVN4				−0.0004***		−0.0003***
				(0.0001)		(0.0001)
NVN5				0.0003**		0.0003*
				(0.0001)		(0.0001)
NVN6				0.0002		0.00004
				(0.0002)		(0.0002)
NVN7				0.0003***		0.0003***
				(0.0001)		(0.0001)
Constant	0.001***	−0.002**	−0.001**	−0.0005	−0.0004	−0.0001
	(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vital nodes	No	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	No	Yes	No	Yes
Network activity	No	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	No	Yes	Yes
Observations	1,633	1,633	1,633	1,633	1,633	1,633
R^2	0.175	0.189	0.239	0.392	0.255	0.435
Adjusted R^2	0.172	0.182	0.228	0.380	0.237	0.416
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 10: Significant variables from models 13 with differing variables include and Lag length 7. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:				
	JV				
	(1)	(2)	(3)	(4)	(5)
VN6		-0.019^{***}	-0.015^{***}		-0.018^{***}
		(0.005)	(0.005)		(0.005)
NVN1			-0.0002^{**}		-0.0002^{**}
			(0.0001)		(0.0001)
NVN2			-0.0002^{**}		-0.0001
			(0.0001)		(0.0001)
NVN3			0.0003^{***}		0.0002
			(0.0001)		(0.0001)
NVN6			0.00003		-0.0001
			(0.0002)		(0.0002)
Constant	-0.001	-0.001^{*}	-0.002^{**}	-0.002^{**}	-0.002^{**}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Vital nodes	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	Yes	No	Yes
Network activity	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	Yes	Yes
Observations	1,633	1,633	1,633	1,633	1,633
R^2	0.022	0.032	0.060	0.102	0.129
Adjusted R^2	0.017	0.022	0.046	0.084	0.103
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 11: Significant variables from models 13 with differing variables include and Lag length 7. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:					
	vol					
	(1)	(2)	(3)	(4)	(5)	(6)
vol2	−0.076***	−0.076***	−0.078***	−0.075***	−0.079***	−0.078***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
vol7	0.050**	0.049**	0.050**	0.052**	0.042*	0.045*
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
VN2			1.221***	1.243***		1.258***
			(0.243)	(0.243)		(0.245)
NVN2				−0.00004		−0.0002
				(0.001)		(0.002)
NVN6				0.001**		0.001*
				(0.0004)		(0.0004)
Constant	0.046**	0.057	0.057*	0.055	0.057	0.062
	(0.020)	(0.035)	(0.035)	(0.035)	(0.042)	(0.042)
Vital nodes	No	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	No	Yes	No	Yes
Network activity	No	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	No	Yes	Yes
Observations	1,633	1,633	1,633	1,633	1,633	1,633
R^2	0.012	0.014	0.031	0.035	0.026	0.046
Adjusted R^2	0.007	0.005	0.017	0.016	0.002	0.013
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 12: Significant variables from models 13 with differing variables include and Lag length 7. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:					
	IV					
	(1)	(2)	(3)	(4)	(5)	(6)
IV1	0.115***	0.101***	0.105***	0.105***	0.087***	0.094***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
IV2	0.161***	0.147***	0.136***	0.135***	0.135***	0.125***
	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)
IV3	0.097***	0.081***	0.088***	0.081***	0.075***	0.076***
	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)	(0.026)
IV4	0.167***	0.154***	0.149***	0.152***	0.145***	0.144***
	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)
IV10	0.058**	0.060**	0.047*	0.052**	0.055**	0.050*
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.026)
IV11	0.060**	0.064**	0.064**	0.082***	0.060**	0.078***
	(0.026)	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)
IV12	0.073***	0.073***	0.081***	0.050**	0.070***	0.048*
	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)	(0.025)
VN1			0.002**	0.002**		0.002**
			(0.001)	(0.001)		(0.001)
VN4			0.004***	0.004***		0.004***
			(0.001)	(0.001)		(0.001)
VN10			−0.0003	−0.0002		−0.0002
			(0.001)	(0.001)		(0.001)
VN11			−0.0002	−0.0003		−0.0003
			(0.001)	(0.001)		(0.001)
VN12			−0.0002	−0.0001		−0.0002
			(0.001)	(0.001)		(0.001)
NVN1				−0.00000		−0.00000
				(0.00000)		(0.00000)
NVN4				−0.00000		−0.00000
				(0.00000)		(0.00000)
NVN7				0.00001***		0.00001***
				(0.00000)		(0.00000)
NVN10				−0.00000		−0.00000
				(0.00000)		(0.00000)
NVN11				0.00003***		0.00003***
				(0.00001)		(0.00001)
NVN12				0.00001**		0.00001**
				(0.00000)		(0.00000)
Constant	0.0003***	−0.0001	−0.0001	−0.0001	0.001	0.001
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.002)	(0.002)
Vital nodes	No	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	No	Yes	No	Yes
Network activity	No	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	No	Yes	Yes
Observations	1,556	1,556	1,556	1,556	1,556	1,556
R^2	0.185	0.207	0.222	0.242	0.225	0.259
Adjusted R^2	0.179	0.194	0.202	0.217	0.192	0.214
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 13: Significant variables from models 13 with differing variables include and Lag length 12. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:				
	JV				
	(1)	(2)	(3)	(4)	(5)
NVN11			0.0002***		0.0002**
			(0.0001)		(0.0001)
Constant	−0.001	−0.001	−0.0002	−0.004	−0.007
	(0.001)	(0.001)	(0.001)	(0.010)	(0.010)
Vital nodes	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	Yes	No	Yes
Network activity	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	Yes	Yes
Observations	1,556	1,556	1,556	1,556	1,556
R^2	0.010	0.010	0.020	0.030	0.038
Adjusted R^2	0.002	−0.007	−0.005	−0.004	−0.013
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 14: Significant variables from models 13 with differing variables include and Lag length 12. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:					
	ret					
	(1)	(2)	(3)	(4)	(5)	(6)
ret2	−0.145***	−0.150***	−0.155***	−0.132***	−0.153***	−0.134***
	(0.047)	(0.047)	(0.047)	(0.049)	(0.048)	(0.049)
ret3	−0.260***	−0.251***	−0.251***	−0.217***	−0.257***	−0.224***
	(0.047)	(0.047)	(0.048)	(0.050)	(0.048)	(0.051)
ret4	−0.180***	−0.176***	−0.174***	−0.158***	−0.174***	−0.158***
	(0.053)	(0.053)	(0.053)	(0.054)	(0.054)	(0.055)
ret5	−0.038	−0.044*	−0.046*	−0.042	−0.048*	−0.045*
	(0.026)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)
ret6	0.047	0.056	0.057	0.078*	0.062	0.084*
	(0.043)	(0.043)	(0.044)	(0.045)	(0.044)	(0.045)
ret10	0.073***	0.070***	0.071**	0.063**	0.065**	0.059**
	(0.027)	(0.027)	(0.027)	(0.028)	(0.028)	(0.028)
NVN7				−0.0001**		−0.0001**
				(0.00003)		(0.00003)
Constant	0.002*	0.0003	0.0003	0.001	−0.018	−0.020
	(0.001)	(0.002)	(0.002)	(0.002)	(0.021)	(0.021)
Vital nodes	No	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	No	Yes	No	Yes
Network activity	No	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	No	Yes	Yes
Observations	1,582	1,582	1,582	1,582	1,582	1,582
R^2	0.057	0.066	0.069	0.078	0.075	0.087
Adjusted R^2	0.050	0.051	0.046	0.047	0.036	0.032
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 15: Significant variables from models 13 with differing variables include and Lag length 12. Insignificant variables are not displayed. Control variables (network statistics and network activity) are included for differing models.

	Dependent variable:					
	vol					
	(1)	(2)	(3)	(4)	(5)	(6)
vol3	−0.028	−0.032	−0.033	−0.035	−0.034	−0.038*
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
vol5	−0.040*	−0.045**	−0.048**	−0.046**	−0.046**	−0.046**
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
vol9	0.057***	0.054**	0.054**	0.056**	0.056**	0.057**
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)
Constant	0.003	0.086*	0.087*	0.099**	0.716	0.656
	(0.018)	(0.045)	(0.045)	(0.046)	(0.470)	(0.475)
Vital nodes	No	No	Yes	Yes	No	Yes
Non-vital nodes	No	No	No	Yes	No	Yes
Network activity	No	Yes	Yes	Yes	Yes	Yes
Network statistics	No	No	No	No	Yes	Yes
Observations	1,582	1,582	1,582	1,582	1,582	1,582
R^2	0.010	0.018	0.021	0.027	0.053	0.061
Adjusted R^2	0.003	0.003	−0.003	−0.006	0.013	0.004
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 16: Return: This table shows that the R^2 and adj. R^2 for the return prediction models regressed onto the observations as by trading hours. The results show that regardless of the trading hour, the changepoints of the VN and NVN strongly improve the modelling of the return series. Notably the R^2 and adj. R^2 are much higher in the hourly regressions than in the models for the entire time series.

		(1)	(2)	(3)	(4)	(5)	(6)
Hour 1	R^2	0.07	0.09	0.1	0.31	0.33	0.49
	adj. R^2	0.03	-0.01	-0.01	0.17	0.1	0.25
Hour 2	R^2	0.1	0.13	0.42	0.58	0.41	0.67
	adj. R^2	0.08	0.08	0.36	0.52	0.29	0.57
Hour 3	R^2	0.29	0.3	0.38	0.46	0.42	0.55
	adj. R^2	0.26	0.25	0.31	0.38	0.31	0.41
Hour 4	R^2	0.52	0.52	0.62	0.65	0.59	0.71
	adj. R^2	0.5	0.49	0.58	0.6	0.51	0.62
Hour 5	R^2	0.13	0.16	0.29	0.55	0.23	0.61
	adj. R^2	0.1	0.1	0.21	0.48	0.07	0.49
Hour 6	R^2	0.16	0.17	0.32	0.36	0.32	0.46
	adj. R^2	0.13	0.11	0.24	0.26	0.17	0.29
Hour 7	R^2	0.18	0.2	0.26	0.29	0.33	0.44
	adj. R^2	0.15	0.14	0.18	0.19	0.19	0.26
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 17: Integrated variance: This table shows the R^2 and adj. R^2 for the integrated variance prediction models regressed onto the observations as by trading hours. The results show that regardless of the trading hour, the changepoints of the VN and NVN strongly improve the modelling of the integrated variance. Notably the R^2 and adj. R^2 are much higher in the hourly regressions than in the models for the entire time series. For Hour 4 we observe a particularly strong improvement from the VN and NVN changepoints.

		(1)	(2)	(3)	(4)	(5)	(6)
Hour 1	R^2	0.53	0.56	0.59	0.68	0.65	0.74
	adj. R^2	0.51	0.51	0.54	0.61	0.53	0.61
Hour 2	R^2	0.65	0.67	0.68	0.84	0.74	0.9
	adj. R^2	0.64	0.65	0.65	0.82	0.68	0.87
Hour 3	R^2	0.62	0.63	0.69	0.76	0.72	0.82
	adj. R^2	0.61	0.61	0.66	0.72	0.66	0.77
Hour 4	R^2	0.52	0.53	0.69	0.9	0.67	0.91
	adj. R^2	0.51	0.5	0.65	0.88	0.6	0.89
Hour 5	R^2	0.6	0.61	0.64	0.8	0.73	0.85
	adj. R^2	0.59	0.58	0.6	0.78	0.68	0.8
Hour 6	R^2	0.78	0.8	0.81	0.89	0.85	0.91
	adj. R^2	0.77	0.79	0.79	0.87	0.82	0.88
Hour 7	R^2	0.69	0.71	0.72	0.82	0.75	0.85
	adj. R^2	0.68	0.68	0.69	0.79	0.7	0.8
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 18: Jump volatility: This table shows the R^2 and adj. R^2 for the jump volatility prediction models regressed onto the observations as by trading hours. The results show that regardless of the trading hour, the changepoints of the VN and NVN strongly improve the modelling of the jump volatility. We observe that contrary to the regression onto the entire time series, the changepoints of VN and NVN strongly improve the modelling of the jump volatility and explains parts the network statistic cannot. This observation is particularly pronounced for Hours 2-5, so the morning and midday trading hours. Further, the R^2 and adj. R^2 are much higher in the hourly regressions than in the models for the entire time series.

		(1)	(2)	(3)	(4)	(5)
Hour 1	R^2	0.14	0.15	0.19	0.34	0.41
	adj. R^2	0.1	0.09	0.09	0.17	0.18
Hour 2	R^2	0.14	0.14	0.38	0.37	0.56
	adj. R^2	0.11	0.09	0.31	0.27	0.45
Hour 3	R^2	0.06	0.1	0.31	0.23	0.41
	adj. R^2	0.03	0.04	0.23	0.11	0.26
Hour 4	R^2	0.04	0.49	0.84	0.44	0.86
	adj. R^2	0	0.45	0.82	0.36	0.83
Hour 5	R^2	0.02	0.05	0.42	0.25	0.57
	adj. R^2	-0.02	-0.01	0.36	0.13	0.46
Hour 6	R^2	0.02	0.05	0.11	0.38	0.43
	adj. R^2	-0.02	-0.02	0.01	0.28	0.28
Hour 7	R^2	0.09	0.1	0.23	0.36	0.49
	adj. R^2	0.06	0.04	0.14	0.26	0.36
	Vital nodes	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	Yes	No	Yes
	Network activity	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	Yes	Yes

Table 19: Trading volume: This table shows the R^2 and adj. R^2 for the trading volume prediction models regressed onto the observations as by trading hours. The results show that at times the changepoints of the VN and NVN improve the modelling of the trading volume, but also often they do not. Notably the R^2 are much higher in the hourly regressions than in the models for the entire time series.

		(1)	(2)	(3)	(4)	(5)	(6)
Hour 1	R^2	0.04	0.07	0.08	0.12	0.2	0.25
	adj. R^2	-0.01	-0.03	-0.03	-0.06	-0.07	-0.1
Hour 2	R^2	0.02	0.03	0.06	0.07	0.11	0.14
	adj. R^2	-0.01	-0.03	-0.04	-0.07	-0.06	-0.11
Hour 3	R^2	0.02	0.04	0.05	0.06	0.11	0.13
	adj. R^2	-0.01	-0.03	-0.05	-0.08	-0.07	-0.13
Hour 4	R^2	0.08	0.11	0.13	0.13	0.16	0.21
	adj. R^2	0.05	0.04	0.04	0.01	-0.01	-0.04
Hour 5	R^2	0.04	0.09	0.28	0.3	0.18	0.37
	adj. R^2	0.01	0.03	0.2	0.19	0.02	0.17
Hour 6	R^2	0.07	0.11	0.17	0.19	0.19	0.27
	adj. R^2	0.04	0.05	0.08	0.07	0.02	0.05
Hour 7	R^2	0.15	0.24	0.27	0.3	0.34	0.39
	adj. R^2	0.12	0.19	0.19	0.19	0.21	0.2
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 20: Return robustness: This table shows that the results are robust towards our vital node definition made in Definition 1. We analyse if the results are robust towards weaker/stricter requirements for submitting a post in 10,15,20 % of the weeks and highest 5,10 % median comments. Note that 'Robustness: 10 - 10' corresponds to the definition of the paper. We table shows the R^2 and adj. R^2. The results show that regardless of the definition of a vital node, the changepoints of the VN and NVN strongly improve the modelling of the return series.

		(1)	(2)	(3)	(4)	(5)	(6)
Robustness: 10 - 5	R^2	0.009	0.02	0.03	0.091	0.045	0.118
	adj. R^2	0.004	0.011	0.016	0.074	0.021	0.088
Robustness: 15 - 5	R^2	0.009	0.02	0.04	0.097	0.045	0.125
	adj. R^2	0.004	0.011	0.026	0.08	0.021	0.095
Robustness: 20 - 5	R^2	0.009	0.02	0.049	0.106	0.045	0.135
	adj. R^2	0.004	0.011	0.036	0.089	0.021	0.105
Robustness: 10 - 10	R^2	0.009	0.02	0.028	0.088	0.045	0.115
	adj. R^2	0.004	0.011	0.014	0.07	0.021	0.084
Robustness: 15 - 10	R^2	0.009	0.02	0.035	0.093	0.045	0.12
	adj. R^2	0.004	0.011	0.022	0.075	0.021	0.089
Robustness: 20 - 10	R^2	0.009	0.02	0.049	0.106	0.045	0.135
	adj. R^2	0.004	0.011	0.036	0.089	0.021	0.105
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 21: Integrated variance robustness: This table shows that the results are robust towards our vital node definition made in Definition 1. We analyse if the results are robust towards weaker/stricter requirements for submitting a post in 10,15,20 % of the weeks and highest 5,10 % median comments. Note that 'Robustness: 10 - 10' corresponds to the definition of the paper. We table shows the R^2 and adj. R^2. The results show that regardless of the definition of a vital node, the changepoints of the VN and NVN strongly improve the modelling of the integrated variance series.

		(1)	(2)	(3)	(4)	(5)	(6)
Robustness: 10 - 5	R^2	0.175	0.189	0.256	0.288	0.255	0.336
	adj. R^2	0.172	0.182	0.245	0.274	0.237	0.313
Robustness: 15 - 5	R^2	0.175	0.189	0.19	0.21	0.255	0.269
	adj. R^2	0.172	0.182	0.179	0.195	0.237	0.244
Robustness: 20 - 5	R^2	0.175	0.189	0.19	0.211	0.255	0.27
	adj. R^2	0.172	0.182	0.179	0.196	0.237	0.245
Robustness: 10 - 10	R^2	0.175	0.189	0.239	0.268	0.255	0.317
	adj. R^2	0.172	0.182	0.228	0.254	0.237	0.293
Robustness: 15 - 10	R^2	0.175	0.189	0.19	0.21	0.255	0.269
	adj. R^2	0.172	0.182	0.179	0.195	0.237	0.244
Robustness: 20 - 10	R^2	0.175	0.189	0.19	0.211	0.255	0.27
	adj. R^2	0.172	0.182	0.179	0.196	0.237	0.245
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 22: Jump volatility robustness: This table shows that the results are robust towards our vital node definition made in Definition 1. We analyse if the results are robust towards weaker/stricter requirements for submitting a post in 10,15,20 % of the weeks and highest 5,10 % median comments. Note that 'Robustness: 10 - 10' corresponds to the definition of the paper. We table shows the R^2 and adj. R^2. The results show that regardless of the definition of a vital node, the changepoints of the VN and NVN strongly improve the modelling of the jump volatility series.

		(1)	(2)	(3)	(4)	(5)	(6)
Robustness: 10 - 5	R^2		0.022	0.034	0.056	0.102	0.133
	adj. R^2		0.017	0.025	0.042	0.084	0.107
Robustness: 15 - 5	R^2		0.022	0.022	0.044	0.102	0.123
	adj. R^2		0.017	0.012	0.03	0.084	0.096
Robustness: 20 - 5	R^2		0.022	0.022	0.044	0.102	0.123
	adj. R^2		0.017	0.013	0.03	0.084	0.097
Robustness: 10 - 10	R^2		0.022	0.032	0.054	0.102	0.132
	adj. R^2		0.017	0.022	0.04	0.084	0.106
Robustness: 15 - 10	R^2		0.022	0.022	0.044	0.102	0.123
	adj. R^2		0.017	0.012	0.03	0.084	0.096
Robustness: 20 - 10	R^2		0.022	0.022	0.044	0.102	0.123
	adj. R^2		0.017	0.013	0.03	0.084	0.097
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes

Table 23: Trading volume robustness: This table shows that the results are robust towards our vital node definition made in Definition 1. We analyse if the results are robust towards weaker/stricter requirements for submitting a post in 10,15,20 % of the weeks and highest 5,10 % median comments. Note that 'Robustness: 10 - 10' corresponds to the definition of the paper. We table shows the R^2 and adj. R^2. The results show that regardless of the definition of a vital node, the changepoints of the VN and NVN strongly improve the modelling of the trading volume series.

		(1)	(2)	(3)	(4)	(5)	(6)
Robustness: 10 - 5	R^2	0.012	0.014	0.036	0.04	0.026	0.051
	adj. R^2	0.007	0.005	0.022	0.021	0.002	0.018
Robustness: 15 - 5	R^2	0.012	0.014	0.039	0.043	0.026	0.055
	adj. R^2	0.007	0.005	0.026	0.025	0.002	0.022
Robustness: 20 - 5	R^2	0.012	0.014	0.057	0.061	0.026	0.073
	adj. R^2	0.007	0.005	0.044	0.043	0.002	0.041
Robustness: 10 - 10	R^2	0.012	0.014	0.031	0.035	0.026	0.046
	adj. R^2	0.007	0.005	0.017	0.016	0.002	0.013
Robustness: 15 - 10	R^2	0.012	0.014	0.036	0.04	0.026	0.051
	adj. R^2	0.007	0.005	0.022	0.021	0.002	0.018
Robustness: 20 - 10	R^2	0.012	0.014	0.057	0.061	0.026	0.073
	adj. R^2	0.007	0.005	0.044	0.043	0.002	0.041
	Vital nodes	No	No	Yes	Yes	No	Yes
	Non-vital nodes	No	No	No	Yes	No	Yes
	Network activity	No	Yes	Yes	Yes	Yes	Yes
	Network statistics	No	No	No	No	Yes	Yes