PLOS ONE

RESEARCH ARTICLE

Dynamics of investor spanning trees around dot-com bubble

Sindhuja Ranganathan[1]*, Mikko Kivelä[2], Juho Kanniainen[1]

1 Laboratory of Industrial and Information Management/Tampere University of Technology, Tampere, Finland, **2** Department of Computer Science, School of Science/Aalto University, Espoo, Finland

* sindhuja.ranganathan@tut.fi

Citation: Ranganathan S, Kivelä M, Kanniainen J (2018) Dynamics of investor spanning trees around dot-com bubble. PLoS ONE 13(6): e0198807. https://doi.org/10.1371/journal.pone.0198807

Editor: Roberta Sinatra, Central European University, HUNGARY

Received: August 10, 2017

Accepted: May 27, 2018

Published: June 13, 2018

Data Availability Statement: The raw data are owned by a third-party organization, Euroclear Finland Ltd and therefore the raw data cannot be made available from the authors under the non-disclosure agreement signed with the data provider. However, academic researchers can purchase the raw data from EuroClear. In this regard, jarkko.heinonen@euroclear.eu can be contacted. Moreover, we made processed data on investors' pairwise correlations public, it is currently available at https://datadryad.org//review?doi=doi:10.5061/dryad.5b8n621 and will be available at https://doi.org/10.5061/dryad.5b8n621.

Abstract

We identify temporal investor networks for Nokia stock by constructing networks from correlations between investor-specific net-volumes and analyze changes in the networks around dot-com bubble. The analysis is conducted separately for households, financial, and non-financial institutions. Our results indicate that spanning tree measures for households reflected the boom and crisis: the maximum spanning tree measures had a clear upward tendency in the bull markets when the bubble was building up, and, even more importantly, the minimum spanning tree measures pre-reacted the burst of the bubble. At the same time, we find less clear reactions in the minimal and maximal spanning trees of non-financial and financial institutions around the bubble, which suggests that household investors can have a greater herding tendency around bubbles.

Introduction

The strategic interaction and collection of individuals or agents in a financial setup can play a key role in determining their financial outcomes. Understanding how investors behave and operate has been a topic of interest in behavioral finance in the recent past. Previously, investor trading strategies and investor behavior were studied at an aggregated level using conventional regression methodologies [1, 2, 3, 4, 5, 6, 7]. In addition, the evolution of networks of stocks and currency rates and their structural change have been successfully analyzed in the existing literature [8, 9, 10, 11, 12, 13]. The effects of the economic and financial bubbles on the stock market have also been analyzed in the literature [14, 15, 16]. However, *investor networks* have received much less attention, and even though complex network methods have been applied to identify investor networks [17, 18, 19, 20], research studying the dynamics of investor networks around a financial crisis is lacking. This paper aims to take the first step toward providing an understanding of investor networks by focusing on the dynamics of investor correlation networks during the dot-com (IT Millennium) bubble using unique investor transaction registry data, which contain all the trades of Finnish households and institutions in Helsinki Exchange. In particular, we focus on the question of how gradual and non-gradual changes in investor network structures are related to the stock price process. This research opens avenues

Funding: The authors receive no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

to increase the understanding of the actual mechanisms of stock markets to identify domino effects that can propagate through investors and propel the stock markets into a crisis state.

In this paper, investor correlation matrices are obtained using time series of investor-specific daily net volumes for Nokia, a major Finnish technology company, around the millennium. At the same time, Nokia is the most liquid stock in our data sample from the Helsinki stock exchange, and there has been other research based on the company's stock market data, for example, by [18, 21, 22]. Investors' correlation matrices are estimated for three main categories of investors: financial institutions, households, and non-financial institutions. Correlation matrices can be interpreted as link-weighted networks, and the links in the resulting networks where all nodes are connected can be filtered with a multitude of different approaches [12, 23, 24, 25]. An elegant and popular method in stock market network analysis is to employ minimal or maximal spanning tree methods to find a "backbone" of the full correlation network [8, 9, 11, 26, 27, 28]. Several more complicated correlation matrix construction and filtering methods have been developed recently [23, 25, 29, 30, 31, 32, 33], but utilizing these methods is left for future research.

The analysis of investor networks dynamics in this paper introduces two theoretical challenges when compared to other financial correlation networks. First, the set of investors is much larger than, for example, the number of stocks, and the set of active investors is strongly time-varying. The vast majority of methods developed for analyzing dynamic, or temporal, networks are based on the assumption that only the links change while the set of nodes is stable [34, 35]. Further, changes in the set of investors limits the applicability of methods based on analyzing each network snapshot separately, as metrics that are sensitive to network size cannot be compared across different time windows, where the number of investors can vary significantly. The second challenge is related to the widely varying sparsity of the time series, where few investors are extremely active and many others trade very infrequently. The active investors could be investigated using high temporal resolution and short observation window lengths, but the infrequent investors must be examined using lower resolution and a longer time window. The conventional correlation analysis performed here requires that a single time resolution level and observation window length be chosen, and this choice must be a compromise between the two extremes.

We construct minimum and maximum spanning trees for networks within six-month time windows, with a displacement of one month. Our results with estimated correlations between households' transactions show that the average weight of maximum spanning tree increases and the average weight of minimum spanning tree decreases before the tipping point of the stock prices (at which stock prices start to decline), after which they remain quite stable. In other words, when the bubble propagates, on average, an investor has more positive correlations with other investors in the maximum spanning tree. At the same time, however, the correlations with the most distant investor, in terms of trading style, become even more negative in the minimum spanning tree. This suggests that households became polarized before the Nokia price crash in 2000. However, as no strong effect can be observed for financial institutions the average weights of the minimum and maximum spanning trees of institutional investors are not as clearly related to the evolution of the financial crisis.

Dot-com bubble

In this paper, we analyze the behavior of Nokia's investors around the dot-com bubble in 2000. Bubbles are a phenomenon where the prices of assets deviate from their fundamental values [36]. Generally, during bubbles, investors purchase shares anticipating future gains. When the bubbles collapse, there is a sudden fall in prices, and that was the case in the dot-

com bubble. Particularly, during the late 1990s, internet-based stocks dominated the equity markets, and there was heavy investment in the internet and technology-based start-ups with extremely optimistic expectations. As investors started pouring money into technology-based start-up companies, the prices of the companies' shares grew very high. Then, in early 2000, investments in these companies reduced drastically, and many of these companies that were expected to generate profits failed, leading to the bursting of the bubble. Consequently, there was panic selling and the market slumped.

Bubbles have been studied quite extensively and from various perspectives. According to [37], market prices during bubbles follow power-law acceleration and have log-periodic oscillations. The dot-com bubble had similar characteristics and resulted in a crash (see, for example, ref. [38]). One perspective is that bubbles occur due to the uncertainty that prevails in the market [39]. In this regard, ref. [40] provides evidence that uncertainty is a plausible cause for a sudden rise in the price of some stocks. Similarly, the high level of uncertainty matched the high prices and high return volatility in the market during the dot-com bubble. The sudden rise and fall in market prices during the dot-com bubble was associated with variations in risks from various sources. Bakshi and Wu [41] show that with the rising valuation of the NASDAQ 100, return volatility as a risk measure increased, estimates for the market price for diffusion risk became negative (from September 21, 1999 to January 5, 2000), and the market price of jump risk became unusually high. Another perspective on bubbles is that they occur when there are new innovations [42] that investors see as opportunity pulls, anticipating high profits in the future. Other reasons for the occurrence of a bubble are a lack of experience on the part of traders [43], investor's emotions [44], investor over-confidence [45], and public announcements [46]. There are several reasons for a bubble to burst. According to [40], one of the reasons that the dot-com bubble burst was that the expected profitability of technology stocks became low. Not all bubbles lead to crashes, but when a bubble does crash, it signals important information to the market. According to [42], a burst signals that there is a need to implement new innovations that occurred during the bubble period. This requires social and economic support to continue the growth of innovations that could benefit the economy.

Results

Next, we describe how we construct a series of correlation networks of investors investing in Nokia stock around the dot-com bubble (1998–2002) and report the basic statistics related to the changes in these networks. We then continue to investigate the minimal and maximal spanning trees we extract from these fully connected networks. We report the results of our analysis separately for Finnish households, financial institutions, and non-financial institutions.

Nodes in the networks: Active investors

Investors form the nodes of the network we construct, and to estimate the correlations between pairs of them, we need to have enough data on their trading behavior. Fig 1a depicts the distributions of investors divided into three categories in the period 1998–2002. Many investors traded for only a few days but relatively few traded for many days, making the data sparse. We take two steps to alleviate the problems related to sparse data in the network construction: First, we only consider *active investors* who have traded for a minimum of 20 days in a given time period. Second, daily net volumes of each active investor are averaged over a week (that is, we apply investor-specific simple moving averages).

We investigate our total sample period 1998–2002 with six-month sliding time windows using a one-month rolling window on it. Using the above definition for active investors for

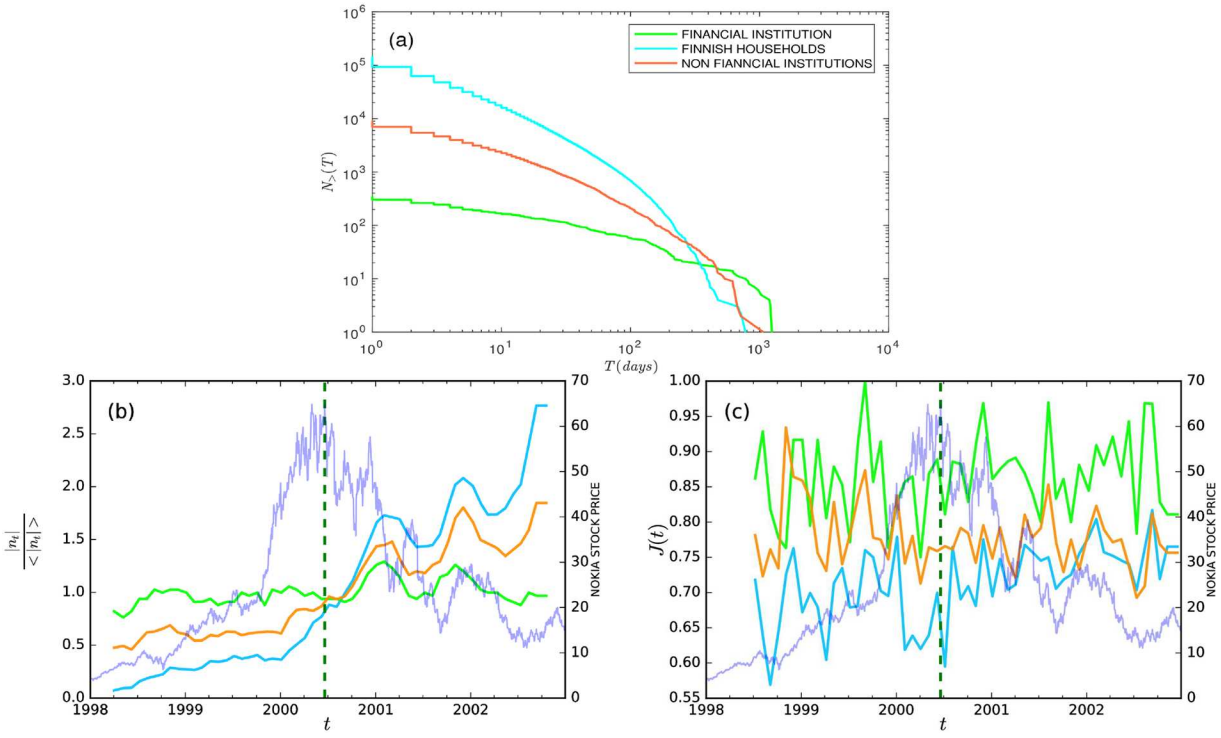

Fig 1. The number of investors in Nokia stocks during the period 1998–2002 and the change in the number of investors across the six-month time windows. (a) The number of investors $N_>(T)$ who traded the Nokia stock at least on T different days during the whole time period 1998–2002 (i.e., a non-normalized complementary cumulative distribution). (b) The evolution of the number of investors trading Nokia in the six-month time windows for households, non-financial institutions, and financial institutions. The numbers of investors in each category $|n_t|$ vary widely across categories, and they are normalized by the average numbers of investors in the full time period $\langle|n_t|\rangle$. (c) The change in the number of investors is measured using the Jaccard coefficient for different investor categories. The value of $J(t)$ is higher (lower) the more (less) similar the consecutive networks are in terms of nodes in them (see Eq 1). Results for each time window in panels (b) and (c) are plotted at the end of the window. That is, each point is estimated with data over the previous 126 trading days (6 months). The estimation windows are rolling by one month, and the resulting points are joined by solid lines. In panels (b) and (c), the green dotted vertical line in the figures represents the highest stock price of Nokia in the sample period, and the blue curves (with axis on the right) represent the Nokia stock price. In all panels, the lime-green curve corresponds to financial institutions, the cyan curve to households, and the orange curve to non-financial institutions.

https://doi.org/10.1371/journal.pone.0198807.g001

each six-month time window, Fig 1b depicts the evolution of the number of active financial institutions, households, and non-financial institutions within these time windows. We see that the numbers of active households and non-financial institutions showed positive trends over the sample period, while the number of active financial institutions remained rather stable. Importantly, the bubble "burst" did not have clear effects on the number of active traders.

Even when the number of investors in each time window is relatively stable, the set of investors can vary significantly. This is indeed the case, as shown by Fig 1c which displays the number of investors overlapping in every subsequent time window measured by the Jaccard index defined as:

$$J(t) = \frac{|n_{t+1} \cap n_t|}{|n_{t+1} \cup n_t|}, \quad (1)$$

where n_t and n_{t+1} represent the sets of nodes in the networks with time windows ending in months t and $t + 1$. For example, for households, the Jaccard index can get values as low as 0.6, which means that 40% of investors in two subsequent time windows were only present in one of them. To put this percentage into perspective, it is worth noting that out of the total of seven months spanned by the two windows, there is a five-month overlap. That is, one could expect

a high Jaccard index even if there are major changes in the sets of investors. Note that the activeness criterion (at least 20 observations in six months) is applied for each estimation period with a displacement of one month, and this filtering has an effect on the Jaccard index values. We observe that the networks of households have lower similarity to each other compared to financial institutions, meaning that the turnover of active household investors is relatively high over time. Thus, especially for the relatively inactive household investors, the networks in different time windows are bound to be very different, and any stability observed in network statics cannot be solely explained by the stability of the networks; other organizing principles in the system also have an effect.

Links in the networks: Correlations in trading patterns

We use the Pearson correlation of trading patterns of investor pairs inside each time window to construct links between the investors (for details, see Methods). The Pearson correlation coefficient has been used extensively in the network analysis of time series of stock prices [8], and it also has some clear advantages in the analysis of individual investor trading. Observations of exceptionally high trading volumes can represent days on which important information has arrived. It is of interest to analyze whether investors react to these information in the same way, and therefore it is desirable that the measure is sensitive to exceptionally large values. In contrast to the Pearson correlation, Kendall and Spearman correlations consider rank-order as opposed to metric information, and thus they do not weight these outlier days appropriately.

The nodes change between the different time windows, and the weights of the links (the correlations) are also relatively unstable. To quantify this, we show the average absolute change in correlations between nodes that remain in two consecutive time windows (see Eq 2 in Methods) and the average correlation between all pairs of nodes in Fig 2. The change in correlations between two consecutive time windows is on average lower (0.04–0.12) than the standard deviation of the correlations inside the time windows (0.14–0.23), but it is still very clearly within the same order of magnitude. That is, the network is relatively unstable in its links, but, as we

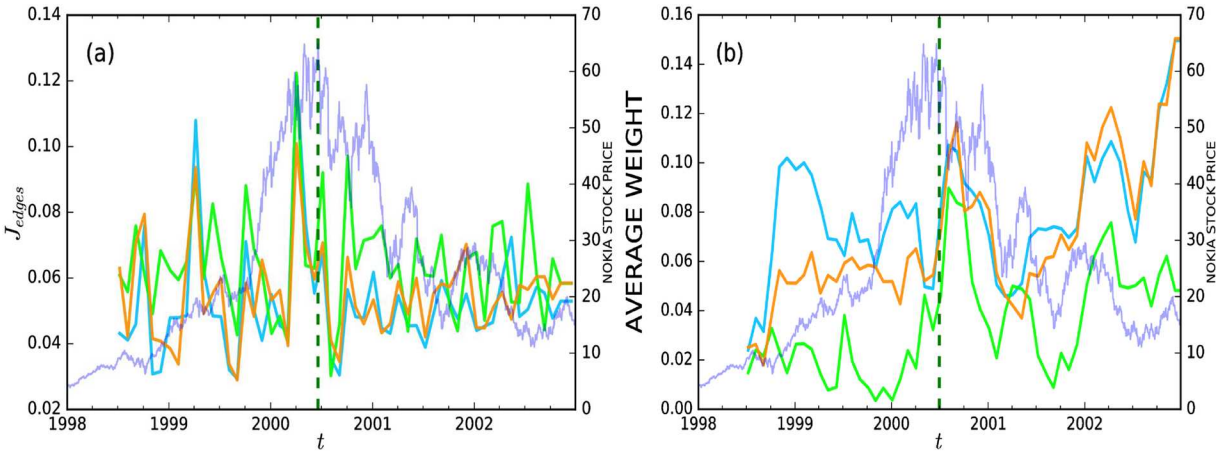

Fig 2. The change in investor correlations of Nokia stock trading across the six-month time windows during 1998–2002. (a) The average change in correlations between two consecutive time windows $J_{edges}(t)$ (see Eq 2 in the Methods section). (b) The average edge weight, or correlation, in each time window. The green dotted vertical line represents the highest stock price of Nokia in the sample period, and the blue curves (with axis on the right) represent the Nokia stock price. The lime-green curves correspond to financial institutions, the cyan curves to households, and the orange curves to non-financial institutions.

https://doi.org/10.1371/journal.pone.0198807.g002

will see below, the global organization of the network and related statistics are still rather stable.

Minimum/Maximum spanning trees

The correlation matrices of investors' net trading volumes can be interpreted as weighed networks where all node pairs (i.e., investors) are connected. Specifically, investors i and j are connected by a weight of $\rho_{ij} \in [-1, 1]$, which is the Pearson correlation coefficient between investors' daily net volumes. Clearly, the topological structure of these fully connected graphs is trivial, and all the information is in the weights. To analyze the structure, one needs to filter out parts of the edges, and there are various approaches for doing so [23, 24, 25, 33]. Following the literature on the analysis of stock prices [8, 9], we employ one of the simplest filtering methods and construct maximum (and minimum) spanning trees of correlation networks.

The idea of maximum spanning tree analysis is to filter out as many edges as possible so that the network is still connected and the highest possible weights (or, correlations) are not filtered out (for details, see the Methods section). According to ref. [17], information links may be identified from realized trades, and thus traders identified with similar trading behavior can have an (private) information channel. In light of this idea about the inference of information transfer in investor networks, the maximum spanning tree would reflect the smallest set of interactions that connect all investors and still have the strongest information flow between them. The interpretation of the empirical investor network as the information network, however, can be questioned, as two investors could certainly trade in the same directions without even knowing each other if they just follow the same investment strategies with the same public information channels. Generally speaking, the maximum spanning tree picks the most similar trading strategies while keeping the graph connected, whether or not it reflects the actual information channels. The average weight of maximum spanning tree shows how investors, on average, are pulled together or dispersed in a connected graph, and this quantity has been previously shown to react to crisis in stock price correlations [9]. The minimum spanning tree, on the other hand, reflects distant trading strategies, and its average weight can be used to analyze divergent trading strategies in a connected graph of investors. Particularly, with the minimum spanning trees, we investigate low, or even negative, correlations between investors' net volumes. Conversely, the negative correlations reflect the fact that investors net volumes are negatively related, and thus they indicate divergent trading.

Fig 3a shows the evolution of the average weight of the minimum spanning tree, L_{min}, for the *merged* network of investors in the three categories. There is an obvious, downward jump in L_{min} just before the tipping point, which is defined as the highest price of the Nokia stock during the sample period. Importantly, L_{min} is estimated using data from the past, and therefore no information about the forthcoming bubble burst was used. That is, the investors pre-reacted to the impending decline in the stock price. Next, we focus on investigating which investor groups are behind this reaction. We visualize the maximum spanning trees in Fig 3b, 3c and 3d. There does not seem to be any clearly visible clustering of categories similar to business sectors in stock networks or geographical regions in currency networks [8, 47, 48]. However, we can see that there might be some local tendency for nodes from the same category to be adjacent; however, this observation is not investigated further here.

Fig 4 displays the average weights of minimum and maximum spanning trees, L_{min} and L_{max}, around the crisis for networks containing nodes only from one of the three investor categories. Again, every data point is estimated with data over the previous 126 trading days (six months), and the estimation windows are rolling by one month. Fig 4a shows that the average weight of the minimum spanning tree, L_{min}, of the household network suddenly jumps down a

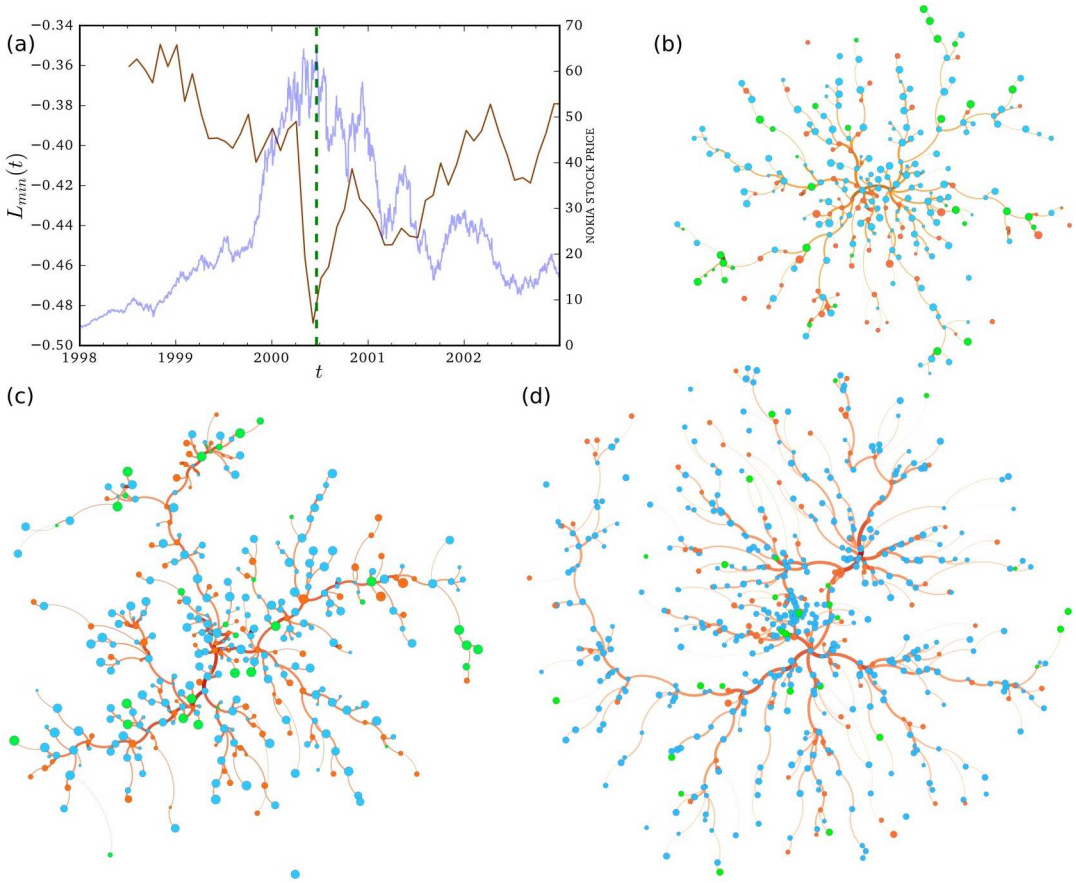

Fig 3. The minimum and maximum spanning trees of all investors. (a) Backward looking average weight of the minimum spanning tree, $L_{min}(t)$, for the merged set of investors with six-month time windows during 1998–2002 (brown line). The green dotted vertical line in the figures represents the highest stock price of Nokia in the sample period, and the blue curves (with axis on the right) represent the Nokia stock price. Maximum spanning trees between (b) July 8, 1999 and January 4, 2000 (before the crisis), (c) January 5, 2000 and July 6, 2000 (during the crisis), and (d) July 7, 2000 and January 4, 2001 (after the crisis). The cyan nodes represent households, the orange nodes non-financial institutions, and the lime-green nodes financial institutions. The sizes of the nodes are based on the volume traded by the investor during the period. However, one should not compare the sizes of nodes between different networks, as the sizes are not comparable across panels.

https://doi.org/10.1371/journal.pone.0198807.g003

few months prior the turning point of the stock price evolution around the crisis. Particularly, the value of L_{min} was -0.32 on April 3, 2000, whereas it was -0.45 on June 6, 2000, after which the stock prices started to burst. Importantly, the difference is considerably large in comparison to other changes in the data sample; however, the estimates, -0.32 and -0.45, are based on partially overlapping estimation data (the length of the estimation period is six months, and the analysis is run with a rolling window of one month). Another important observation is that the level of L_{min} does not recover back to its level as prior to the tipping point during the following two years. For non-financial and financial institutions, we see no obvious patterns in L_{min} around the crisis. Overall, weights in the minimum spanning trees among households are, on average, abnormally negative just around the turning point for households. This means that households, on average, have neighbors in the minimum spanning tree who are trading in an abnormally opposite way.

The dynamics of the maximum spanning trees in Fig 4b provide a slightly different story compared to the minimum spanning tree dynamics. In particular, we see that the average weight of the maximum spanning tree, L_{max}, for households shows a clearly positive trend

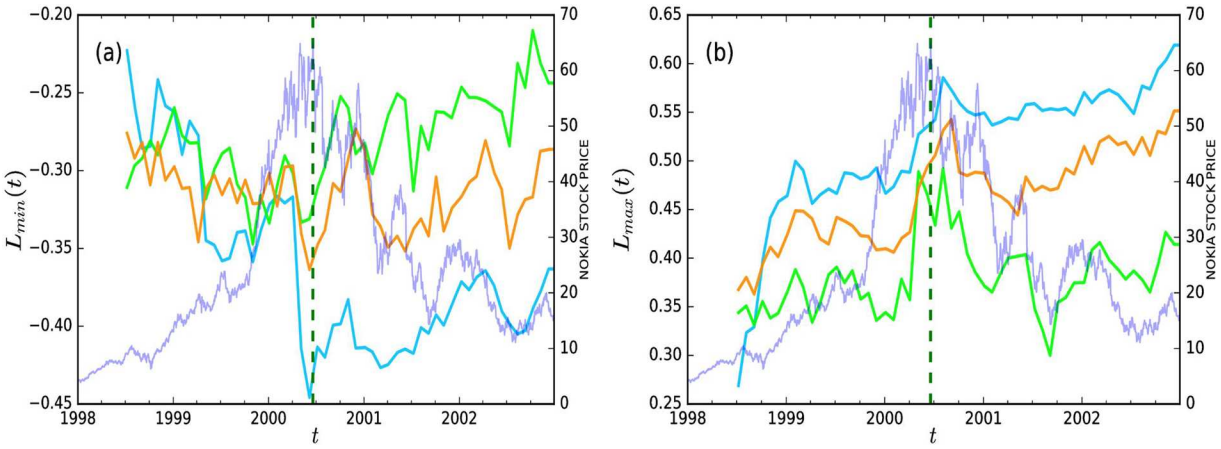

Fig 4. Backward looking average weight of the (a) minimum spanning tree, $L_{min}(t)$, (b) maximum spanning tree, $L_{max}(t)$ for different investor categories with six-month time windows during 1998–2002. The green dotted vertical line in the figures represents the highest stock price of Nokia in the sample period. The lime-green curve corresponds to the plot for Finnish financial institutions, the cyan curve corresponds to the plot for Finnish households, and the orange curve corresponds to the plot for Finnish non-financial institutions.

https://doi.org/10.1371/journal.pone.0198807.g004

prior the spike of February 2000, after which it remains quite stable. Specifically, its value was 0.27 in 1998, increasing to almost 0.6 in two years in bull markets, which is an increase of *122%*. This means that there are investors that were trading more similarly when the bubble was building up. A positive pre-trend and rather stable post-trend can also be identified for non-financial institutions, although it is weaker compared to households. Financial institutions, however, behave differently regarding L_{max}–there is a peak in L_{max} for financial institutions just before the tipping point, which lasts half a year, but otherwise L_{max} is relatively stable over the period. Note that the average weights of the networks displayed in Fig 2b do not display peaks at the same times or of the same magnitude.

In light of private information channels that investors use in trading in stock markets (see [17]), our maximum spanning tree analysis results suggest that household investors' connections to their most important neighbors in the connected graph became increasingly important when the techno bubble was building up, which could indicate herding in the stock market Moreover, the existing literature provides evidence that spanning trees for different financial networks react around financial crises, although with different data sets (and thus with different networks) compared to the present research (see [9, 49] with data on stock returns, [13] with data on stock market indexes, and [50] with data on currency exchange rates).

Discussion

This paper examines the behavior of Finnish investors using shareholding registration records for Nokia stock on the Helsinki stock exchange from 1998–2002, which includes the period of the dot-com bubble. Analyses for households, non-financial institutions, and financial institutions are conducted using minimum and maximum spanning trees constructed from correlations between investor-specific net-volumes. We find that the spanning tree measures reflected the bubble with the data for households, and, in fact, they pre-reacted to forthcoming bear markets, whereas non-financial and financial institutions show no equally clear reactions. In particular, the average correlations of households' minimum spanning trees clearly jumped down a couple of months before the Nokia price started to show a negative trend. Conversely, the average correlation in households' maximum spanning tree dynamics did not jump

suddenly right before the burst of the bubble—rather, the average correlation had a considerably large upward trend in bull markets, increasing from 0.27 to almost 0.60 in the two years before the stock price crash, after which it remained quite stable. The analysis was also conducted with 12-month rolling windows, and the results were the same (the results for this robustness check are available upon request). This result on maximum spanning trees could reflect information channels between individual household investors—investors' connections to their most important neighbors in the connected graph becoming increasingly important when the techno bubble was building up, which could indicate herding in stock markets, especially among household investors. Based on our results, it could be argued that households were mainly responsible for the bubble and its burst. This question, however, needs more research with alternative approaches. For example, agent-based models estimated with actual transaction data could be used to elaborate the role of households in more detail.

There are some restrictions in our research on correlated investor networks, which are mainly related to how the networks are constructed. We used data on investors' transactions with only one stock, as the other stocks in the our data set were too illiquid to have enough data estimating investor-specific networks. In future studies, multiple similar stocks could be pooled together or methods that function better with sparse data could be used. Another limitation is the way the Pearson correlation was used between the investment time series to calculate the similarities between nodes. There are more sophisticated ways of inferring the latent relationships between the nodes in the literature [29, 30, 31, 32], but the particular challenge in investor networks is the high variation in the transaction frequencies between investors. High-frequency nodes can be analyzed with much higher temporal resolution than low-frequency ones, and choosing a single resolution level involves a compromise between these two extremes. Finally, the spanning tree analysis discards valuable data in a very aggressive way to make the system less complex, and there are multiple alternatives in the literature where more data is kept [23, 24, 25, 33]. In future research, we aim to build the network in a more sophisticated way, which will allow us to analyze a large number of stocks with alternative methods.

The network of investors is changing dynamically, and the approach taken here—which is in line with the literature on stock correlation networks—was intended to calculate various static network metrics on snapshots of the network and then inspect how those metrics change over time. Methods that do not rely on static networks but measure the dynamics of networks have been developed in the field of temporal networks [34, 35], but most of these approaches have been constructed for networks where the links change dynamically but the nodes are relatively stable. There are, of course, other systems with long temporal data and large changes in the set of nodes, such as citation networks and collaboration networks [51, 52, 53]. In some systems, such contact networks of customers, the patterns of nodes' leaving and entering the system can even be of primary interest [54, 55, 56]. However, there are relatively few methods for analyzing networks in which both nodes and links change, and the temporal investor networks introduced here could serve as a good example for network analysis in future research. To faciliate this we have made the investor correlation matrices public (see the Methods section).

Additionally, in the present paper, the set of investors was organized based on the status of household, financial institution, or non-financial institution and activeness, which is a rather arbitrary way to classify investors. Also, one could argue that the observations of investor trading events are just realizations of a non-observable (psychological) process, making the identified temporal network unstable. In our future research, we will develop sampling methods to overcome these potential problems. In addition, filtering and community-detection methods [57] as well as alternative inference techniques for the estimation of network edges are expected in our future research.

Materials and methods

Data

The data used in this study come from the central register of shareholdings for Finnish stocks from the Finnish central depository, provided by Euroclear Finland. The data set includes all the major publicly traded Finnish stocks from 1995. It consists of shareholdings of all Finnish and non-Finnish investors traded in the Helsinki stock exchange on a daily-level basis. The data contain investors' trades and portfolios, including all Finnish household investors, Finnish institutions, and foreign institutions. The records are exact duplicates of the official certificates of ownership and trades, and hence they are very reliable. The Book Entry System entails compulsory registration of holdings for Finnish individuals (referred to as households) and institutions. Foreigners are partially exempt from registration, as they can opt for registration in a nominee name, and thus they cannot be separated from each other. Thus, data about foreigners' trades is excluded in the present paper. A more detailed descriptions of the data set is provide in [1, 18].

Our sample data consist of marketplace transactions of **Nokia** stock, consisting of investor transactions from January 1, 1998 to December 2002. Each data record contains the following information: stock ticker, owner id, trading date, transaction registration date, number of shares traded, the price of trade, buy/sell transaction type, and other investor-specific fields, such as investors' sector code, language code, gender, date of birth, and postal code. We have considered investors from different categories who have traded actively with Nokia in our analysis. Information about having the status of household, financial institution or non-financial institution was directly available from the data provider. Each investor has a unique investor ID, and for each ID, certain attributes are assigned, such as category. This information is self-provided by identifiable investors.

Links in the network

Net volume traded by an investor i on day t is given as $V_{i,t} = V^b_{i,t} - V^s_{i,t}$, where $V^b_{i,t}$ is the number of Nokia shares bought by investor i on day t, and $V^s_{i,t}$ is the number of Nokia shares sold by investor i on day t. In comparison to the inference method introduced in [18], we do not scale the net volumes by $V^b_{i,t} + V^s_{i,t}$, as the scaled approach does not measure the magnitude of trades; that is, the level of the scaled variable does not reflect exceptionally high or low traded net volumes. For example, suppose that on a given day for a given stock, investor A buys one share and sells zero while investor B buys exceptionally many shares, say, 1,000,000 and sells zero. Then, both investors' scaled net volumes would equal +1, although their trading behavior has been very different. The dependency between two investors, i and j, is measured with the Pearson correlation for M different time windows of fixed width W. In our study, W is set to 126 trading days (six months), and the analysis is run with six-month sliding time windows using a one-month (21 trading days) rolling window. As the total number of days in our data is 1252, these choices give us $M = 54$ time windows for the overall six-month time window.

Note that the data studied here are very sparse in the sense that, for many investors, most days are without any activity (see Fig 1a), although these silent days are here considered as intentional decisions not to trade. That is, the inactive days are not considered as missing data in our calculation of the Pearson correlation coefficient. In our notation, $\rho^{(ij)}_t$ denotes the Pearson correlation coefficient between investors i and j estimated from daily net volumes of W days, counted backwards from the day t. That is, the correlations between nodes at time t are

defined as,

$$\rho_t^{(ij)} = \frac{\mathbf{Cov}(\vec{V}_{i,t}, \vec{V}_{j,t})}{\sqrt{\mathbf{Var}(\vec{V}_{i,t})\mathbf{Var}(\vec{V}_{j,t})}},$$

where $\vec{V}_{i,t} = \{V_{i,\tau}\}_{\tau=t-W}^{t}$. One could also use the daily net volumes of W/2 days in the past and W/2 days in the future, but we prefer to use the data in the past instead of using the data in the future in order to analyze pre-reactions in the networks so that no information about the forthcoming bubble burst is unused.

The average absolute change in correlations between nodes that remain in two consecutive time windows is defined as

$$J_{edges}(t) = \frac{1}{|e_t \cap e_{t+1}|} \sum_{(i,j)\in e_t \cap e_{t+1}} \left(|\rho_{t+1}^{(ij)} - \rho_t^{(ij)}| \right), \quad (2)$$

where e_t denotes the set of edges in the network at time t (i.e., all pairs of nodes $e_t = \{(u, v)|u, v \in n_t, u \neq v\}$ where n_t represent the sets of nodes in the network where the time window ends at month t). Data about correlations between investor pairs is available at https://doi.org/10.5061/dryad.5b8n621.

Minimum and maximum spanning trees

For a network with $|n_t|$ nodes and edge set e_t, a maximum spanning tree is a connected sub-network with the same nodes and a subset of $|n_t| - 1$ edges $e_t^{max} \subseteq e_t$ such that the sum of the edge weights (here correlations), $\sum_{(i,j)\in e_t^{max}} \rho_t^{(ij)}$, is maximized. Similarly, for a minimal spanning tree, we find a set of edges e_t^{min} such that the sum of the edge weights is minimized.

Note that we do not transform the correlations into distance using $d_{ij} = \sqrt{2(1-\rho_t)}$, which would turn minimal spanning trees into maximal ones and vice-versa—spanning tree structure is otherwise invariant to this transformation because this transformation only reverses the rank-order of the edge weights. We also construct minimum spanning trees, which are complementary to the maximum ones.

The average weights of maximum and minimum spanning trees are defined as:

$$L_{max}(t) = \frac{1}{(N_t - 1)} \sum_{(i,j)\in e_t^{max}} \rho_t^{(ij)}.$$

and

$$L_{min}(t) = \frac{1}{(N_t - 1)} \sum_{(i,j)\in e_t^{min}} \rho_t^{(ij)},$$

respectively.

The spanning trees were computed using Kruskal's algorithm, which is a standard algorithm for finding minimum and maximum spanning trees for graphs that have positive weights. We used NetPython (available from https://github.com/CxAalto/netpython) network analysis software for computations.

Author Contributions

Conceptualization: Sindhuja Ranganathan, Mikko Kivelä.

Data curation: Sindhuja Ranganathan.

Formal analysis: Sindhuja Ranganathan.

Investigation: Sindhuja Ranganathan.

Methodology: Sindhuja Ranganathan.

Supervision: Mikko Kivelä, Juho Kanniainen.

Writing – original draft: Sindhuja Ranganathan.

Writing – review & editing: Sindhuja Ranganathan, Mikko Kivelä, Juho Kanniainen.

References

1. Grinblatt M, Keloharju M. The investment behavior and performance of various investor types: A study of Finland's unique data set. Journal of Financial Economics. 2000; 55(1): 43–67. https://doi.org/10.1016/S0304-405X(99)00044-6
2. Odean T. Are investors reluctant to realize their losses? The Journal of Finance. 1998; 53(5): 1775–1798. https://doi.org/10.1111/0022-1082.00072
3. Brennan MJ, Cao HH. International portfolio investment flows. The Journal of Finance. 1997; 52(5): 1851–1880. https://doi.org/10.1111/j.1540-6261.1997.tb02744.x
4. Kaniel R, Saar G, Titman S. Individual investor trading and stock returns. The Journal of Finance. 2008; 63(1): 273–310. https://doi.org/10.1111/j.1540-6261.2008.01316.x
5. Barrot JN, Kaniel R, Sraer D. Are retail traders compensated for providing liquidity? Journal of Financial Economics. 2016; 120(1): 146–168. https://doi.org/10.1016/j.jfineco.2016.01.005
6. Hoffmann AO, Post T, Pennings JM. Individual investor perceptions and behavior during the financial crisis. Journal of Banking & Finance. 2013; 37(1): 60–74. https://doi.org/10.1016/j.jbankfin.2012.08.007
7. Chiang TC, Zheng D. An empirical analysis of herd behavior in global stock markets. Journal of Banking & Finance. 2010; 34(8): 1911–1921. https://doi.org/10.1016/j.jbankfin.2009.12.014
8. Mantegna RN. Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems. 1999; 11(1): 193–197. https://doi.org/10.1007/s100510050929
9. Onnela JP, Chakraborti A, Kaski K, Kertesz J. Dynamic asset trees and Black Monday. Physica A: Statistical Mechanics and its Applications. 2003; 324(1): 247–252. https://doi.org/10.1016/S0378-4371(02)01882-4
10. Naylor MJ, Rose LC, Moyle BJ. Topology of foreign exchange markets using hierarchical structure methods. Physica A: Statistical Mechanics and its Applications. 2007; 382(1): 199–208. https://doi.org/10.1016/j.physa.2007.02.019
11. Heimo T, Kaski K, Saramäki J. Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks. Physica A: Statistical Mechanics and its Applications. 2009; 388(2): 145–156. https://doi.org/10.1016/j.physa.2008.10.007
12. Emmert-Streib F, Dehmer M. Influence of the time scale on the construction of financial networks. PLoS One. 2010; 5(9): e12884. https://doi.org/10.1371/journal.pone.0012884 PMID: 20949124
13. Song DM, Tumminello M, Zhou WX, Mantegna RN. Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. Physical Review E. 2011; 84(2): 026108. https://doi.org/10.1103/PhysRevE.84.026108
14. Zhou WX, Sornette D. A case study of speculative financial bubbles in the South African stock market 2003–2006. Physica A: Statistical Mechanics and its Applications. 2009; 388(6): 869–880. https://doi.org/10.1016/j.physa.2008.11.041
15. Zhou WX, Sornette D. 2000–2003 real estate bubble in the UK but not in the USA. Physica A: Statistical Mechanics and its Applications. 2003; 329(1): 249–263. https://doi.org/10.1016/S0378-4371(03)00600-9
16. Jiang ZQ, Zhou WX, Sornette D, Woodard R, Bastiaensen K, Cauwels P. Bubble diagnosis and prediction of the 2005–2007 and 2008–2009 Chinese stock market bubbles. Journal of Economic Behavior & Organization. 2010; 74(3): 149–162. https://doi.org/10.1016/j.jebo.2010.02.007
17. Ozsoylev HN, Walden J, Yavuz MD, Bildik R. Investor networks in the stock market. Review of Financial Studies. 2014; 27(5): 1323–1366. https://doi.org/10.1093/rfs/hht065

18. Tumminello M, Lillo F, Piilo J, Mantegna RN. Identification of clusters of investors from their real trading activity in a financial market. New Journal of Physics. 2012; 14(1): 013041. https://doi.org/10.1088/1367-2630/14/1/013041

19. Gualdi S, Cimini G, Primicerio K, Di Clemente R, Challet D. Statistically validated network of portfolio overlaps and systemic risk Scientific reports. 2016; 6: 39467. https://doi.org/10.1038/srep39467 PMID: 28000764

20. Baltakys, K and Kanniainen, J and Emmert-Streib, F. Multilayer Aggregation with Statistical Validation: Application to Investor Networks arXiv:1708.09850. 2018

21. Kalev PS, Nguyen AH, Oh NY. Foreign versus local investors: Who knows more? Who makes more? Journal of Banking & Finance. 2008; 32(11): 2376–2389. https://doi.org/10.1016/j.jbankfin.2007.12.031

22. Lillo F, Miccichè S, Tumminello M, Piilo J, Mantegna RN. How news affects the trading behaviour of different categories of investors in a financial market. Quantitative Finance. 2015; 15(2): 213–229. https://doi.org/10.1080/14697688.2014.931593

23. Tumminello M, Aste T, Di Matteo T, Mantegna RN. A tool for filtering information in complex systems. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(30): 10421–10426. https://doi.org/10.1073/pnas.0500298102 PMID: 16027373

24. Serrano MÁ, Boguná M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences. 2009; 106(16): 6483–6488. https://doi.org/10.1073/pnas.0808904106

25. Chi KT, Liu J, Lau FC. A network perspective of the stock market. Journal of Empirical Finance. 2010; 17(4): 659–667. https://doi.org/10.1016/j.jempfin.2010.04.008

26. Vandewalle N, Brisbois F, Tordoir X. Non-random topology of stock markets. Quantitative Finance. 2001; 1(3): 372–374. https://doi.org/10.1088/1469-7688/1/3/308

27. Wang GJ, Xie C, Stanley HE. Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks. Computational Economics. 2018; 51(3): 607–635. https://doi.org/10.1007/s10614-016-9627-7

28. Birch J, Pantelous AA, Soramäki K. Analysis of correlation based networks representing DAX 30 stock price returns. Computational Economics. 2016; 47(4): 501–525. https://doi.org/10.1007/s10614-015-9481-z

29. Kenett DY, Preis T, Gur-Gershgoren G, Ben-Jacob E. Dependency network and node influence: Application to the study of financial markets. International Journal of Bifurcation and Chaos. 2012; 22(07): 1250181. https://doi.org/10.1142/S0218127412501817

30. Qian XY, Liu YM, Jiang ZQ, Podobnik B, Zhou WX, Stanley HE. Detrended partial cross-correlation analysis of two nonstationary time series influenced by common external forces. Phys Rev E. 2015; 91: 062816. https://doi.org/10.1103/PhysRevE.91.062816

31. Nakajima J, West M. Dynamic network signal processing using latent threshold models. Digital Signal Processing. 2015; 47: 5–16. https://doi.org/10.1016/j.dsp.2015.04.008

32. Musmeci N, Nicosia V, Aste T, Di Matteo T, Latora V. The multiplex dependency structure of financial markets. arXiv:160604872 [physicssoc-ph]. 2016.

33. Kwapień J, Oświecimka P, Forczek M, Drożdż S. Minimum spanning tree filtering of correlations for varying time scales and size of fluctuations. Physical Review E. 2017; 95(5): 052313. https://doi.org/10.1103/PhysRevE.95.052313 PMID: 28618491

34. Holme P, Saramäki J. Temporal networks. Physics Reports. 2012; 519(3): 97–125. https://doi.org/10.1016/j.physrep.2012.03.001

35. Holme P. Modern temporal network theory: A colloquium. The European Physical Journal B. 2015; 88(9): 1–30. https://doi.org/10.1140/epjb/e2015-60657-4

36. Kindleberger CP. Bubbles. In: The World of Economics. Springer; 1991. p. 20–22.

37. Johansen A, Sornette D. Log-periodic power law bubbles in Latin-American and Asian markets and correlated anti-bubbles in Western stock markets: An empirical study. arXiv preprint cond-mat/9907270. 1999.

38. Johansen A, Sornette D. The Nasdaq crash of April 2000: Yet another example of log-periodicity in a speculative bubble ending in a crash. The European Physical Journal B-Condensed Matter and Complex Systems. 2000; 17(2): 319–328. https://doi.org/10.1007/s100510070147

39. Oechssler J, Schmidt C, Schnedler W. On the ingredients for bubble formation: Informed traders and communication. Journal of Economic Dynamics and Control. 2011; 35(11): 1831–1851. https://doi.org/10.1016/j.jedc.2011.05.009

40. Pástor L, Veronesi P. Was there a Nasdaq bubble in the late 1990s? Journal of Financial Economics. 2006; 81(1): 61–100. https://doi.org/10.1016/j.jfineco.2005.05.009

41. Bakshi G, Wu L. The behavior of risk and market prices of risk over the Nasdaq bubble period. Management Science. 2010; 56(12): 2251–2264. https://doi.org/10.1287/mnsc.1100.1256

42. Perez C. The double bubble at the turn of the century: Technological roots and structural implications. Cambridge Journal of Economics. 2009; 33(4): 779–805. https://doi.org/10.1093/cje/bep028

43. Dufwenberg M, Lindqvist T, Moore E. Bubbles and experience: An experiment. The American Economic Review. 2005; 95(5): 1731–1737. https://doi.org/10.1257/000282805775014362

44. Andrade EB, Odean T, Lin S. Bubbling with excitement: An experiment. Review of Finance. 2015; 20(2): 447–466. https://doi.org/10.1093/rof/rfv016

45. Abreu D, Brunnermeier MK. Bubbles and crashes. Econometrica. 2003; 71(1): 173–204. https://doi.org/10.1111/1468-0262.00393

46. Corgnet B, Kujal P, Porter D. The effect of reliability, content and timing of public announcements on asset trading behavior. Journal of Economic Behavior & Organization. 2010; 76(2): 254–266. https://doi.org/10.1016/j.jebo.2010.06.014

47. Heimo T, Kumpula JM, Kaski K, Saramäki J. Detecting modules in dense weighted networks with the Potts method. Journal of Statistical Mechanics: Theory and Experiment. 2008(08): P08007.

48. Wang GJ, Xie C, Chen YJ, Chen S. Statistical properties of the foreign exchange network at different time scales: evidence from detrended cross-correlation coefficient and minimum spanning tree. Entropy. 2013; 15(5): 1643–1662. https://doi.org/10.3390/e15051643

49. Coelho R, Gilmore CG, Lucey B, Richmond P, Hutzler S. The evolution of interdependence in world equity markets—Evidence from minimum spanning trees. Physica A: Statistical Mechanics and its Applications. 2007; 376: 455–466. https://doi.org/10.1016/j.physa.2006.10.045

50. Jang W, Lee J, Chang W. Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree. Physica A: Statistical Mechanics and its Applications. 2011; 390(4): 707–718. https://doi.org/10.1016/j.physa.2010.10.028

51. Martin T, Ball B, Karrer B, Newman MEJ. Coauthorship and citation patterns in the Physical Review. Phys Rev E. 2013; 88: 012814. https://doi.org/10.1103/PhysRevE.88.012814

52. Wu S, Das Sarma A, Fabrikant A, Lattanzi S, Tomkins A. Arrival and departure dynamics in social networks. In: Proceedings of the sixth ACM international conference on Web search and data mining. ACM; 2013. p. 233–242.

53. Hric D, Kaski K, Kivelä M. Stochastic Block Model Reveals the Map of Citation Patterns and Their Evolution in Time. arXiv:170500018 [physicssoc-ph]. 2017;.

54. Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, et al. Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology. ACM; 2008. p. 668–677.

55. Kawale J, Pal A, Srivastava J. Churn prediction in MMORPGs: A social influence based approach. In: Computational Science and Engineering, 2009. CSE'09. International Conference on. vol. 4. IEEE; 2009. p. 423–428.

56. Saramäki J, Moro E. From seconds to months: An overview of multi-scale dynamics of mobile telephone calls. The European Physical Journal B-Condensed Matter and Complex Systems. 2015; 88(6): 1–10.

57. MacMahon M, Garlaschelli D. Community detection for correlation matrices. Physical Review X. 2015; 5: 021006. https://doi.org/10.1103/PhysRevX.5.021006