Journal of Economic Dynamics & Control 131 (2021) 104217

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

journal homepage: www.elsevier.com/locate/jedc

Identification of information networks in stock markets

Margarita Baltakienė*, Juho Kanniainen, Kęstutis Baltakys

Statistical Data Analytics/Unit of Computational Sciences, Tampere University, Finland

ARTICLE INFO

Article history:
Received 22 December 2020
Revised 11 August 2021
Accepted 12 August 2021
Available online 25 August 2021

JEL classification:
D8
G10

Keywords:
Information channels
Information transfer
Investor network
Network inference
Private information
Public information

ABSTRACT

We introduce a novel method to identify information networks in stock markets, which explicitly accounts for the impact of public information on investor trading decisions. We show that public information has a clear effect on the empirical investor networks' topology. Most importantly, our method strengthens the identified relationship between investors' network centrality and returns. Furthermore, when less significant links are removed, the association between centrality and returns becomes statistically and economically stronger. Findings suggest that our approach leads to a more precise representation of the information network.

1. Introduction

Financial studies support the existence of private information channels in the stock markets (see, for example, Ahern, 2017; Brown et al., 2008; Ozsoylev and Walden, 2011; Ozsoylev et al., 2013). It has been estimated that more than two-thirds of the market price movements are not connected with public news (Cutler et al., 1988), suggesting that information is incorporated through other channels. While the diffusion of private information is intuitive, the pathways it takes are not directly observable. This motivates us to ask how information transfer between investors can be reliably identified from their trading patterns. A particular problem exists when the trading co-occurrences appear for other reasons than the communication of private information. For example, two investors without an actual social connection can trade in the same direction at the same time because of the same trading strategies, the same interpretation of the arrival of new information, or just by chance. While several investor information network inference methods already exist (Baltakys et al., 2018b; Ozsoylev et al., 2013; Tumminello et al., 2012), the impact of public information on similar trading patterns between investors is not explicitly addressed. For this reason, the objective of this paper is to take into account the impact of public information on investor behavior in the identification of the information network in stock markets.

Unlike many other types of financial networks, such as networks of financial systems (Haldane and May, 2011; Hautsch et al., 2015; Ladley, 2013), investor networks do not typically have an observable structure. Instead, the research objective is to reveal it. While the literature on investor information networks has made a significant contribution to analyzing the topic (Ahern, 2017; Berkman et al., 2014; Colla and Mele, 2010; Mantegna, 2020; Ozsoylev et al., 2013), there is

* Corresponding author.
E-mail address: margarita.baltakiene@tuni.fi (M. Baltakienė).

https://doi.org/10.1016/j.jedc.2021.104217

a lack of robust and reliable methods to identify private information transfer from investor trading data. In addition to Ozsoylev et al. (2013) and Tumminello et al. (2012), other network inference methods exploit the graph filtering via mutual information and transfer of entropy (Gutiérrez-Roig et al., 2019) or a probabilistic interaction model (Nadini et al., 2020). Existing approaches are primarily limited to identifying overall trading synchronization from investor transactions. A significant drawback of these methods in the context of private information channels is that information transfer signals cannot be separated from the use of public information in decision making in stock markets. In this regard, "public information" can be considered information that is not exclusively but publicly available such as public company announcements and observations on past returns.

The main methodological contribution of this paper is to introduce a new framework for more reliable identification of the underlying investor information network. The framework is developed to enhance private information transfer signals, taking into consideration individual investment strategies. Importantly, our proposed methodology could also help mitigate misidentification of links due to different investment strategies influenced by public signals, e.g., style or quantitative investing. The underlying idea is that the decisions of an investor are driven by the market conditions, the investor's own experience, and private or public information. Our proposed approach can determine the synchronization between two sparse time series after removing the effects of exogenous information publicly available to all the market participants. It can also be applied to lead-lag networks (Challet et al., 2018; Cordi et al., 2019). This procedure nests the existing network inference method with statistical validation introduced in Tumminello et al. (2012), and it can be seen as an extension of the inference method (without statistical validation) introduced by Ozsoylev et al. (2013). The proposed framework relies on the assumption that public information can be captured via the regression model. We acknowledge that the linear regression we implement is limited in its capacity to achieve this goal. First, we are not able to acquire all possible variables that would exhaustively capture the public information. Second, investors do not necessarily consistently react to public news and may evaluate their expectations relative to the news. For example, earnings announcements that beat analyst estimates may trigger the investors to sell if their expectations were higher, while they would be buying if the announcements were in line with their expectations.

In the empirical part, we apply our method to the investors' trading data from the Helsinki Stock Exchange (HSE). While existing methods require a number of data points for each investor to filter spurious trade synchronization, our proposed framework requires even more observational data. This is due to the proposed regression model used to establish a baseline trading behavior and identify trades as abnormal for each investor. This, naturally, introduces a selection bias, where the networks map out only a partial information network for the most active and important investors in the market. However, by varying the investor selection criteria, we found no significant changes to the results presented in the article (see the online Appendix).

Our empirical contribution is multifold. First, we show that taking public news into account in the network inference has a clear and statistically significant impact on the investor information network topology. Second, we show that accounting for public information in the network inference reveals a better representation of the information network in terms of a stronger and statistically more significant association between investors' centrality and their earned returns. Particularly, central investors in information networks are better informed as they receive information signals earlier, which, in turn, may affect their trading behavior and profitability (Colla and Mele, 2010; Grossman and Stiglitz, 1980; Hellwig, 1980; Kyle, 1985). This is closely related to Ozsoylev et al. (2013), who find that central investors earn higher returns than peripheral investors. However, differently from the aforementioned study, we explicitly consider the influence of public information arrivals on investors' trades when setting up the networks. Moreover, instead of inferring a single network as in the latter study, we start by estimating information networks for each of the 22 investigated securities, separately for buy and sell sides. Third, our findings show that the positive association between network centrality measures and future returns in the stock market becomes both statistically and economically stronger when network links are validated using harsher thresholds. Among other robustness checks, we find that harsher validation is associated with lower distances between investors connected in the resulting information networks. This is a novel and intuitive finding, well supported by research literature (Backstrom et al., 2010; Baltakys et al., 2018a; Liben-Nowell et al., 2005; Preciado et al., 2012), which suggests that the resulting networks better resemble the underlying information transfer channels.

Stock markets are driven by information, which can be public or private. If private information channels are reliably identified, questions regarding price formation (Jackson, 1991), investor welfare and asset pricing (Ozsoylev and Walden, 2011), insider trading (Ahern, 2017), and volatility dynamics (Walden, 2019) can be addressed in relation to private information transfer. In this regard, it is important to have a solid methodological foundation for the identification of information networks. This paper provides advanced methods to identify private information channels, which can serve answering the questions mentioned above in future research.

2. Data set

We use a world-wide unique investor-level stock market transaction data set provided by Euroclear Finland. The data set contains all transactions executed in the Helsinki Stock Exchange by Finnish investors between 1995 and 2009. Each transaction in the data set is characterized by the investors' anonymized ID, trade and registration dates, security ISIN code, traded volume, investor's sector code, postal code, and investor's birth year and gender for household investors. Using the sector code information, we can classify investors into six main categories:

- households (private individuals),
- non-financial corporations,
- financial and insurance corporations,
- government,
- non-profit institutions,
- the rest of the world (foreign private individuals or institutions).

Each domestic investor has a unique ID associated with all of her transactions. In contrast, foreign investors may choose to keep their holdings and execute their trades through the nominee accounts of financial intermediaries. In such cases, the transactions are recorded with the financial intermediary's investor ID and a special flag to indicate that a specific transaction belongs to some client of the intermediary behind the investor ID. Because of this, we cannot keep track of foreign investor transactions using the same financial intermediary. As our analysis focuses on individual rather than aggregate investor trading decisions, we have eliminated nominee transactions to avoid biased results[1].

In this study, we analyze the trading data for 22 securities that composed OMX Helsinki 25 (OMXH25) benchmark index as of December 2009[2] (see Table Appendix A.1). As discussed in Baltakienė et al. (2019) the database contains incorrectly dated transactions between 1998 and 2004. Therefore, we select our analysis period to span between 1 August 2005 and 13 November 2009.[3]

We limit the investor set to the most active ones who traded at least ten out of 22 investigated securities for at least ten trading days during the analysis period[4], and traded Nokia stock. In the selected set of 2245 investors, there are 1755 households, 305 non-financial, 87 financial-insurance, 33 general-government, 39 non-profit institutions, and 26 foreign investors. On the one hand, such filtering may remove the most active stock-specific investors, which may influence the agent centrality in the network. On the other hand, such investor selection allows for a meaningful comparison of trading strategy coefficients and guarantees sufficient input for the regression analysis[5].

By requiring that investors trade at least ten days each of the ten securities, the resulting networks have a relatively stable set of investors. This way, the conclusions about the network differences after incorporating public information are more comparable and generalizable over all securities. Importantly, we have performed robustness checks using different thresholds – ten days and nine/ten/12/15 securities. The results were consistent for all threshold choices and are shown in Fig. Appendix H.1.

3. Methodology

Identifying information flows in investor networks from trading data is non-trivial because the social links are not directly observable. Similar trade timing could indicate the existence of a social link between a pair of investors. However, it can also be a consequence of investors' using the same public information sources and similar trading strategies. For example, a pair of investors can have the same kind of implementation of the contrarian strategy, i.e., buying (selling) a security if its price drops (increases) by a certain amount. To mitigate this issue, we incorporate public information and other exogenous influences on investors' behavior in the network inference procedure. We do this by leveraging the constrained multivariate linear regression model to specify investors' expected trading volume based on the existing public information. The idea is that non-public information drives investors' behavior that cannot be explained by public information.

Using the model's output, we can classify investor's daily trading behavior into abnormal buying and abnormal selling states. We define the abnormal buying (selling) state as buying more/selling less (selling more/buying less) than conditionally expected by investor-specific regression models based on the public information available at the moment. For example, a pair of investors may utilize more or less different strategies on how to use public information. However, they can adjust their trading to the same direction if they have mutual (private) information transfer. Our method thus helps to identify these adjusted (abnormal) reactions, influenced by non-public information.

The identified co-occurrences of investors' abnormal trading states are then used to establish links in the information network in a statistical manner. Thus, our method discounts the effects of the undesired observable market variables to investor trade synchronization. In this paper, we capture the public information with investor net traded volumes, security and market returns, volatilities, and information on company announcements, which are further described in Section 3.2. Naturally, the list of our used variables does not exhaust all possible ways to include public information. As a matter of fact, our proposed procedure does not depend on a specific model. In future research, it can be replaced by any model that predicts investor trading behavior, e.g., differently specified multivariate and/or non-linear regression models with different explanatory variables on public information in stock markets or neural networks.

[1] More information about the data set can be found in other research publications (see, e.g., Baltakienė et al., 2019; Baltakys, 2019; Baltakys et al., 2018a; 2018b; 2020; Ilmanen and Keloharju, 1999; Ranganathan et al., 2018; Siikanen et al., 2018).

[2] 3 securities were dropped since they had their IPOs during the analysis period.

[3] This is the last day when non-aggregate marketplace transactions are recorded in the data set for the analyzed ISINs and investors.

[4] If an investor traded more than ten securities, her 10th least traded security should have been traded on at least 10 days.

[5] See the *Methodology* section.

In addition, to exclude spurious links, we statistically validate the links between investors using the hypergeometric test with a null hypothesis of random co-occurrence of investors' trading events. We use various threshold levels to prune the network.

3.1. Investor abnormal trading state categorization

For each investor i and her/his traded security k on the trading day t, we calculate the scaled euro net-volume as

$$v_{i,k,t} = \frac{V^b_{i,k,t} - V^s_{i,k,t}}{V^b_{i,k,t} + V^s_{i,k,t}} \tag{1}$$

where $V^b_{i,k,t}$ and $V^s_{i,k,t}$ are the daily buy and sell euro volumes, respectively, aggregated from investor i transactions of the stock k that occurred on the day t, which are observed from the shareholder registration data. In Tumminello et al. (2012), investors' buy and sell trading states are defined using the scaled euro net-volume as follows:

$$\begin{cases} b - \text{buying state, when } v_{i,k,t} > \theta, \\ s - \text{selling state, when } v_{i,k,t} < -\theta, \end{cases} \tag{2}$$

where $\theta > 0$.

In order to incorporate public information in the investor network inference, we calculate the scaled euro net-volume predicted by a model based on the use of public information:

$$\widehat{v}_{i,k,t} = \begin{cases} \frac{\widehat{V}^b_{i,k,t} - \widehat{V}^s_{i,k,t}}{\widehat{V}^b_{i,k,t} + \widehat{V}^s_{i,k,t}}, & \text{if } \widehat{V}^b_{i,k,t} + \widehat{V}^s_{i,k,t} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\widehat{V}^b_{i,k,t}$ and $\widehat{V}^s_{i,k,t}$ are the corresponding buy and sell euro volumes predicted by a model using *public* data that has been available a day before the trade. In this paper, we employ constrained linear regressions for the predictions with eight different types of variables discussed in the next section. Importantly, our model choice can be replaced by a more sophisticated non-linear model with different explanatory variables. We assign daily states to each investor using the difference between the real and predicted scaled volumes for a selected threshold $\theta > 0$ as follows:

$$\begin{cases} b^* - \text{abnormal buying state, when } v_{i,k,t} - \widehat{v}_{i,k,t} > \theta, \\ s^* - \text{abnormal selling state, when } v_{i,k,t} - \widehat{v}_{i,k,t} < -\theta. \end{cases} \tag{4}$$

In this paper, we use $\theta = 0.10$. Note that we do not determine $v_{i,k,t}$ and $\widehat{v}_{i,k,t}$ for the days that had no trading activity, and therefore, on such days, we do not assign an abnormal trading state. Similarly, a trading state is not assigned when the net traded volume $v_{i,k,t}$ closely follows the regression model prescribed strategy, i.e., $-\theta \leq v_{i,k,t} - \widehat{v}_{i,k,t} \leq \theta$. Additionally, when the model predicted net traded volumes $\widehat{v}_{i,k,t}$ equal 0, i.e., investor behavior is independent of the public information, our proposed method reduces to the default investor trading state categorization, introduced by Tumminello et al. (2012).

3.2. Predicting investor-specific euro volumes with public information variables

To obtain $\widehat{v}_{i,k,t}$, we use a constrained linear regression model (Coleman and Li, 1996) with positive output variables. Particularly, we seek to explain how day t buy (sell) euro volume $V^b_{i,k,t}$ ($V^s_{i,k,t}$) of investor i for a given security k is associated with the stock and market returns, volatility, and company news announcements. We base our choice of the explanatory variables on the financial literature findings. For example, Grinblatt and Keloharju (2001b) find that past returns and the price movements influence investor trading decisions.

We denote security k's daily log return as $r_{k,t}$ and define:

- yesterday's return as $r^{(d)}_{k,t-1} = r_{k,t-1}$,
- last week's return (excluding yesterday's return) as $r^{(w)}_{k,t-1} = \sum_{\tau=t-2}^{t-5} r_{k,\tau}$,
- last month's return (excluding last week's return) as $r^{(m)}_{k,t-1} = \sum_{\tau=t-6}^{t-21} r_{k,\tau}$.

In addition, we estimate the daily volatilities denoted as $\sigma_{k,t}$ from daily returns using the GARCH model (Bollerslev, 1986):

$$\begin{aligned} r_{k,t} &= \mu_k + \sigma_{k,t} e_{k,t} \\ \sigma^2_{k,t} &= \omega_k + \alpha_k \sigma^2_{k,t-1} e^2_{k,t-1} + \beta_k \sigma^2_{k,t-1}, \end{aligned} \tag{5}$$

where $e_{k,t} \sim N(0, 1)$ and σ_0 is an unconditional standard deviation of returns estimated from the daily returns in July 2005. Then we define:

- yesterday's volatility as $\sigma^{(d)}_{k,t-1} = \sigma_{k,t-1}$,

– last week's volatility (excluding yesterday's volatility) as $\sigma^{(w)}_{k,t-1} = \sum_{\tau=t-2}^{t-5} \sigma_{k,\tau}$,
– last month's volatility (excluding last week's volatility) as $\sigma^{(m)}_{k,t-1} = \sum_{\tau=t-6}^{t-21} \sigma_{k,\tau}$.

We calculate the returns and volatilities for all the 22 securities and for the composite index OMXH25, denoted by $\{\tilde{r}_t, t = 1, 2, \dots\}$ and $\{\tilde{\sigma}_t, t = 1, 2, \dots\}$, respectively.

According to Barber and Odean (2008), in addition to large returns, trading volumes and media spotlight attract investors' attention. In light of this, we explain investor's trading behavior by the euro volume turnover, $\tilde{V}_{k,t}$, which we use to control for the impact of total yesterday's, last week's and last month's euro volume. Moreover, to take into account scheduled and non-scheduled company media announcements, we denote yesterday's announcement dummy variable $A_{k,t-1}$ and we analyze its moderating effect on yesterday's return $r^{(d)}_{k,t-1}$. The product of these two variables indicates whether the announcement was positive or negative. Moreover, we add dummy variables for today's and tomorrow's scheduled announcements, $A^*_{k,t}$ and $A^*_{k,t+1}$.

There can be an autoregressive structure in investors' trading behavior. For this reason, we control the regression by investor's trading with the investor's yesterday's, last week's, and last month's total net traded euro volume. We denote the investor-specific net traded euro volume in a specific security as $\mathrm{NV}_{i,k,t} = V^b_{i,k,t} - V^s_{i,k,t}$ and define:

– yesterday's net euro volume as $\mathrm{NV}^{(d)}_{i,t-1} = \mathrm{NV}_{i,t-1}$,
– last week's net euro volume (excluding yesterday's volume) as $\mathrm{NV}^{(w)}_{i,t-1} = \sum_{\tau=t-2}^{t-5} \mathrm{NV}_{i,\tau}$,
– last month's net euro volume (excluding last week's volume) as $\mathrm{NV}^{(m)}_{i,t-1} = \sum_{\tau=t-6}^{t-21} \mathrm{NV}_{i,\tau}$.

Furthermore, to control for the stock-specific strategies, we take the dummy variables for traded securities $\mathbb{1}_k(l)$, where l is one of the 22 investigated securities (excluding Nokia, as its effects will be captured with the baseline dummy α_0).

Using the variables defined above, for each investor i, we define the regression for buy and sell euro volumes in the security k as follows:

$$\begin{aligned} V^{[b,s]}_{i,k,t} = \alpha_0 &+ \sum_x^{[d,w,m]} \alpha_{i,r,x} r^{(x)}_{k,t-1} + \sum_x^{[d,w,m]} \alpha_{i,\sigma,x} \sigma^{(x)}_{k,t-1} + \sum_x^{[d,w,m]} \alpha_{i,\mathrm{NV},x} \mathrm{NV}^{(x)}_{i,k,t-1} \\ &+ \sum_x^{[d,w,m]} \alpha_{i,\tilde{r},x} \tilde{r}^{(x)}_{t-1} + \sum_x^{[d,w,m]} \alpha_{i,\tilde{\sigma},x} \tilde{\sigma}^{(x)}_{t-1} + \sum_x^{[d,w,m]} \alpha_{i,\tilde{V},x} \tilde{V}^{(x)}_{k,t-1} \\ &+ \alpha_{i,A_1} A_{k,t-1} \cdot r^{(d)}_{k,t-1} + \alpha_{i,A_2} A^*_{k,t} + \alpha_{i,A_3} A^*_{k,t+1} \\ &+ \sum_l \alpha_{i,\mathbb{1},l} \mathbb{1}_k(l) + \varepsilon_{i,k,t}, \end{aligned} \tag{6}$$

where $\varepsilon_{i,k,t}$ is i.i.d. random variable. The regressions are run separately for all the investors, on both buy and sell sides.[6] In that way, we obtain the model predicted buying $\widehat{V}^b_{i,k,t}$ and selling $\widehat{V}^s_{i,k,t}$ euro volumes.

3.3. Information networks

In Ozsoylev et al. (2013), a pair of investors is linked if they have same direction trade co-occurrences within a given time window. In our proposed approach, we infer security-specific information networks from the *abnormal* trading state synchronization, linking investors i and j if they had the same *abnormal* trading state on at least one trading day Eq. (4). We perform this procedure separately for the abnormal buying and selling trading state types. Altogether we end up with an ensemble of 44 information networks, having two networks for each security.

In the results Section 4.3, when investigating the association between investor centralities in the information networks and their trading returns, we will merge the security-specific information networks inferred from buying and selling behaviors by taking the union of their links. This will be done separately for networks inferred from trading state and our proposed abnormal trading state co-occurrences.

To see how our proposed categorization changes the networks compared to the networks that do not discount the effect of public information, we use non-adjusted net scaled volumes Eq. (1) to construct the networks using trading states defined by Eq. (2). We refer to investor networks as *trade co-occurrence networks* if they have been constructed by not taking the public information into account, i.e., the conventional approach with Eq. (2). In fact, they represent the networks proposed by Ozsoylev et al. (2013) and Tumminello et al. (2012). If public information has been taken into account, the networks are referred to as *abnormal trade co-occurrence networks*, i.e., our proposed approach with Eq. (4). The dichotomy between the discussed network types is illustrated in Table 1.

3.4. Validating information networks

To exclude the spurious links from the abnormal trading state synchronization, we leverage the hypergeometric test (Tuminello et al., 2011). For security k, we estimate four parameters for each pair of investors i and j in the information

[6] Using a standard PC, it took roughly 8 seconds on average to estimate a multivariate regression model with our data. Since the model is predicted for each investor separately, the estimation of the investor regression variables can be performed in parallel in large scale markets.

Table 1
A summary of the differences between various investor network inference methods.

	Proxies for investor information networks	
	Trade co-occurrence networks	**Abnormal trade co-occurrence networks**
Public information	Not considered	*Considered*
Links estimated based on synchronization of	Trading states	*Abnormal* trading states
Network links non-validated	Referred to as empirical investor network (EIN) in Ozsoylev et al. (2013)	Introduced in Section 3.3
Network links validated	Referred to as statistically validated (investor) network (SVN) in Tumminello et al. (2012)	Introduced in Section 3.4

networks: $T_k, N^P_{i,k}, N^P_{j,k}, N^P_{i,j,k}$. First, we assume that each investor could have traded on any trading day during the analyzed period. Consequently, the length of the joint trading period T_k for each pair of investors is 1081 days. Then, for each investor i we denote $N^P_{i,k}$ as the number of days when investor i had an abnormal buy or sell state, $P \in \{b^*, s^*\}$, respectively. Finally, we count the number of occurrences when investors i and j were in the same state on the same days, and denote it by $N^P_{i,j,k}$.

Under the null hypothesis of random abnormal trading state co-occurrences for investors i and j trading a stock k with a given abnormal trading state $P \in \{b^*, s^*\}$, the probability of observing X co-occurrences in T_k observations can be estimated by the hypergeometric distribution $H(X|T_k, N^P_{i,k}, N^P_{j,k})$. Each link between investors i and j is associated with a p-value, which equals the probability of having at least $N^P_{i,j,k}$ trade co-occurrences under randomness, i.e.:

$$p(N^P_{i,j,k}) = \text{Prob}\big(X \geq N^P_{i,j,k}\big) = 1 - \sum_{Y=0}^{N^P_{i,j,k}-1} H\big(Y|T_k, N^P_{i,k}, N^P_{j,k}\big). \tag{7}$$

As there are multiple links to be validated in the network, i.e., multiple hypotheses are tested, the chances of keeping spurious links increases. In order to analyze the impact of the accumulation of false positive errors, we investigate the properties of investor networks by using 100 different p-thresholds, equally spaced on a log scale from 10^{-10} to 1. Moreover, we report some of the results for the case of Bonferroni multi-test correction in the main analysis and the Appendix. In the case of Bonferroni multi-test correction, the statistical significance α is adjusted by dividing it by the total number of tests performed, i.e., $\alpha_b = \alpha / n_{\text{tests}}$.

4. Results

In this section, we first provide summary statistics on the regression results over all the investors Eq. (6). Then, we consider to which extent the incorporation of the public information in the network inference changes investor networks' topologies. In particular, we estimate the networks based on the trading states and abnormal trading states (not using and using public information). If the resulting network topologies are statistically different, then the incorporation of public information matters in the inference. Specifically, suppose some of the links disappear with the use of public information. In that case, the existing approaches can yield false-positive links, potentially because investors' similar reactions to the arrival of public information can be incorrectly interpreted as evidence about a private information channel. On the other hand, our method can also introduce new links that cannot be identified with the existing methods. This can happen if the behavior of two investors is seemingly different, which does not necessarily exclude the possibility of private information transfer. They can behave differently because of different strategies and reactions to public information arrivals, but once these have been filtered out, their "residual behaviors" can coincide.

Next, we analyze the relationship between investors' centrality and returns they earn. Ozsoylev et al. (2013) identifies a positive association using data on Istanbul Exchange with non-validated networks. Given that a positive relationship between investor centrality and returns reflects the existence of information pathways within the identified investor network, we may verify our method by testing if the incorporation of public information in the network inference strengthens this relationship. We conduct this test without and with network validation with various p-thresholds.

4.1. Summary statistics on the significance of regression parameters

For every investor we estimate investor-specific regression coefficients defined in Eq. (6). Table 2 shows the percentage of statistically significant regression coefficients over 2245 investors for buying and selling euro volume models. Our findings are consistent with the literature (see, e.g., Barber and Odean, 2008; Grinblatt and Keloharju, 2001b) and show that traded euro volumes, scheduled announcements, and, to a lesser extent, returns drive investor trading actions. However, the model does not provide evidence about the importance of volatility, market volatility, market returns, yesterday's announcements', and tomorrow's scheduled announcements.

Table 2
Statistically significant coefficients for market variables driven by the public information at the significance level 0.05. The values are shown as percentage of statistically significant coefficients out of 2245 investor models. See the notation of the column names (regression variables) in the Section 3.2. Indices i and k are dropped for readability.

	$\mathrm{NV}_{t-1}^{(w)}$	$\mathrm{NV}_{t-1}^{(d)}$	$\mathrm{NV}_{t-1}^{(m)}$	A_t^*	$\tilde{V}_{t-1}^{(m)}$	$\tilde{V}_{t-1}^{(d)}$	$\tilde{V}_{t-1}^{(w)}$	$r_{t-1}^{(m)}$	$r_{t-1}^{(w)}$	$r_{t-1}^{(d)}$	$\tilde{r}_{t-1}^{(m)}$
V_t^b	62%	60%	52%	29%	18%	13%	11%	11%	11%	11%	8%
V_t^s	63%	57%	62%	12%	16%	10%	11%	10%	10%	9%	6%

	$\tilde{r}_{t-1}^{(w)}$	$\tilde{\sigma}_{t-1}^{(m)}$	$\tilde{\sigma}_{t-1}^{(d)}$	$\sigma_{t-1}^{(m)}$	A_{t+1}^*	$\tilde{r}_{t-1}^{(d)}$	$\tilde{\sigma}_{t-1}^{(w)}$	A_{t-1}	$\sigma_{t-1}^{(d)}$	$\sigma_{t-1}^{(w)}$
V_t^b	7%	7%	5%	5%	4%	4%	4%	2%	2%	2%
V_t^s	6%	6%	3%	3%	4%	3%	2%	2%	1%	1%

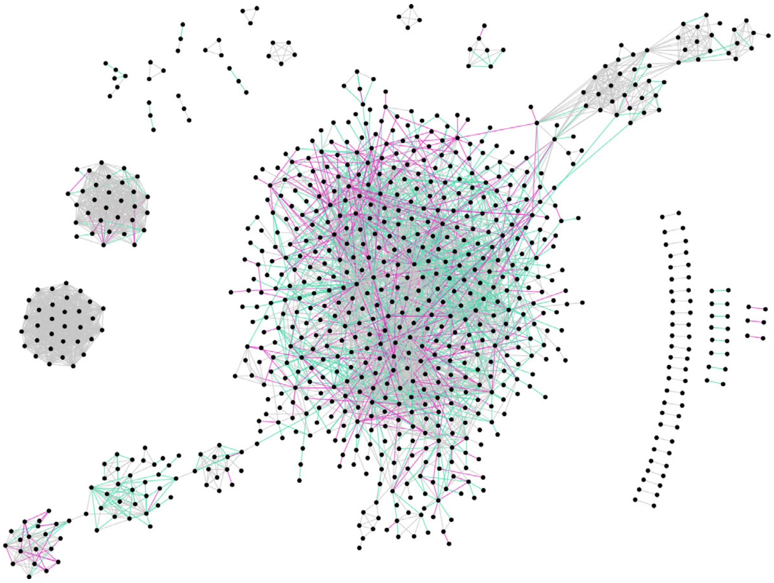


Fig. 1. Union of Bonferroni-validated information networks inferred from buying and abnormal buying state synchronization for Nokia. **Legend:**🟩 - links of validated information network inferred from buying states, 🟪 - links of validated information network inferred from abnormal buying states, ⬜ - common links, ● - investors present in at least one of the networks.

Further justification of the selected model parameters is shown in clustering analysis of investor trading strategies (see Table Appendix C.1 and Fig. Appendix C.1).

4.2. Impact of public information on the information network topology

In this section, we show that taking into account public information leads to substantially different information networks. We do this by comparing the results in the network inference procedures with and without considering public information. We compare the changes in the trading states, as well as resulting networks, both validated and non-validated.

First, compared to the trading state categorization of investor behavior used in previous research (see Table 1), the trading state categorization that takes into account the public information Eq. (4) yields changes in 10.69% of all trading states and 10.62% of trading states on average per ISIN. Second, for each security, we link a pair of investors if they had at least one day with the same (abnormal) trading state in the information networks. Third, we perform the network link validation. We validate information networks obtained using trading state and the new – abnormal trading state – assignment rules. Importantly, the inference with the public information on the validated networks results in a notable network rewiring, adding new links, and removing a portion of the old ones. For example, let us take the Bonferroni-validated information network inferred from buying behavior synchronization in Nokia. There are 3365 links in the network inferred without considering the public information, and 857 links are replaced by 436 new links when taking public information into account. Similar changes for all securities are documented in Table Appendix D.1. Generally, there are fewer links when taking the public information into account in the network inference, which suggests that *the existing methods tend to show links that, in fact, represent investors' synchronous reactions to public information rather than private information transfer*.

To demonstrate the resulting investor network topologies graphically, we continue the previous example with Nokia. Grey links in Fig. 1 represent links common to both methods, whereas links present only in the network inferred without considering public information are shown as solid black lines, and links that appeared after our proposed modification –

Table 3
Local and global similarity measures between information networks inferred from trading and abnormal trading states. The results are provided only for validated networks inferred from buying and selling behaviors, because the measures are appropriate for sparser networks. Local similarity is defined as a statistically significant overlap between nodes neighborhoods in two networks. Global similarity is defined as the Jaccard similarity index between the links in two networks. *Avg.* (*Std.*) is the average (standard deviation) of the similarity measures for 22 companies.

Company	Buying		Selling	
	Local	Global	Local	Global
Cargotec	0.38	0.79	0.42	0.82
Elisa	0.49	0.88	0.49	0.88
Fortum	0.46	0.80	0.51	0.87
Kesko (B)	0.32	0.69	0.47	0.91
KONE	0.44	0.84	0.47	0.78
Konecranes	0.55	0.91	0.59	0.86
Metso	0.45	0.82	0.43	0.86
Neste	0.47	0.76	0.43	0.79
Nokia	0.51	0.66	0.45	0.72
Nokian Renkaat	0.42	0.77	0.46	0.75
Nordea Bank	0.43	0.76	0.45	0.85
Outokumpu	0.54	0.82	0.50	0.88
Pohjola Bank (A)	0.31	0.60	0.29	0.69
Rautaruukki	0.47	0.85	0.51	0.90
Sampo (A)	0.46	0.88	0.51	0.94
Sanoma	0.60	0.93	0.60	0.88
Stora Enso (R)	0.48	0.83	0.45	0.90
TeliaSonera	0.49	0.85	0.50	0.94
Tieto	0.42	0.78	0.51	0.70
UPM-Kymmene	0.48	0.81	0.43	0.84
Wärtsilä	0.28	0.59	0.37	0.51
YIT	0.42	0.82	0.39	0.80
Avg.	0.45	0.79	0.47	0.82
Std.	0.08	0.09	0.07	0.10

dashed black. In this illustration, we can observe that our method mostly rewired the central giant component while the disconnected and peripheral communities were preserved.

4.2.1. Global and local similarity measures

To take a closer look into the magnitude of difference between network topologies obtained using the different methods, we investigate the global and local information network similarities, see Table 3. As a global similarity measure, we calculate the Jaccard coefficient as the ratio of the common and total network links produced by the two methods[7]. We report these measures for validated networks only, as they are appropriate to analyze sparser networks. The Jaccard coefficient indicates that, on average, 79% (82%) of the links overlap in the buy (sell) information networks.

As a local network similarity measure, we define the ratio of the nodes with statistically significantly overlapping neighborhoods; that is, we ask how many of the investors remain connected to a similar neighborhood. Here we again employ the hypergeometric test to evaluate the neighborhood similarity. We find that 45% (47%) of investors in information networks inferred from buying (selling) state synchronization keep similar connections in their neighborhoods (see local similarity measure in Table 3). While the global similarity is relatively high, the local similarity is notably lower.

4.2.2. Impact of public information on network link weights

To further confirm whether the incorporation of public information in the inference has a significant influence on the network structure, we compare link weights, Katz, and closeness centralities using the Wilcoxon signed-rank test. The Wilcoxon signed-rank test is based on the null hypothesis that the paired value differences' median is zero. We opt for the test with Pratt modification, which involves zero-differences of paired observations in the ranking process, but drops the ranks of the zeros (Pratt, 1959). We define the link weights using the p-value of the hypergeometric tests for all link pairs as follows:

$$w^{P}_{i,j,k} = 1 - p\left(N^{P}_{i,j,k}\right). \tag{8}$$

To investigate the effect of public information on the resulting networks, we run the paired Wilcoxon (Pratt) signed-rank test, separately for the information networks inferred from the buying and selling trading states, on the link weights Eq. (8) in non-validated (Table 4) and Bonferroni-validated information networks (Table Appendix E.1). For each test, we define the number of paired observations N as the union of observed links in the information networks inferred using the

[7] The number of links in the network inferred from trading and abnormal trading states, and their overlap for each ISIN are given in Table Appendix D.1.

Table 4
Wilcoxon signed-rank test with Pratt modification for observed link weights in **non-validated** networks. Here, ΔMedian is the difference of the weight medians in the networks inferred from trading states and abnormal trading states, *N* is the number of observations, *N.w.z* is the number of observations excluding zero-differences, *p* is the *p*-value of the Wilcoxon (Pratt) signed-rank test. The information network comparison between networks inferred from trading and abnormal trading is presented separately for buying and selling networks.

Company	Buying					Selling				
	ΔMedian	*N*	*N.w.z*	*p*		ΔMedian	*N*	*N.w.z*	*p*	
Cargotec	4.55e-03	190726	160864	3.86e-48	***	−6.82e-04	146472	125939	1.55e-136	***
Elisa	1.04e-02	250683	197358	3.55e-302	***	−6.12e-04	142972	117888	0	***
Fortum	6.61e-03	738966	642177	0	***	−1.48e-03	451083	386135	0	***
Kesko (B)	1.04e-02	232091	184257	0	***	2.87e-03	137232	116343	0	***
KONE	2.15e-02	196516	183732	1.85e-19	***	−7.46e-03	166127	159717	4.05e-243	***
Konecranes	4.82e-03	196624	166951	2.12e-19	***	−9.20e-03	168224	149412	1.60e-31	***
Metso	1.24e-02	728656	660427	2.85e-155	***	−8.19e-03	499355	455767	2.91e-186	***
Neste	1.74e-02	815407	798953	1.99e-277	***	6.69e-03	583760	566274	4.94e-167	***
Nokia	1.67e-02	1227666	1192884	9.06e-172	***	−7.57e-03	946953	909957	1.00e-62	***
Nokian Renkaat	2.20e-02	593910	545189	0	***	−3.30e-03	395643	362376	1.16e-22	***
Nordea Bank	1.33e-02	587149	532025	4.94e-14	***	7.32e-03	433685	397123	2.00e-196	***
Outokumpu	5.93e-03	655518	531609	0	***	−2.18e-03	491690	423337	0	***
Pohjola Bank (A)	1.78e-02	217125	190563	6.88e-184	***	−2.69e-04	143691	129464	2.14e-06	***
Rautaruukki	1.10e-02	534820	482416	6.58e-46	***	−1.14e-02	390714	356090	1.36e-90	***
Sampo (A)	9.17e-03	576255	479171	0	***	−4.62e-03	387637	336705	0	***
Sanoma	1.02e-02	85911	64166	1.59e-66	***	5.91e-03	60928	46699	2.30e-88	***
Stora Enso (R)	8.28e-03	118263	100900	1.50e-57	***	3.60e-03	100868	90632	8.71e-44	***
TeliaSonera	9.72e-03	521506	451074	0	***	−1.90e-04	312414	256948	0	***
Tieto	1.42e-02	170061	141868	3.16e-133	***	−1.07e-03	127218	110955	8.38e-34	***
UPM-Kymmene	9.90e-03	405850	354723	4.70e-80	***	−3.76e-03	270916	237695	1.96e-94	***
Wärtsilä	3.17e-02	413348	405382	3.14e-68	***	2.09e-02	304060	297739	6.26e-04	***
YIT	1.54e-02	431844	392657	6.42e-08	***	−1.93e-03	326199	297321	1.30e-34	***

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

trading states and the abnormal trading states. If one of the corresponding links does not exist in one of the networks, we set its weight to zero.

We report our results in Table 4 for non-validated networks. Corresponding results for validated networks, which are rather consistent, are available in Table Appendix E.1. By investigating the resulting *p*-values of the paired tests for information networks inferred from buying behavior (the left-hand side), we find that all networks reject the null hypothesis that the median of paired differences is zero at the significance level of 0.05. This indicates that links between investors become weaker if we incorporate public information in the inference of buy-networks. This is expected, as investors' joint buying behavior is most likely to be driven by their partially similar reaction to public information. In this case, taking it into account in the inference, the weights of the links are reduced.

By analyzing the weights in terms of investors' selling behavior (the right-hand side), we also find that all networks have statistically different weights (in terms of medians) after incorporating the public information in the network inference. Differently from the buying behavior, the median weights are *larger* if the public information is considered. This observation might be related to the fact that investors have different strategies to employ public information for their sell-decisions. Moreover, investors can only sell the securities they already own. Therefore, acting on public or private information should be different from buying behavior (Grinblatt and Keloharju, 2001a).

Overall, we observed low *p*-values from the link weight paired tests in all types of information networks. This suggests that we end up with considerably different investor networks by taking the public information into account in the inference procedure.

4.2.3. Impact of public information on network centrality measures

Our previous observation of the significant changes in the link weights after the public information is taken into account raises the question if other key network properties are affected in the binary networks[8]. To investigate this, we take a look at the distributions of two prominent network centrality measures. In particular, we assess the changes in Katz and closeness centralities. The results of the paired tests for the information networks are summarized in Table Appendix E.2. For information networks inferred from abnormal trading behavior (considering the effect of public information), Katz centralities are generally higher. In comparison, the closeness centralities are lower than in the networks inferred without considering public information. This means that in the networks resulting from our proposed method, the nodes have more connections locally. On the other hand, some of the shortcuts that make distinct parts of the network closer are removed, decreasing the closeness centralities.

[8] To make the links binary, we retain a link if the *p*-value from the hypergeometric test is lower than the significance level of 0.05 adjusted with the Bonferroni correction.

Overall, the results consistently confirm the significant changes in networks. We conclude that the effect of public information on the structure of the information network is substantial and statistically significant. While we provided some intuition for why the public information may cause the networks to change in the ways we have observed, a more in-depth study is outside the scope of this article and will be performed in the future.

4.3. Investor centrality and returns

One of the expected properties in the information networks is the positive relationship between the investors' centrality and their profitability (Ozsoylev and Walden, 2011). More central investors are better informed; therefore, they are expected to make higher profits, which was empirically confirmed by Ozsoylev et al. (2013). Moreover, Walden (2019) shows that the investor's profitability and centrality are closely related and justifies the choice of Katz centrality[9] theoretically. We argue that by filtering the impact of public information on trading behavior, the co-occurrences of the abnormal buying or selling trading states are more likely to be related to the transfer of private information and, thus, better estimates for the information network. To test this, we regress the centrality of the investors against the returns they earn. We expect that information networks inferred with our method by eliminating public information in investor trading synchronization will show a stronger positive relationship between investor centrality and returns.

We aggregate stock-specific networks by taking the union of networks inferred from the buying and selling behavior. This is done separately for the networks inferred using the existing methods and our approach, which eliminated the impact of public information. As a centrality measure, suitable for both directed and undirected networks, we choose Katz centrality[10]. We denote Katz centrality for investor i in the information network derived from her/his trading synchronizations in security k as $C^K_{i,k}$. Next, we follow Ozsoylev et al. (2013) to determine the realized profit. For each investor i and her/his executed trade z in security k, let $Q_{i,k,z}$ be the quantity of shares traded, $P_{i,k,z}$ the transaction price, and $V_{i,k,z} = Q_{i,k,z} \times P_{i,k,z}$ the traded euro volume. Differently from the notation in Eqs. (1–6), where the traded euro volume is calculated for each trading day t, here the volumes are calculated for each transaction z. Next, for each transaction z we define the ΔT-day log return as

$$\mu_{i,k,z} = \text{sign}(z) \times \log\left(P^{\Delta T}_{k,z}/P_{i,k,z}\right), \tag{9}$$

where for security k, $P^{\Delta T}_{k,z}$ is the closing price $\Delta T = 21$ trading days after the transaction z was executed and sign(z) is equal to -1 if z was a sale, and 1 if it was a purchase transaction.[11] Here, $\mu_{i,k,z}$ captures the returns that are generated within a month after a trade. Then we can define the value weighted average return for the investor i in security k as

$$\mu_{i,k} = \frac{\sum_z \mu_{i,k,z} \cdot V_{i,k,z}}{\sum_z V_{i,k,z}}. \tag{10}$$

Finally, we can regress investor centralities $C^K_{i,k}$ against the average value-weighted returns $\mu_{i,k}$:

$$\mu_{i,k} = \beta_0 + \beta_K C^K_{i,k} + \textstyle\sum_l \beta_{\mathbb{1},l} \mathbb{1}_k(l) + \varepsilon_{i,k}, \tag{11}$$

where $\mathbb{1}_k(l)$ are the dummy variables for traded securities and l is one of the 22 investigated stocks (excluding Nokia, as its effects will be captured with the baseline dummy β_0) and $\varepsilon_{i,k}$ is i.i.d. random variable. Additionally, as control variables we take the degree centrality $C^D_{i,k}$, the total number of transactions $Q_{i,k} = \sum_z Q_{i,k,z}$, and the total traded euro volume $V_{i,k} = \sum_z V_{i,k,z}$. We run regressions with ISIN dummy variables with each of the control variables separately and additionally combine the uncorrelated variables into a more complex multivariate regression model. Since Katz and degree centralities are strongly correlated, we only include the Katz centrality in the multivariate regressions (see Table Appendix F.1 for correlations between regression variables). Similarly, the number of trades and the total traded euro quantity are strongly correlated. Therefore we only include one of them in the multivariate regression.

For non-validated networks, coefficients in all regressions are positive, confirming the positive association between returns and investor centralities in information networks observed in the Istanbul Stock Exchange (Ozsoylev et al., 2013). More importantly, when the effects of public information are accounted for in the network inference, the association between future investor returns and their centralities is stronger, with a lower p-value (Table 5). Therefore, the consideration of public information in the network inference not only changes the network topology but also strengthens the relationship between investors' centrality and returns.

Having obtained significant positive associations between centralities and returns, as a robustness check we wanted to see how the association changes with respect to different validation thresholds applied to the p-value from the hypergeometric test Eq. (7). We took 100 threshold levels equally spaced on a log scale between 10^{-10} and 1, used them to filter out the network links and performed the regression analysis (Fig. 2a).

First, Fig. 2a shows that the regression coefficient (on the relation between centralities and future returns) across the different threshold levels is higher when public information is filtered out. We test this with the paired t-test across the threshold levels and find that the observed association between centralities and future returns is statistically higher in

[9] Katz centrality is the special case of the eigenvector centrality (see, e.g., Friedkin, 1991).

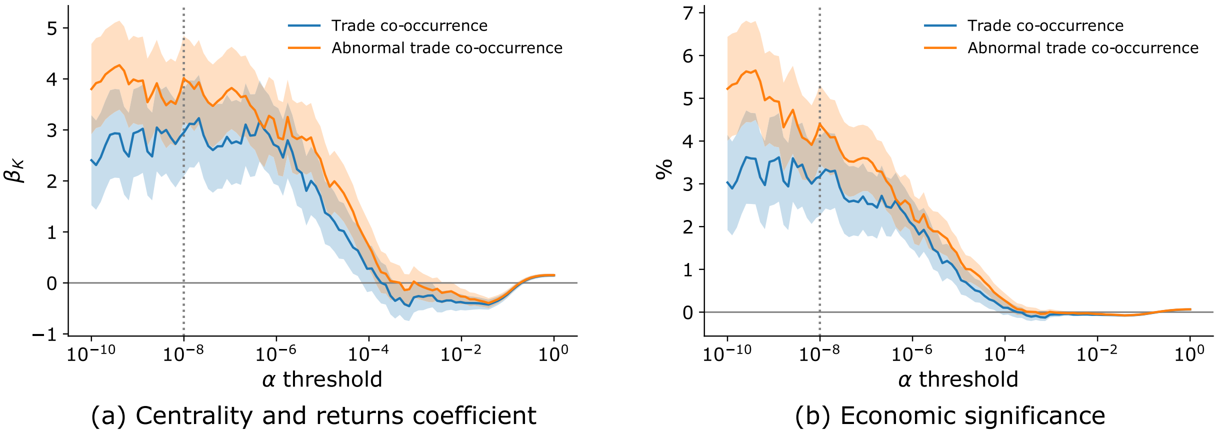
[10] Results are consistent with Eigenvector, Katz, and PageRank centralities.

[11] Here, P_t is not corrected for market splits, therefore we chose to remove the transactions that fell into the $[t-2; t+23]$ trading days period around a split of a stock. This removed approximately 0.5% of transactions.

Table 5
Centrality and returns in **non-validated** networks. The table displays results from regressions of value-weighted returns $\mu_{i,k}$ on Katz centrality $C^K_{i,k}$, degree centrality $C^D_{i,k}$, the normalized quantity $Q_{i,k}$, and normalized traded value $V_{i,k}$ in information networks. Each column represents a linear regression with ISIN dummy variables (coefficients are omitted). The first row displays coefficients while the second row displays the p-values. Degree $C^D_{i,k}$ centrality measures the number of node's links normalized by the maximum possible degree $\max_{i,k}(C^D_{i,k})$ in corresponding networks. Quantity $Q_{i,k}$ is the total number of transactions for each investor in a given security. Value $V_{i,k}$ is investor i euro sum of all transactions executed with a given security. R^2 is the explained variation. N is the number of observations.

	Panel A: Trade co-occurrence networks				
Model	I	II	III	IV	V
Katz centrality, $C^K_{i,k}$	0.1463 (1.53e-3)				0.1312 (5.26e-3)
Degree, $C^D_{i,k}$		0.0035 (1.15e-3)			
Quantity, $Q_{i,k}$			0.0099 (2.29e-2)		0.0076 (8.74e-2)
Value, $V_{i,k}$				0.0092 (2.97e-2)	
R^2	0.023	0.023	0.023	0.023	0.023
N			41965		

	Panel B: Abnormal trade co-occurrence networks				
Model	I	II	III	IV	V
Katz centrality, $C^K_{i,k}$	0.1528 (4.94e-4)				0.1388 (1.9e-3)
Degree, $C^D_{i,k}$		0.0036 (3.72e-4)			
Quantity, $Q_{i,k}$			0.0094 (2.47e-2)		0.0068 (1.1e-1)
Value, $V_{i,k}$				0.0086 (3.2e-2)	
R^2	0.025	0.025	0.025	0.025	0.025
N			41766		



(a) Centrality and returns coefficient

(b) Economic significance

Fig. 2. (a) Centrality and returns coefficient, β_K from Eq. (11), with respect to the threshold applied in network link validation with hypergeometric test (see Eq. 7). (b) Economic significance of β_K, with respect to the threshold applied in network link validation. Coefficients for both network inference methods – from trade co-occurrence Eq. (2) and abnormal trade co-occurrence Eq. (4) – are accompanied with 50% confidence intervals in (a) and (b). The dotted vertical line marks the Bonferroni validation threshold.

networks inferred using our proposed method than the network inference where the public information is not taken into account.

Second, we find that strict network filtering with p-threshold less than 10^{-4} leads to a very strong association between the investor centralities and their future returns. On the other hand, with an intermediate region, roughly between 10^{-3} and 10^{-1}, the association becomes negative. We hypothesize that in this region, the validation procedure primarily removes the true positive links, i.e., making type II errors. In comparison, a stronger filtering starts revealing even though a smaller, but more accurate backbone of the actual information network. Moreover, stricter network filtering also leads to an economically

Table 6
Properties of networks, inferred using trade co-occurrences and abnormal trade co-occurrences for different validation thresholds. Panel A shows the number of communities in the networks, using Clauset-Newman-Moore greedy modularity maximization method. Panel B shows the Network centrality index. Panel C shows the stability of the networks, split into two equally-long periods: before and after 20 September 2007. The total possible number of links, taking the investors present in both periods is 2027091 in the networks inferred from trading and 1819278 in the networks inferred from abnormal trading co-occurrences. Panel D shows the average distance between investors in the network.

Validation theshold		10^{-8}	10^{-6}	10^{-4}	10^{-2}	1
Panel A: Number of communities						
Trade co-occurrence		66	46	7	3	3
Abnormal trade co-occurrence		59	38	6	3	2
Panel B: Network centrality index						
Trade co-occurrence		0.131	0.196	0.340	0.417	0.077
Abnormal trade co-occurrence		0.133	0.202	0.336	0.422	0.078
Panel C: Network stability in time						
Trade co-occurrence	Number of links in first half, k_1	2171	5786	32534	392715	1544072
	Number of links in second half, k_2	2884	8302	46342	476488	1704287
	Number of overlaps, y	1162	1857	6251	131910	1323178
	Expected random overlap $E[y]$	3	24	744	92312	1298186
	Realized and expected overlaps ratio, $y/E[y]$	376.205	78.365	8.404	1.429	1.019
Abnormal trade co-occurrence	Number of links in first half, k_1	1658	4586	26358	333702	1353406
	Number of links in second half, k_2	2390	6483	36611	395545	1496286
	Number of overlaps, y	901	1517	5029	105847	1137915
	Expected random overlap $E[y]$	2	16	530	72553	1113124
	Realized and expected overlaps ratio, $y/E[y]$	413.658	92.827	9.481	1.459	1.022
Panel D: Geographical average distances between investors (km)						
Trade co-occurrence		192.635	208.110	215.465	216.044	216.831
Abnormal trade co-occurrence		190.861	207.923	215.369	216.479	216.977

more significant association between node centrality and returns (Fig. 2b). When Bonferroni multi-test correction is used, one standard deviation change in the investor centrality leads to a 4.4% increase in the expected future returns.

We perform an additional analysis and robustness check by splitting investors into two categories, households and institutions, and estimate the regression coefficient in 100 log spaced thresholds (see Fig. Appendix I.1). The previously observed shape of a curve is repeated in households (see Fig. Appendix I.1a), while institutions seem to have a slightly different pattern. Our findings show a stronger association between centralities and returns in households rather than institutions. This can indicate that private information is more important to households, or this can be due to how institutional data are aggregated. In particular, an institutional ID represents all the traders of that institution, who can have different data sources, different strategies and different social connections. For that reason we refrain from making a generalized comparison between institutions and individual households.

4.4. Further analysis of information networks

Similarly to Ozsoylev et al. (2013), to check whether our inferred investor networks exhibit expected properties of information networks, we perform multiple analyses and compare results obtained from networks inferred with and without taking public information into account. For different link validation thresholds, we investigated (i) the number of communities (Clauset et al., 2004), (ii) the network centrality index (NCI) (Freeman, 1978), (iii) the stability of networks inferred across different periods, and (iv) geographical distances between the investors. Given that investors have private information transfer channels that can be detected by observing their trading synchronization, we would expect (i) to see a number of communities in information networks, (ii) the topology of such networks not to exhibit properties of high centralization, i.e. topology with only a few highly connected hubs, (iii) networks to be fairly stable over time, and (iv) the average distance between investors connected in the inferred information network to be lower than if the connections were randomly shuffled among the same investors.

Our findings are not directly comparable to those observed in (Ozsoylev et al., 2013) due to different data set resolution and analyzed investor set choice. When we compare the results against the networks inferred without taking public information into account, we do not observe any significant differences in the first two analyses (Table 6).

In an information network, with non-centralized information diffusion, we would expect to see multiple communities. However, due to the small size of our investor set and daily resolution of the transaction data, for non-validated networks, we only detect two communities with our proposed method (three when public information is not considered in the network inference) (Table 6, Panel A). This is quite expected as we have chosen a relatively small number of investors that are sufficiently active over multiple securities. When the validation threshold approaches 10^{-8}, the community detection algorithm finds roughly up to 59 (66) communities. However, when looking at NCI in a non-validated network (Table 6, Panel B),

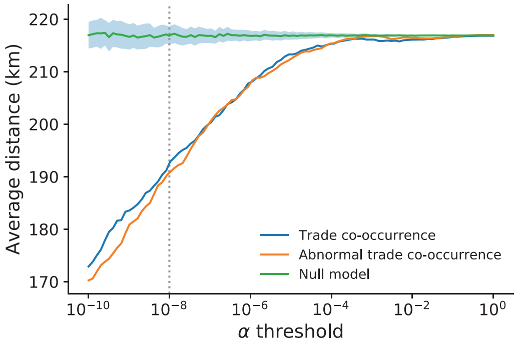


Fig. 3. The average distance in kilometers between investors with respect to network validation threshold applied in network link validation with hypergeometric test (see Eq. 7). The results are calculated from trade co-occurrence Eq. (2) and abnormal trade co-occurrence networks Eq. (4). The number of links in the null model network coincides with the number of links in the network inferred from abnormal trade co-occurrence in each validation threshold. The average distance in null model is accompanied with the standard deviation bands. The dotted vertical line marks the Bonferroni validation threshold.

we see a relatively low value, conflicting with a star-like network structure, and suggesting a non-centralized information diffusion. NCI both for non-validated networks and for validated networks with very harsh thresholds is around 0.1. In the intermediate region where we do not observe the positive association between returns and centralities it increases up to 0.45. In general, NCI ranges from 0 to 1, and it equals 0 in a regular lattice-type network, where all the nodes have the same number of connections, and 1 in a star-like network, where all nodes have only a single connection to one central node.

To test whether the inferred information networks are stable, we split the period into two non-overlapping consecutive periods and infer investor networks for each of them separately. We check how stable those networks are relatively to the Erdős–Rényi random network model (Erdős and Rényi, 1960). The ratio between the observed and expected number of overlapping network links is higher for our model, always larger than one, and increases with harsher threshold levels (Table 6, Panel C). This means that the networks resulting after public information is taken into account are more stable.

Notably, the financial research literature suggests that distance between investors is negatively related to the opportunities to exchange information and share insights about investments. As the physical proximity is one of the major factors in social networks (Backstrom et al., 2010; Baltakys et al., 2018a; Brown et al., 2008; Preciado et al., 2012), we expect to see shorter distances between investors in the better proxies of the information network. Following Baltakys et al. (2018a), we estimate the geographical distance between the investor pairs using information about their postal codes. In addition, for each given threshold we create an empirical null model where the same number of links is randomly drawn 100 times from all possible connections between investors, and each time the average distance between randomly connected investors is calculated. From the obtained empirical distributions, we estimate the means and the standard deviations. Our results show that the average distances between connected investors decrease from roughly 217 to 170 km using stricter validation thresholds, and they are considerably lower than expected under the null model (Fig. 3).

This observation should not be connected to a local bias (Grinblatt and Keloharju, 2001a; Ivković and Weisbenner, 2005; Seasholes and Zhu, 2010; Zhu, 2002), because local bias should be related to the portfolio composition rather than trade timing. Nor should it be localized public information signals, because company news is widespread and should not have local effects only.

5. Discussion and conclusions

Different investors adopt different trading strategies based on the available resources – time, capital, education, personal experiences, and available public and/or private information. People seeking to reduce the cognitive burden while making investment decisions may decide to follow other investors' trading strategies. Indeed, as multiple studies have shown, social influence increases market participation (Brown et al., 2008; Heimer, 2014; Hong et al., 2004), and, moreover, individuals are more likely to act on public information when private personal information supports it (Katz and Lazarsfeld, 1955; Rogers et al., 1962). These observations reinforce our intuition that investors may be inclined to use private information channels that they trust to gain better profits at a smaller cost.

In this paper, we proposed a new approach to infer information networks from investor trading histories. The novelty of our approach is in taking the effects of public information into account in the investor network inference procedure. This is done by assigning abnormal trading states to investors, which allows us to link investor pairs based on their abnormal trading co-behavior. By comparing the information network resulting from our approach with the two most prominent investor network inference methods (Ozsoylev et al., 2013; Tumminello et al., 2012), we found that our approach yields a significantly different network topology. Moreover, the use of our method strengthens the association between investor centrality and their future returns across different validation thresholds. Furthermore, we found that, with some exceptions, such association becomes both statistically and economically more significant for networks validated with harsher thresholds.

Overall, the use of our method yields significantly different network topology and stronger relation between investor's centrality and returns. At the same time, no major differences between our and existing methods are observed when analyzing properties of information networks. In particular, the use of either method shows that the topological structure of the resulting information networks is in stark contrast with the one expected under a centralized information diffusion: it contains (i) multiple communities, (ii) low network centralization measures, (iii) significant link persistence, and (iv) short physical distance between connected investors.

Our approach requires investor-specific models to predict their trading volumes. Depending on the complexity of the model used, a significant number of trading observations may be required for each investor for reliable model estimation. In our future research, we will investigate if more general models can describe investor behavior, which would allow us to use investor group-level instead of individual-level models. While public shareholder registration data sets are available to regulators, we hope data sets from other markets will become accessible to academic researchers.

Acknowledgements

M.B. received funding from the doctoral school of Tampere University. M.B. is grateful for the grants received from The Foundation for the Advancement of Finnish Securities Markets, The Finnish Foundation for Share Promotion, and The Nordea Bank Foundation. K.B. would like to thank The Foundation for the Advancement of Finnish Securities Markets, The Finnish Foundation for Share Promotion, and Finnish Foundation for Technology Promotion for the received funding. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Authors would like to thank the anonymous referee for excellent comments that significantly improved our paper.

Research Data Statement

The data set analyzed in the current study is not publicly available and cannot be distributed by the authors because it is a proprietary database of Euroclear Finland. The database can be accessed for research purposes under the nondisclosure agreement by asking permission from Euroclear Finland.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jedc.2021.104217.

References

Ahern, K.R., 2017. Information networks: Evidence from illegal insider trading tips. J. Financ. Econ. 125 (1), 26–47.
Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International conference on World Wide Web. ACM, pp. 61–70.
Baltakienė, M., Baltakys, K., Kanniainen, J., Pedreschi, D., Lillo, F., 2019. Clusters of investors around initial public offering. Palgrave Commun. 5 (1), 1–14.
Baltakys, K., 2019. Investor networks and information transfer in stock markets, http://urn.fi/URN:ISBN:978-952-03-1280-0.
Baltakys, K., Baltakienė, M., Kärkkäinen, H., Kanniainen, J., 2018. Neighbors matter: geographical distance and trade timing in the stock market. Finance Res. Lett..
Baltakys, K., Kanniainen, J., Emmert-Streib, F., 2018. Multilayer aggregation with statistical validation: application to investor networks. Sci. Rep. 8 (1), 8198.
Baltakys, K., Kanniainen, J., Saramäki, J., Kivela, M., 2020. Trading Signatures: Investor Attention Allocation in Stock Markets. Available at SSRN.
Barber, B.M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Rev. Financ. Stud. 21 (2), 785–818.
Berkman, H., Koch, P., Westerholm, P.J., 2014. Inside the Director Network: When Insiders Trade Outside Stocks. Technical Report. SSRN Working Paper, http://ssrn. com/abstract= 2424527.
Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31 (3), 307–327.
Brown, J.R., Ivković, Z., Smith, P.A., Weisbenner, S., 2008. Neighbors matter: causal community effects and stock market participation. J. Finance 63 (3), 1509–1531.
Challet, D., Chicheportiche, R., Lallouache, M., Kassibrakis, S., 2018. Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. Adv. Complex Syst. 21 (08), 1850019.
Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70 (6), 066111.
Coleman, T.F., Li, Y., 1996. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. SIAM J. Optim. 6 (4), 1040–1058.
Colla, P., Mele, A., 2010. Information linkages and correlated trading. Rev. Financ. Stud. 23 (1), 203–246.
Cordi, M., Challet, D., Kassibrakis, S., 2019. The market nanostructure origin of asset price time reversal asymmetry. arXiv preprint arXiv:1901.00834.
Cutler, D.M., Poterba, J.M., Summers, L.H., 1988. What Moves Stock Prices? Technical Report. National Bureau of Economic Research.
Erdős, P., Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci 5 (1), 17–60.
Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Social Netw. 1 (3), 215–239.
Friedkin, N.E., 1991. Theoretical foundations for centrality measures. Am. J. Sociol. 96 (6), 1478–1504.
Grinblatt, M., Keloharju, M., 2001. How distance, language, and culture influence stockholdings and trades. J. Finance 56 (3), 1053–1073.
Grinblatt, M., Keloharju, M., 2001. What makes investors trade? J. Finance 56 (2), 589–616.
Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Am. Econ. Rev. 70 (3), 393–408.
Gutiérrez-Roig, M., Borge-Holthoefer, J., Arenas, A., Perelló, J., 2019. Mapping individual behavior in financial markets: synchronization and anticipation. EPJ Data Sci. 8 (1), 10.
Haldane, A.G., May, R.M., 2011. Systemic risk in banking ecosystems. Nature 469 (7330), 351–355.
Hautsch, N., Schaumburg, J., Schienle, M., 2015. Financial network systemic risk contributions. Rev. Finance 19 (2), 685–738.
Heimer, R.Z., 2014. Friends do let friends buy stocks actively. J. Econ. Behav. Organ. 107, 527–540.
Hellwig, M. F., 1980. On the aggregation of information in competitive markets.
Hong, H., Kubik, J.D., Stein, J.C., 2004. Social interaction and stock-market participation. J. Finance 59 (1), 137–163.

Ilmanen, M., Keloharju, M., 1999. Shareownership in Finland. Finnish J. Bus. Econ. 48 (1), 257–285.

Ivković, Z., Weisbenner, S., 2005. Local does as local is: information content of the geography of individual investors' common stock investments. J. Finance 60 (1), 267–306.

Jackson, M.O., 1991. Equilibrium, price formation, and the value of private information. Rev. Financ. Stud. 4 (1), 1–16.

Katz, E., Lazarsfeld, P.F., 1955. Personal Influence, Glencoe, ILL. The FreePress.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 1315–1335.

Ladley, D., 2013. Contagion and risk-sharing on the inter-bank market. J. Econ. Dyn. Control 37 (7), 1384–1400.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., 2005. Geographic routing in social networks. Proc. Natl. Acad. Sci. 102 (33), 11623–11628.

Mantegna, R.N., 2020. Clusters of traders in financial markets. In: Complexity, Heterogeneity, and the Methods of Statistical Physics in Economics. Springer, pp. 203–212.

Nadini, M., Rizzo, A., Porfiri, M., 2020. Reconstructing irreducible links in temporal networks: which tool to choose depends on the network size. J. Phys. 1 (1), 015001.

Ozsoylev, H.N., Walden, J., 2011. Asset pricing in large information networks. J. Econ. Theory 146 (6), 2252–2280.

Ozsoylev, H.N., Walden, J., Yavuz, M.D., Bildik, R., 2013. Investor networks in the stock market. Rev. Financ. Stud. 27 (5), 1323–1366.

Pratt, J.W., 1959. Remarks on zeros and ties in the Wilcoxon signed rank procedures. J. Am. Stat. Assoc. 54 (287), 655–667.

Preciado, P., Snijders, T.A., Burk, W.J., Stattin, H., Kerr, M., 2012. Does proximity matter? Distance dependence of adolescent friendships. Social Netw. 34 (1), 18–31.

Ranganathan, S., Kivelä, M., Kanniainen, J., 2018. Dynamics of investor spanning trees around dot-com bubble. PloS one 13 (6), e0198807.

Rogers, E.M., et al., 1962. Diffusion of innovations.. Diffus. Innov..

Seasholes, M.S., Zhu, N., 2010. Individual investors and local bias. J. Finance 65 (5), 1987–2010.

Siikanen, M., Baltakys, K., Kanniainen, J., Vatrapu, R., Mukkamala, R., Hussain, A., 2018. Facebook drives behavior of passive households in stock markets. Finance Res. Lett..

Tuminello, M., Micciche, S., Lillo, F., Piilo, J., Mantegna, R.N., 2011. Statistically validated networks in bipartite complex systems. PloS one 6 (3), e17994.

Tumminello, M., Lillo, F., Piilo, J., Mantegna, R.N., 2012. Identification of clusters of investors from their real trading activity in a financial market. N. J. Phys. 14 (1), 013041.

Walden, J., 2019. Trading, profits, and volatility in a dynamic information network model. Rev. Econ. Stud. 86 (5), 2248–2283.

Zhu, N., 2002. The Local Bias of Individual Investors. Yale ICF Working Paper No. 02-30.