

# Investor Networks in the Stock Market\*

Han Ozsoylev<sup>†</sup>    Johan Walden<sup>‡</sup>    M. Deniz Yavuz<sup>§</sup>    Recep Bildik<sup>¶</sup>

February 18, 2012

## Abstract

We study the trading behavior of investors in an entire stock market. Using an account level dataset of all trades on the Istanbul Stock Exchange in 2005, we identify traders with similar trading behavior as linked in an empirical investor network (EIN). Consistent with the theory of information networks, we find that central investors earn higher profits and trade earlier than peripheral investors with respect to information events. Overall, our results support the view that information diffusion among the investor population influences trading behavior and profits.

**Keywords:** *Information networks, Information diffusion, heterogeneous investors, portfolio choice, trading.*

---

\*We thank seminar participants at the 2012 meetings of the American Finance Association, the Berkeley finance lunch seminar, the Berkeley-Stanford seminar, HKUST, the 2011 Indiana State Conference, Johns Hopkins University, the Swedish Institute for Financial Research (SIFR), University of Minnesota, University of Miami, and UT Austin. The Department of Information Technology at Uppsala University, and especially Per Lötstedt have been tremendously supportive. Thanks to Niclas Eriksson and Ludvig Larruy at the same department for excellent research assistance with the program implementation and computational analysis of the data, and also to Andreas Kieri and Joakim Saltin for continuing their work. We also thank Mehmet Ihsan Canayaz for research assistance on news collection and Sabanci University School of Management for their hospitality while part of this work was carried out. We are grateful to John Beshears, Jonathan Cohn, Jennifer Conrad, Nicolae Garleanu, Simon Gervais, Joel Hasbrouck, Craig Holden, Ron Kaniel, Shimon Kogan, Christine Parlour, Raghu Rau, Mark Seasholes, Sheridan Titman, Laura Veldkamp and Jeffrey Zwiebel for valuable comments and suggestions. Finally, we thank Gil Shallom for developing the initial code.

<sup>†</sup>Saïd Business School, University of Oxford, Park End Street, Oxford, OX1 1HP, United Kingdom. *E-mail:* han.ozsoylev@sbs.ox.ac.uk, *Phone:* +44-1865-288490. *Fax:* +44-1865-278826.

<sup>‡</sup>Corresponding author: Haas School of Business, University of California at Berkeley, 545 Student Services Building #1900, CA 94720-1900. *E-mail:* walden@haas.berkeley.edu, *Phone:* +1-510-643-0547. *Fax:* +1-510-643-1420. Support from the Institute for Pure and Applied Mathematics (IPAM) at UCLA is gratefully acknowledged.

<sup>§</sup>Purdue University, Krannert School of Management, 403 West State Street, West Lafayette, IN 47907. *E-mail:* myavuz@purdue.edu

<sup>¶</sup>Istanbul Stock Exchange. *E-mail:* recep.bildik@imkb.gov.tr

# 1 Introduction

What determines the heterogeneous trading behavior of individual investors in the stock market? Several suggestions have been made in the literature. First, it has been argued that some investors are more sophisticated than others. For example, unsophisticated investors underperform by trading too often, in the wrong stocks and by investing too much or too little (Odean (1999); Barber et al. (2009); Fedyk et al. (2010)). Second, variations in liquidity needs make it possible for some investors to earn a premium by acting as market makers (Grossman and Miller (1988)). Third, investors may prefer different investment “styles” (Brown and Goetzmann (1997); Barberis and Shleifer (2003)), for example, preferring value stocks over growth stocks.<sup>1</sup> Fourth, investors have different information; well-informed investors are able to make better investment decisions than less informed investors (Hellwig (1980); Grossman and Stiglitz (1980); Kyle (1985)).

In this paper, we focus on the fourth determinant and study investor behavior through the lens of information networks (Colla and Mele (2010), Ozsoylev and Walden (2011), Han and Yang (2011)). Loosely speaking, an information network describes how heterogeneous information diffuses over time among a population of agents. Agents who are centrally placed in such a network tend to receive information signals early, whereas agents who are in the periphery tend to receive them later. As a result, if such networks are present among agents in a stock market, their trading behavior and profitability are influenced by their position in the network. Further, the dynamics of aggregate asset prices depend on general topological properties of the network. In its most general form, an information network describes how agents sequentially act upon information over time, without any further specification about the channels through which the diffusion of information occurs. Those channels could, e.g., be direct communication between individuals, Internet discussion boards and news papers. Regardless of the specific underlying communication channel, however, it is interesting to learn whether information networks plays any role in aggregate price dynamics and profitability of network participants.

Identifying the underlying information network in the entire stock market is of course a major challenge. Our first contribution in this paper is to develop a method to proxy for the market’s information network, using observable data. The general idea is that information links may be identified from realized trades, since traders who are directly linked in the network will tend to trade in the same direction in the same stock at a similar point in time. Using this approach, we identify an Empirical Investor Network (EIN), and in a stylized model we use simulations to show that the true information network is indeed well estimated by the EIN.

We calculate the EIN, based on account level trading data that covers all trades in the Istanbul Stock Exchange (in Turkey) in 2005. We first verify that the EIN is fairly stable over time. We test the stability by dividing our sample period into two six month sub-periods and define an EIN for each of these periods. The overlap between the two EINs is strongly significantly different from the overlap of two randomly generated networks. We then verify that some investors systematically trade before their neighbors, and that such early trading is positively related to centrality. Together, these results support the view that the EIN captures information diffusion. This provides our second contribution.

---

<sup>1</sup>There is also a large literature that explains heterogeneous portfolio holdings with hedging motives (see, e.g., Mayers (1973); Bodie et al. (1992); Massa and Simonov (2006); Parlour and Walden (2011), Betermeier et al. (2011)). For example, investors who differ in the human capital risk they face will hedge this risk in capital markets and thereby choose different investment portfolios. Such considerations therefore also lead to different trading behavior over time, although typically at a low frequency of rebalancing. Similarly heterogeneous preferences, e.g., different risk aversion, also induce trading but also typically at a low frequency of rebalancing. We include such trading motives in our broad definition of “investment styles.”

We study the relationship between centrality and profits of investors in the network, and find substantial support for a positive relationship. We find that a one-standard deviation increase in centrality, all else equal, leads to a 0.3-2.8% increase in profits (over a 30-day period), per unit traded depending on the specification. That is, an investor who purchased and sold stocks for a total of 10,000 Turkish Lira in 2005 made on average 30-280 Turkish Lira higher profits than an investor with one standard deviation lower centrality, all else equal. These results are obtained after controlling for other variables such as trading volume, so the tests distinguish between investors who are central in the information network from investors who just trade a lot. Showing the positive relationship between centrality and profits is our third contribution.

Finally, as a fourth contribution, we document that centrality is directly related to acting early on information. We identify several information events in the news that caused stock prices to move, and find that central agents in the network tended to trade — in the right direction — before peripheral agents. Again, the results are statistically and economically significant. We also show that our results are robust to several variations, providing further support for the information diffusion story over alternative explanations, like liquidity provision, style investing, momentum strategies, and market microstructure effects.

Our study supports a view of the stock market as a place where information is incorporated into asset prices through gradual decentralized diffusion. Information networks provide an intermediate information channel, in between the public arena where news events and prices themselves make some information available to all investors, and the completely local arena of private signals and inside information. It may then not be so surprising that large stock market movements often are unaccompanied by public news events, as noted in Cutler et al. (1989), and Fair (2002), and that stock market returns, trading volume and number of trades vary substantially over time, as analyzed in Gabaix et al. (2003).

Our paper belongs to the literature on heterogeneous information, networks, and trading in stock markets. There is extensive evidence of frequent communication among stock market investors, and this evidence suggests that investors exchange information about the stocks that they trade. Shiller and Pound (1989) vividly demonstrates this: the authors survey 131 institutional investors in the NYSE and ask them what prompted their most recent stock purchase or sale. The majority asserts that it was their discussions with their peers. Ivković and Weisbenner (2007) find similar evidence for households, and Hong et al. (2004) provide further evidence that fund managers' portfolio choices are influenced by word-of-mouth communication. Feng and Seasholes (2004) find that Chinese trades are highly correlated when divided geographically. Cohen et al. (2008) posit that there is communication via shared education networks between fund managers and corporate board members, manifested in the abnormal returns managers earn on firms they are connected to through their network (see also Das and Sisk (2005), Fracassi (2009) and Pareek (2009)). Shive (2010) develops an epidemic model of investor behavior and shows that it predicts individual investor trading, using Finnish data. Duffie et al. (2009) develop a dynamic equilibrium search model, in which information diffusion occurs when agents with heterogeneous information meet randomly. Our paper is also related to the literature that studies the relationship between news, investor behavior and stock returns, e.g., Tetlock (2007) and Tetlock (2011). Our study contributes to the literature by providing — to the best of our knowledge — the first market-wide study of information diffusion in a stock market, and its effects on the entire investor population's behavior and trading profits.

The rest of the paper is organized as follows: In the next section, we introduce a stylized information network model to motivate the connection between agents' centrality and trading profits, and the

methodology used to create the EIN. In Section 3, we describe the data and also show some summary statistics. In Section 4 we present our main findings on the stability of networks, order in which investors trade, relationship between centrality and profits, and between centrality and trading early on information events. Concluding remarks are made in Section 5. Some additional analyses are delegated to the appendix.

## 2 Model

We introduce a simple model of information diffusion in a stock market, to describe the concept of an information network, to study the relationship between network centrality and trading profits, and to understand how an information network may be detected from observed trading behavior.

Let us for simplicity use a network structure according to Figure 1, in which there are  $N_I = 21$  agents in a network, as an example. Each node (circle) represents an agent (investor, trader), and each edge (line) represents a direct link between two agents, i.e., that the two agents are connected. In other words, linked agents are neighbors in the network. These connections are bidirectional, i.e., if agent  $i$  is connected to agent  $j$ , then  $j$  is connected to  $i$ . It is straightforward to generalize the model to networks with unidirectional links. For technical reasons, we always assume that an agent is connected to himself.

In addition to the agents in the network, we assume that there is a large number,  $N_U$  of uninformed noise traders, whose trading motives we do not model and who randomly take on opposite sides of trades.<sup>2</sup> These are represented by the isolated nodes in the lower right corner of the figure. Altogether there are  $N = N_I + N_U$  traders in the model.

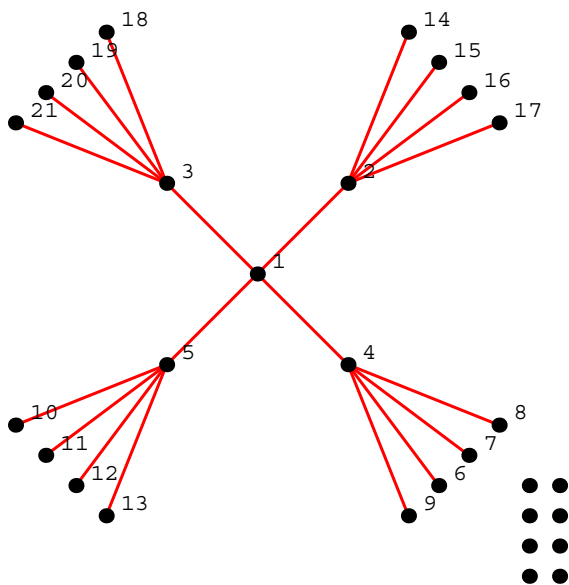


Figure 1: Network of 21 agents in a market. Each agent is represented by a node (a filled circle). An edge (line) between two agents represents that these are connected in the network, i.e., that they are neighbors. In addition, there is a large number of liquidity traders, represented by the isolated nodes at the lower right corner.

<sup>2</sup>We could alternatively have excluded the noise traders from the model, and instead assumed that agents in the network are hit by random liquidity shocks, which forces them to take on opposite sides of trades, with similar results.

Trading occurs at discrete times,  $t = 0, 1, 2, \dots$ . At each point in time, each of the  $N_I$  agents in the network receives a distinct signal about stocks in the market, i.e., agent  $i$  receives signal  $s_i^t$  at time  $t$ . We denote the set of signals agent  $i$  has received at time  $t$  by  $\mathcal{I}_i^t$ , and it then follows that  $s_i^{t'} \in \mathcal{I}_i^t$  for  $t' \leq t$ .<sup>3</sup>

For simplicity, we assume that only one signal in each time period, agent  $n_t$ 's signal, is valuable. Thus, at time  $t$ , agent  $n_t$  receives a signal and trades. All the other signals at time  $t$  are worthless, the other agents understand this, and therefore do not trade. The expected profits from agent  $n_t$ 's trade is  $\pi > 0$ . We do not specify further what types of trades the signal leads to. For example, it could involve buying one specific stock, a portfolio of stocks, short-selling, etc. The important fact is that the expected profits of the trade are positive, above and beyond expected market returns, including the inferences that uninformed investors may draw about the signal when observing the market. For simplicity, we assume that there is a noise trader willing to take the opposite position in the trade (with expected profits of  $-\pi$ ), whereas agents in the network only trade when they receive information.<sup>4</sup>

Now, agent  $n_t$  may “share” his signal with one of his neighbors between  $t$  and  $t + 1$ . Specifically, for each of his neighbors, there is a probability of  $q_1$  that agent  $n_t$  shares his information. For example, given the network in Figure 1, if agent 1 received the initial signal, the probability is  $q_1$  that he will share it with any one of agents 2, 3, 4 and 5. For simplicity, we assume that he never shares it with more than one of his neighbors. We could, for example, think of a situation in which the time period is one day, and for each of agent  $n_t$ 's neighbors there is a probability of  $q_1$ , that  $n_t$  and the neighbor have lunch together after agent  $n_t$  received the signal and traded on it, in which case he shares his information. Other diffusion mechanisms are also possible, e.g., Internet discussion boards. In its most general form, an information network describes how agents sequentially act upon information over time. It is therefore equivalent to a specification of information diffusion among a population in that an information diffusion process *defines* an information network, regardless of the channel through which information diffusion occurs. This is our interpretation of what constitutes an information network, and we stress that although it may represent information diffusion through social communication channels — through a “social network” — other channels are also possible.<sup>5</sup>

Given that information is shared, the receiving agent — let us call him  $n_t^2$  — then trades at  $t + 1$ , however his expected trading profit is lower than that of agent 1, say  $\pi/2$ , in line with the assumption that, as time passes, the expected profits from trading on the information declines. This could, e.g., be because agent  $n_t$  has already traded and some of his information is already incorporated into prices. The signal may also be slowly diffusing into the market through other channels. We thus have that  $s_{n_t^2}^t \in \mathcal{I}_{n_t^2}^{t+1}$ .

In a similar manner, agent  $n_t^2$  shares his signal between  $t + 1$  and  $t + 2$ , but for simplicity we assume that he shares the signal with all his neighbors (except with agent  $n_t$ , who already knows it) with probability one. They then trade at  $t + 2$  and realize lower expected profits than agent  $n_t^2$ , say  $\pi/4$ . At  $t + 3$ , the signal is completely incorporated into the stock market's prices and no further profits can be made. It

---

<sup>3</sup>More generally,  $\mathcal{I}_i^t$ , represents the signals that agent  $i$  uses in his trading decisions at time  $t$ . The distinction is relevant in cases where agents may need time to process signals, in which case they may act upon signals at later points in time than when actually receiving them. In this paper we use the terminology that an agent “receives” a signal when it becomes available for him to use in his trading decisions, which could thus be at a later point than when the signal is actually received.

<sup>4</sup>These assumptions are, of course extremely stylized, but as shown in Ozsoyev and Walden (2011) the framework can be developed under much greater generality in general equilibrium with agents who fully incorporate the information provided by the stock market in their trading decisions. Our partial equilibrium setting allows for a much simplified analysis.

<sup>5</sup>As another example, exactly the same dynamics would arise if the nodes represented different news outlets — or more correctly, agents receiving their information from such outlets — the leaves being local newspapers and the central node being a national news agency. News may reach a local news paper first and then diffuse to the news agency, but for any given local newspaper, the direction of the diffusion is usually the opposite.

would of course be easy to extend the model to longer sequences of information diffusion.

What are the expected profits per unit time of each agent in the network in this case? Let us define agent  $n$ 's (first-order) *degree*,  $D_n$  as his number of neighbors (including himself, according to our technical convention). Thus, agent 1's degree is  $D_1 = 5$ , since he has 5 neighbors (including himself), whereas agent 2's degree with the same argument is  $D_2 = 6$ , and agent 6's degree is  $D_6 = 2$ . Further, let us define an agent's second order degree,  $D_n^2$  as the number of agents within a "distance" of two from the original agent, i.e., taking into account not only neighbors but also neighbors of neighbors. Thus, the second order degree of agent 1 is  $D_1^2 = 21$ , whereas the second order degree of agent 2 is  $D_2^2 = 9$  and that of agent 6 is  $D_6^2 = 6$ . It then follows immediately that the expected profits of trades at any time  $t > 1$  for any agent,  $n$ , in the network is

$$P_n \stackrel{\text{def}}{=} E[\Pi_n] = \frac{\pi}{N_I} \left( 1 + \frac{q_1}{2}(D_n - 1) + \frac{q_1}{4}(D_n^2 - D_n) \right). \quad (1)$$

Therefore, agent 1 will have the highest expected profit per unit time ( $P_n = \frac{\pi}{21} (1 + 24\frac{q_1}{4})$ ), followed by agents 2-5 (who have  $P_n = \frac{\pi}{21} (1 + 13\frac{q_1}{4})$ ), and agents 6-21 will have the lowest expected profits ( $P_n = \frac{\pi}{21} (1 + 6\frac{q_1}{4})$ ).

We note that agent 1 has a higher expected profit than agent 2, although agent 2 actually has a higher degree, i.e., has more direct neighbors. This is because agent 1 is more *central* than agent 2. Although agent 1 has fewer direct neighbors, his neighbors are more connected (on average) than those of agent 2, and he will therefore capture more benefits from indirect information than agent 2. That more centrally placed agents in an information network, in general make higher profits also follows from the general equilibrium model in Ozsoylev and Walden (2011).

There is a large literature on how to measure centrality. We will use the notion of *eigenvector centrality*, which is a widely used measure with many application areas, see, e.g., Chung and Li (2006) and references therein. To define this measure we first need to introduce some notation. A general network of  $N$  agents can be represented by a neighborhood (adjacency) matrix,  $\mathcal{E} \in \{0, 1\}^{N \times N}$ , with  $\mathcal{E}_{ij} = 1$  if investors  $i$  and  $j$  are directly connected and  $\mathcal{E}_{ij} = 0$  otherwise.<sup>6</sup> The bidirectionality of connections implies that  $\mathcal{E}$  is symmetric (i.e.,  $\mathcal{E}_{ij} = \mathcal{E}_{ji}$  for all  $i$  and  $j$ ). Moreover, under the assumption that an agent is connected with himself,  $\mathcal{E}_{ii} = 1$  for all  $i$ .

The total market of investors will thus be represented by an  $N \times N$  neighborhood matrix, where we use the convention that the first  $N_I$  agents are the ones in the network, and the remaining  $N_U$  are the noise traders (each of which is only connected to himself). The market's neighborhood matrix can therefore be decomposed into

$$\mathcal{E} = \begin{bmatrix} \mathcal{E}^I & \mathbf{0} \\ \mathbf{0}^T & \mathbf{I} \end{bmatrix}, \quad (2)$$

where  $\mathbf{I}$  is a  $N_U \times N_U$  identity matrix, representing that noise traders are connected to themselves, but have no other noise traders as neighbors in the network,  $\mathbf{0}$  is a  $N_I \times N_U$  matrix of zeros representing that there are no connections between agents in the network and noise traders, and  $\mathcal{E}^I$  is a  $N_I \times N_I$  matrix representing the connections among agents in the network. For example, the matrix representation of the network in Figure 1, is given in Figure 2 below, where it is assumed that there are  $N_U = 29$  noise traders, so that the total number of traders is  $N = 21 + 29 = 50$ . In the figure, dots represent connections,

---

<sup>6</sup>We use the following matrix notations: A matrix is defined by the  $[\cdot]$  operator on scalars, e.g.,  $\mathcal{E} = [e_{ij}]_{ij}$ . We write  $(\mathcal{E})_{ij}$  for the scalar in the  $i$ th row and  $j$ th column of the matrix  $\mathcal{E}$ , or, if there can be no confusion, we write it as  $\mathcal{E}_{ij}$ .

i.e., elements for which  $\mathcal{E}_{ij} = 1$ . The upper left part of the matrix represents the agents in the network,  $\mathcal{E}^I$ . For example, focusing on the first row, the five first elements are nonzero, showing that agent 1 is connected to himself, and agents 2-5, respectively. The lower right part of the matrix (elements 22-50) is diagonal, representing the unconnected noise traders.

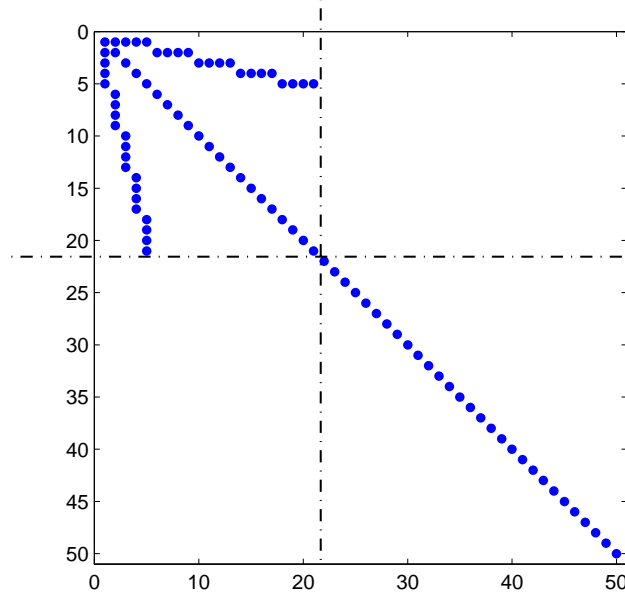


Figure 2: *The neighborhood matrix for the network with  $N_I = 21$  agents who trade for information purposes and  $N_U = 29$  noise traders.*

The neighborhood matrix,  $\mathcal{E}$  contains all relevant information about the network. For example, a vector with the degree of each agent — the number of direct neighbors that agent has — is constructed by summing all the elements in each row of the neighborhood matrix,

$$D_i = \sum_j \mathcal{E}_{ij}, \quad \text{or in vector form} \quad D = \mathcal{E}\mathbf{1}, \quad (3)$$

where  $\mathbf{1}$  is a vector of ones. More generally, the  $m$ th order degree vector, which contains the number of agents that are at most  $m$  links away from each agent, is constructed as follows: Define the  $m \geq 1$  degree neighborhood matrix,  $\mathcal{E}^m \in \{0, 1\}^{N \times N}$ , as

$$\mathcal{E}^m = \chi(\underbrace{\mathcal{E} \times \mathcal{E} \cdots \times \mathcal{E}}_{m \text{ times}})$$

Here,  $\mathcal{E} \times \mathcal{E}$  denotes standard matrix multiplication, and  $\chi(\mathbf{A})$  is the “indicator matrix” of an arbitrary matrix  $\mathbf{A}$ , such that  $(\chi(\mathbf{A}))_{ij} = 1$  if  $\mathbf{A}_{ij} \neq 0$  and  $(\chi(\mathbf{A}))_{ij} = 0$  if  $\mathbf{A}_{ij} = 0$ . Then the  $m$ th-order degree vector has elements

$$D_i^m = \sum_j \mathcal{E}_{ij}^m, \quad \text{or in vector form} \quad D^m = \mathcal{E}^m \mathbf{1}. \quad (4)$$

We note that two agents,  $i$  and  $j$  are at distance  $m$  from each other in the network if  $\mathcal{E}_{ij}^m - \mathcal{E}_{ij}^{m-1} = 1$ ,

where we use the convention that  $\mathcal{E}^0 = I$  is the identity matrix. We use this to define the distance function,  $d$ , or  $d^{\mathcal{E}}$  when we wish to stress the network that the distance function is based upon, and thus  $d^{\mathcal{E}}(i, j) = m$  in this case. Here we use the convention that if there is no path between agents  $i$  and  $j$ , i.e., if  $\mathcal{E}_{ij}^m = 0$  for all  $m$ , then the distance between these agents equals infinity,  $d^{\mathcal{E}}(i, j) = \infty$ .

We are now in a position to formally define an information network, given an information diffusion mechanism among agents:

**Definition 1** Consider a population of agents among which heterogeneous information signals,  $s_i^t$ , diffuse over time. Then  $\mathcal{E}$  defines the population's information network, if for all agents  $i, j$  and times  $t, t'$ , the probability that  $s_i^t \in \mathcal{I}_j^{t'}$  is

- zero when  $d^{\mathcal{E}}(i, j) > t' - t$ ,
- greater than zero when  $d^{\mathcal{E}}(i, j) \leq t' - t$ .

It is easy to check that given the information diffusion mechanism between agents just described, the information network is indeed the one shown in Figure 1.

Similarly to the degree vector, a vector  $C$  where the  $i$ th element represents agent  $i$ 's (eigenvector) centrality can be constructed. Formally, let  $C_i$  denote the centrality of investor  $i$ . By letting  $i$ 's centrality score be proportional to the sum of the scores of all the investor's neighbors, we derive

$$C_i = \frac{1}{\lambda} \sum_j \mathcal{E}_{ij} C_j, \quad \text{or in vector form} \quad C = \frac{1}{\lambda} \mathcal{E} C. \quad (5)$$

The proportionality constant,  $\lambda$ , is an eigenvalue of  $\mathcal{E}$  and  $C$  is the corresponding eigenvector. The eigenvector corresponding to the largest eigenvalue is the centrality vector.<sup>7</sup> For large matrices, an efficient way of solving (5) is by using *power iterations*.<sup>8</sup>

Higher order degrees and centrality both proxy for the amount of information an investor obtains from longer distances than direct links in the network and we would therefore expect both to be related to agents' trading profits. Specifically, (5) implies that agent  $i$ 's centrality does not only depend on how many neighbors he has, but also on how central these agents are (in contrast to the degree measure). Thus, centrality takes into account not only one's number of neighbors but also each neighbor's number of neighbors and each neighbor's neighbor's number of neighbors and so on, just like we argued that the information advantage an agent has in the market does not only depend on that agent's direct neighbors, but also on how connected he is at higher orders. If information spreads over longer distances, centrality may therefore be a better summary of information advantage than any fixed-order degree. We will therefore mainly work with  $C$ .<sup>9</sup>

---

<sup>7</sup>For an  $N \times N$  matrix,  $\mathbf{A}$ , the set of eigenvalues contains the roots of an  $N$ th order polynomial equation, given by the determinant  $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ , where  $\mathbf{I}$  is an  $N \times N$  identity matrix. The neighborhood matrix,  $\mathcal{E}$ , has only nonnegative elements. It therefore follows from the Perron-Frobenius theorem that it has a real maximal eigenvalue, and that the associated eigenvector has only nonnegative elements. This is the centrality vector. The only potential issue is uniqueness, since  $\mathcal{E}$  may not be irreducible, but this has not caused a problem in our tests.

<sup>8</sup>Specifically, given an estimate of the centrality vector,  $C^k$ , an updated estimate is obtained by performing the iteration  $C^{k+1} = \frac{1}{\|C^k\|} \mathcal{E} C^k$ , where  $\|C^k\|$  is some suitable chosen normalization of  $C^k$  (e.g., the mean-square norm). If  $\mathcal{E}$  contains relatively few non-zero elements—in other words, if the matrix is *sparse*—and the largest eigenvalue is significantly larger than the second largest eigenvalue, then each iteration can be calculated quickly and convergence to the true centrality vector is obtained in few iterations.

<sup>9</sup>There are also purely technical advantages of working with centrality: to calculate the  $m$ th degree vector,  $\mathcal{E}^m$  needs to

A closely related measure is the *rescaled centrality*,  $C/D$ , i.e., the ratio between centrality and degree. As we shall see when discussing how to estimate the neighborhood matrix from empirical data, this measure may be more robust than pure centrality, since there may be noise traders who trade a lot for other reasons than information. These traders will typically come out as very connected, which will influence their centrality positively too. However, their rescaled centrality will be low, as it should be, since those traders are not central in the information network.

In Table 1, we have listed the centrality, first-order degree, second order degree, rescaled centrality, and expected profits per unit time (denoted by  $P$ ) for all the agents in the network, given that  $\pi = 10N_I$ ,  $q_1 = 0.3$ , and  $N = 100$ . We see that second order degree and centrality rank agents according to their expected profits correctly. These measures rank agent 1 the highest, followed by agents 2 – 5, 6 – 21 and then finally the noise traders, 22 – 100. The reason why agent 1 gets a higher centrality score than agent 2 is that although he has fewer neighbors than agent 2 (5 versus 6, including himself), agent 1’s neighbors are “worth more” than agent 2’s, because they in turn are on average more central than agent 2’s neighbors, who mainly are in the periphery of the network. In contrast, the first-order degree measure ranks agents 2-5 above agent 1, since they have more direct neighbors. Our general point, however, is that when information travels over large distances through gradual information diffusion, measures of higher order connectivity ( $C$  or  $D^2$ ) will be more closely related to trading profits than the measure of direct connections ( $D$ ). Rescaled centrality does not provide the right measure in this example, since it ranks investor 6 over investor 2. As mentioned, the main value of rescaled centrality is for empirical estimates when the true information network is not observable. We have also included  $w$  in the table, defined as the average probability that an agent trades before his neighbors. We note that, also in line with the previous arguments, more centrally placed agents in the information network tend to trade before their neighbors more often than less peripheral agents.

Agent	$C$	$D$	$D^2$	$C/D$	$P$	$w$
1	0.5	5	21	0.1	2	0.762
2 – 5	0.354	6	9	0.059	1.41	0.759
6 – 21	0.125	2	6	0.0625	1.04	0.111
22 – 100	0	1	1	0	-0.31	0.5

Table 1: Centrality, first-order degree, second order degree, rescaled centrality, and expected profits per unit time for agents in network according to Figure 1, when  $\pi = 10N_I = 210$ ,  $q_1 = 0.3$  and  $N = 100$ .

## 2.1 Estimating the neighborhood matrix

In practice the information network is not observable, but based on the key observation that agents who are connected in the network will tend to trade in similar stocks in the same direction at similar points in time, we propose the following method to construct an empirical proxy — an *Empirical Investor Network* (EIN):

**Definition 2** *The Empirical Investor Network,  $\mathcal{E}^{\Delta t, M}$ , in a stock market which operates over some finite time period, is defined such that for each pair of investors,  $i, j \neq i$ ,  $\mathcal{E}^{\Delta t, M} = 1$  if and only if agents  $i$  and*

---

be calculated, which is a major task if  $N$  is large. The reason is that even though  $\mathcal{E}$  is a sparse object,  $\mathcal{E}^m$  is much less sparse, leading to severe memory requirements. Moreover,  $m$  matrix by matrix multiplications are needed to calculate  $\mathcal{E}^m$ , leading to severe CPU requirements. In contrast, to calculate the largest eigenvector using the power method, only matrix by vector calculations using  $\mathcal{E}$  are needed, and convergence typically occurs within less than fifty iterations.

$j$  traded in the same stock in the same direction within a time window of  $\Delta t$ , at least  $M$  times over the full time period.

In the rest of this section, we focus on the case where the threshold for a connection is  $M = 1$ , and the time period is one unit of time, so that agents who trade within the period  $[t, t + 1]$  for some  $t$  are connected, i.e., we focus on  $\mathcal{E}^{1,1}$ . In our empirical tests in the next section, we will vary  $\Delta t$  from a minute up to a day, and we will also explore using a higher threshold,  $M = 10$ .<sup>10</sup>

We calculate the EIN for a slightly generalized version of the information diffusion process just described. Specifically, we assume that the agent who gets the initial signal, with probability  $q_1$  shares the information with each of his neighbors, and that the events of sharing with different neighbors are independent. Also, we assume that any receiver of the initial agent’s signal, shares it with probability  $q_2$  (independently, just like for the initial receiver) with each of his neighbors. Thus, second stage receivers may share with only a subset of their neighbors, or even with nobody.

Both these generalization make the inference problem of backing out the true information network from observed trades more difficult, since false links will be drawn between agents who trade, not because they are directly connected, but because they are both connected to agents who provided them with signals at the same point in time. If the EIN still provides a good proxy of the true network, this then provides stronger support for the validity of the approximation.

We simulate trades in the network in Figure 1, with  $N = 100$  agents, over 50 trading periods, with probabilities  $q_1 = 0.25$ ,  $q_2 = 0.5$ . We choose a higher per-agent probability for information diffusion at the second stage, since it seems natural to assume that agents are pickier in who they share information with early on, when information is more proprietary.

An example of a realized EIN is shown in Figure 3. We see that the general structure of the true network is identified, although not every link is correct. For example, in the upper left part of the EIN, which represents the agents in the information network, there are several elements just off the diagonal that are nonzero, representing links between agents, although no such links exist in the true information network. This is, for example the case for agents 20 and 21, who are incorrectly linked in the EIN. The reason is that agent 3 received a signal that he shared with agents 20 and 21, who then traded simultaneously and who were thereby falsely identified as directly linked, although they are in practice only indirectly linked through their common connection with agent 3. Similarly, erroneous links occur in the part of the matrix with uninformed agents. These links arise when two agents happen to take the opposite position of their informed counterparties, at similar points in time.

In the informed part of the network matrix (the first  $N_I \times N_I$  in the upper left corner), there are 42 agents, who are incorrectly identified as being linked. Also, there are 4 agents who are actually linked, but who are not estimated to be linked. Thus, in total, 46 elements are misclassified, corresponding to about 10% of the total number (441) of elements of  $\mathcal{E}^I$ . In the noise trader part of the network, there are 126 incorrect links, scattered randomly, corresponding to about 2% of the total number of elements (6,241).

Does the EIN provide a “good” proxy of the true underlying network? It turns out that it does. Not only is the fraction of misclassified elements low, but the randomness of these elements also helps in that the misclassified elements do not systematically distort the key properties of the network. The EIN, of

---

<sup>10</sup>Note that our definition of EIN is different from the one taken in Adamic et al. (2010), who identify two investors as connected if they traded with each other. Such traders are on the opposite side and will thus not be viewed as connected in our model. Also, note that in line with the previous section, all links in the EIN are bidirectional, i.e., if investor  $i$  is linked to  $j$ , then  $j$  is also linked to  $i$ . Finally, recall that per definition, agents are self connected,  $\mathcal{E}_{ii}^{\Delta t, M} = 1$ .

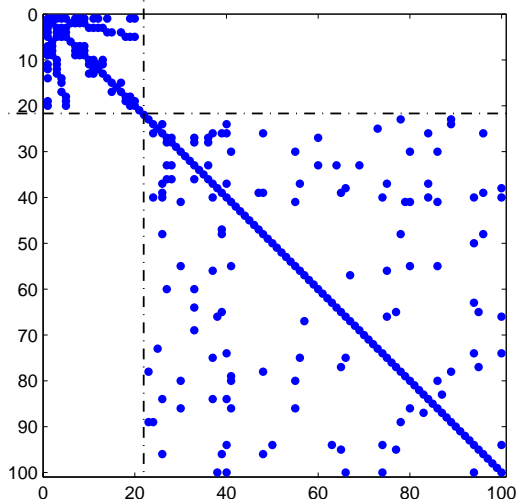


Figure 3: *The empirical investor network, EIN, generated in a simulation, when the true network has  $N = 100$  traders.*

course, depends on the exact (random) realization of signals and trades, which we simulate with 1,000 Monte Carlo simulations.

Our simulations show that the centrality vector of the true network,  $C$  and that of the EIN,  $C_{EIN}$  are highly correlated, on average  $Corr(C, C_{EIN}) = 0.64$ . Further, it is immediately clear from Figure 3 that the EINs will be a lot more similar across simulations, compared with what would be the case if agents traded randomly.

There is also a significantly positive average correlation between agents' profits and centrality,  $Corr(C_{EIN}, P) = 0.51$  on average. The relationship between the EIN's degree and trading profits,  $P$ , is on average much weaker,  $Corr(D_{EIN}, P) = 0.19$ . This is in line with the idea that centrality is more important for agents' trading profits than their first-order connections. Of course, the correlation between true centrality and trading profits is even higher,  $Corr(C, P) = 0.76$  on average, but the EIN captures a significant part of this positive relation. Finally, rescaled centrality is slightly more correlated with profits than centrality when these empirical measures are used,  $Corr(C_{EIN}/D_{EIN}, P) = 0.55$  on average, providing some support for the argument that the measure is more robust because it "punishes" noise traders, who trade a lot but are not central in the information network.

We note that an alternative representation of how "close" two agents are may be provided through some type of a correlation measure. The first principal component of the agents' correlation matrix would then correspond to our centrality measure. For our data, the network approach has some clear advantages: First, whereas our EIN will typically be a sparse object, correlation matrices are typically dense, which would make the analysis computationally infeasible. Second, it is unclear how such a correlation matrix would even be defined, given that we do not know agents' portfolios. Instead, our approach is to define agents as being "close" if they carry out similar trades at similar points in time. Given that the vast majority of agents in our sample trade quite infrequently, the trades of such agents will overlap at most a few times, and any ad hoc definition of a correlation matrix of trades will therefore look very similar to our EIN. The EIN has a clear and simple interpretation, and we therefore prefer the network approach in this application.

### 3 Description of the data

#### 3.1 The Istanbul Stock Exchange

The Istanbul Stock Exchange (ISE) was founded as an autonomous, professional organization in early 1986. The ISE is the only corporation in Turkey established to provide trading in equities, bonds and bills, revenue-sharing certificates, private sector bonds, foreign securities and real estate certificates as well as international securities. All ISE members are incorporated banks and brokerage houses. There were exactly 100 ISE members in 2005.

The ISE is an order-driven, multiple-price, continuous auction market with no dedicated market makers or specialists. A computerized system automatically matches buy and sell orders on a price and time priority basis. The buyers and sellers enter the orders through their workstations located at the ISE building and also in their headquarters. It is a blind order system with trading ISE members identified upon matching. The system enables members to execute several types of orders such as “limit,” “limit value,” “fill or kill,” “special limit,” and “good till date” type orders. Members can enter buy and sell orders with various validity periods of up to one trading day. Unmatched orders without a specific validity period are canceled at the end of the trading session.

The stock trading activities are carried out in two separate sessions, 9:30-12:00 for the first session and 14:00-16:30 for the second session, on workdays. Settlement of securities traded in the ISE is realized by the ISE Settlement and Custody Bank Inc. (Takasbank), which is the sole and exclusive central depository in Turkey. Turkey has a liberal foreign exchange regime with a fully convertible currency. In 2005, the value of one Turkish Lira (TL) varied between 0.7-0.8 USD. Since August 1989, the Turkish stock and bond markets have been open to foreign investors without any restrictions on the repatriation of capital and profits. At the start of our sample period, the vast majority (94.7%) of the institutional investors in our sample were foreigners.

ISE ranks 19th across the world with market capitalization of 201 billion dollars in 2005 (World Development Indicators). The average daily trading volume ranged between approximately USD 300 and 700 million. The turnover ratio of the ISE was 155% in 2005, which was comparable to the turnover ratio of %129 for the US. There is only one time zone in Turkey.

#### 3.2 The raw data

Our dataset contains all the trades on the ISE over a 12 month period, January 1-December 31, 2005. During this period there were 303 stocks actively traded in the market. In the data, each trader is identified by a unique account number, and for each trade the following information is available: time of trade, stock ticker, number of shares traded, price, account number of purchaser and seller, purchaser and seller types (private, institutional or brokerage house trading on its on own account), and whether the trade was a short sale. In total there were 580,142 active accounts during the time period. Of these, 489 were classified as institutional accounts and the remaining 579,653 were classified as individual accounts. On average about 200,000 trades were executed on a day when the ISE was open.

#### 3.3 The Empirical Investor Network

We calculate the EIN for the market, using the threshold  $M = 1$  and varying the length of the time window,  $\Delta t$ , between one minute and 30 minutes. We will subsequently extend the time window to a

whole day. The results are similar for different window lengths. This is quite unsurprising, since two investors who are directly linked when a window length of  $\Delta T$  is chosen, are typically also indirectly connected (at a higher degree than one) with a window length  $\Delta T' < \Delta T$ .<sup>11</sup> The main difference is that we find a stronger relationship between centrality and profitability for longer window lengths.

By using a window length of no more than one day, we separate information driven “fast” trading from other types of trading such as portfolio rebalancing, momentum investing, style investing etc., which we typically think of as occurring over lower frequencies. For example, momentum strategies are typically implemented over a 3-month to 1-year horizon, whereas the impact of value and size strategies are rarely studied over shorter than monthly horizons. In contrast, the EIN is constructed to capture information diffusion effects at horizons of, say, up to about a week, taking higher order effects into account (i.e., degrees of higher order than one). We verify the robustness of the results for longer time windows in Section 4.5.

In Table 2, we provide summary statistics for the EIN, using window lengths between 1 minute to half an hour. We see that the total number of links is substantial, e.g., 1.66 Billion for the EIN with a 15 minute window length, or about 2,860 on average per investor. Also, some investors are extremely highly connected; the most connected investor when the 15 minute window length is used has almost a quarter million links, and is thereby directly connected to over 40% of the other investors. We suspect that these are (unofficial) market makers that provide liquidity—an investor group that is not part of our theoretical model—and therefore come out as extremely connected although they are not part of the information network. We have no direct way of identifying such traders, but we believe that a good proxy is given by identifying traders with extremely high trading frequencies, which we will control for. In total, the matrix is very sparse. Even with the 30-minute time window, agents are on average only connected to a small fraction, 0.7% (3,881/580,142), of the population.

$\Delta T$	1 min	5 min	15 min	30 min
# links	0.51B	1.01B	1.66B	2.25B
Average #links	876	1,736	2,860	3,881
Fraction of links	0.15%	0.3%	0.5%	0.7%
# investors with one link	18,927	1,939	192	32
maximum # links	153,923	204,635	249,064	312,807

Table 2: Summary statistics for EIN, with window length 1, 5, 15 and 30 minutes.

Once we have generated the EIN, first-order degrees,  $D$ , higher order degrees,  $D^m$ , and centrality,  $C$ , are straightforward to calculate, using (3-5). The distributions of  $D$  and  $C$  are quite heavy-tailed, and their ranges are over several orders of magnitude. For our statistical tests, we will therefore also work with their logarithms, i.e., we define the vectors  $d$  and  $c$ , where  $d_i = \log(D_i)$  and  $c_i = \log(C_i)$ .

### 3.4 Trading volume, number of trades, and trading profits

We need a measure of individual agents’ trading profits, represented by a vector  $P$ , where the  $i$ th element is the realized profit of agent  $i$ . We also wish to control for agents’ trading volumes and number of trades. We therefore construct a vector of unsigned trading volumes,  $V$ , where  $V_i$  is the total total value (in TL)

<sup>11</sup>The only time that this is not the case is when there was a a time period larger than  $\Delta T'$ , during which a stock was not traded in the market. In this case the chain of links is broken, and two traders will be directly linked in the  $\mathcal{E}^{\Delta T'}$  network, but not even indirectly linked in the  $\mathcal{E}^{\Delta T}$  network.

of purchases and sales that investor  $i$  executes over the full time period (1 year). Similarly, we define the vector of number of trades of each individual investor,  $N$ , over the total time period. We also define the log-counterparts,  $v$  and  $n$ , i.e., vectors with  $v_i = \log(V_i)$  and  $n_i = \log(N_i)$ .

To measure trading profits, we use the method developed in Barber et al. (2009), but focusing on individual investors' trades rather than on investor groups. Briefly, we define a window length,  $\Delta T_P$  which we set to 30 days (we have experimented with shorter lengths, e.g., using one week windows, with qualitatively similar results). For each trade,  $z$ , we calculate the realized profit as

$$P_{iz} = (\text{\#shares purchased}) \times (Q^{t+\Delta T_P} - Q^t),$$

where  $Q^{t+\Delta T_P}$  is the closing price of the stock 30 days after the trade (or, if the market is closed on that day, the closing price the nearest open day after),  $Q^t$  is the price at which the stock was traded, and ( $\text{\#shares purchased}$ ) is negative for an investor on the sell side of a trade. Here,  $Q$  is corrected for stock splits, and takes dividend payments into account.

We then define the total profit of an investor as the sum of all profits from individual trades,  $P_i = \sum_z P_{iz}$ . Throughout the paper, we use the index,  $z$ , to enumerate trade events, both for individual trades (e.g.,  $P_{iz}$  for investor  $i$ 's trade  $z$ ), and for aggregate trades (e.g.,  $P_z$  for trade  $z$  in the market).

We note that our data does not contain any information about investors' portfolios, so we can not calculate the return on these portfolios. Neither can we calculate the total value of an investor's portfolio. In principle, over a long enough period, we could "build" the portfolios by adding up investors' trades, but our sample period is not long enough to do this. Another limitation is that we can not identify a trader who uses multiple accounts.

Our profit measure captures returns that are generated within a month after a trade. Given our focus on information that is diffused into the market relatively quickly, we believe that this window is long enough. Returns over longer time horizons will not be captured by this profit measure, but investors who trade and realize returns at higher frequencies will be measured correctly, on average. For example, assume that an investor has positive information about a stock, buys it (this is trade  $z$  at  $t$ ), and that it subsequently generates high returns over the next week after which the investor sells it (this is trade  $z'$  at  $t'$ ). These two trades then contribute to the profit measure by

$$\begin{aligned} P_{iz} + P_{iz'} &= (\text{\#shares purchased}) \times \left( (Q^{t+\Delta T_P} - Q^t) - (Q^{t'+\Delta T_P} - Q^{t'}) \right) \\ &= (\text{\#shares purchased})(Q^{t'} - Q^t) + (\text{\#shares purchased})(Q^{t+\Delta T_P} - Q^{t'+\Delta T_P}). \end{aligned}$$

Given that the trade was profitable, the first term will be positive, whereas the second term on average will be zero, so given that current information shocks are uncorrelated with future information shocks, profits realized over a shorter period than a month will also be captured.

The profit measure  $P$  will also capture market movements, i.e., a trader may be profitable because the market happened to go up during the period in which he traded, although he had no valuable information about the stock. To adjust for market movements, we also define  $P^e$  as the excess profit over market returns, using a market corrected price at  $t + \Delta T_P$  of  $\tilde{Q}^{t+\Delta T_P} = Q^{t+\Delta T_P} \times \frac{Q_M^t}{Q_M^{t+\Delta T_P}}$ , where  $Q_M^t$  is the price of the market at time  $t$ .<sup>12</sup> It is unclear whether market returns *should* be included in the profit measure or not, since it could be that the valuable stock information happened to actually apply to all firms in

---

<sup>12</sup>Here, we use the ISE 100 index as a proxy for the market.

the market. We therefore use both the excess profits and the gross profits in our analysis.

The profit vectors,  $P$  and  $P^e$ , do not take trading volumes into account. We divide these profits by the total trading volume, to get normalized profit vectors,  $\mu$  and  $\mu^e$ , i.e.,

$$\mu_i = \frac{P_i}{V_i}, \quad \text{and} \quad \mu_i^e = \frac{P_i^e}{V_i}, \quad (6)$$

describe the normalized gross and excess profits for trader  $i$ , respectively.

### 3.5 Summary statistics

We provide summary statistics for the variables in Table 3, where we have used the 30-minute window for the EIN,  $\Delta T = 30$  minutes, and the 30-day window for profits,  $\Delta T_P = 30$  days. Several observations are in place: (i) The way that profits are defined, the mean of  $P$  and  $P^e$  are both identically equal to zero, since there are always investors on both sides of a trade; (ii)  $C$ ,  $D$ , and  $N$  are all severely right skewed which can be seen from the mean being much higher than the median. Also, their standard deviations are high, consistent with heavy-tailed distributions. We verify that the distributions are indeed heavy-tailed in Appendix C; (iii)  $C$  and  $D$ , as well as  $c$  and  $d$ , are highly correlated. Nevertheless, we shall see that the additional information provided by centrality beyond what is provided by connectedness is important in explaining investor performance.

	Mean	Std. dev.	Median	$C$	$D$	Corr $N$	$P$	$P^e$	$V$
$C$	5.86	11.7	1.17	1					
$D$	3,881	9,327	673	0.97	1				
$N$	149	1,467	8	0.40	0.51	1			
$P$	0	2.0E5	-19.0	0.015	0.025	0.13	1		
$P^e$	0	1.3E5	-2.4	0.013	0.023	0.12	0.69	1	
$V$	9.1E5	20.4E6	11,340	0.21	0.30	0.83	0.21	0.18	1
	Mean	Std. dev.	Median	$c$	$d$	Corr $n$	$\mu$	$\mu^e$	$v$
$c$	0.13	2.36	0.16	1					
$d$	6.53	2.03	6.51	0.86	1				
$n$	2.41	2.00	2.08	0.75	0.87	1			
$\mu$	-0.014	0.085	-0.038	0.052	0.055	0.083	1		
$\mu^e$	-0.058	0.074	-0.0012	-0.0004	-0.0053	0.018	0.86	1	
$v$	9.34	2.95	9.34	0.66	0.77	0.84	0.050	0.0064	1

Table 3: Summary statistics for degree ( $D$ ), centrality ( $C$ ), Number of trades ( $N$ ), profits ( $P$ ,  $P^e$ ), Volume ( $V$ ), log-degree ( $d$ ), log-centrality ( $c$ ), log-number of trades ( $n$ ), normalized profits ( $\mu$ ,  $\mu^e$ ) and log-volume ( $v$ ).

The distributions of  $C$ ,  $D$ ,  $N$ ,  $P$  and  $V$  are intimately related to asset pricing dynamics, and may therefore be informative about the aggregate behavior of stock markets, as highlighted by Gabaix et al. (2003). According to their arguments, aggregate stock market dynamics can be explained by the so-called Pareto distributions in the cross section of investors. Here, a vector whose elements are drawn from a distribution that satisfies  $\mathbb{P}(X \geq x) \sim x^{-\alpha}$  in the tail of the distribution is said to be *power-law* or *Pareto* distributed, with *tail exponent*  $\alpha$ .

In Figure 4, we show the tails of the empirical cumulative distribution functions of  $C$ ,  $D$ ,  $N$ ,  $P$  and  $V$ , plotted in log-log coordinates. In log-log coordinates, a Pareto distribution will show up as a

straight line, and the slope is (minus) the tail exponent. A thinner-tailed distribution, e.g., log-normal or exponential, will show up as a concave function. We see that profits ( $P$ ) and trading volume ( $V$ ) seem to be fairly well approximated by a straight line. Centrality ( $C$ ) and degree ( $D$ ) on the other hand seem to be more curved, whereas the number of trades ( $N$ ) lies somewhere between. In the appendix, we verify these results with statistical analysis, and further relate the observed distributions to the literature on heavy-tails in stock markets, identifying interesting challenges for future research.

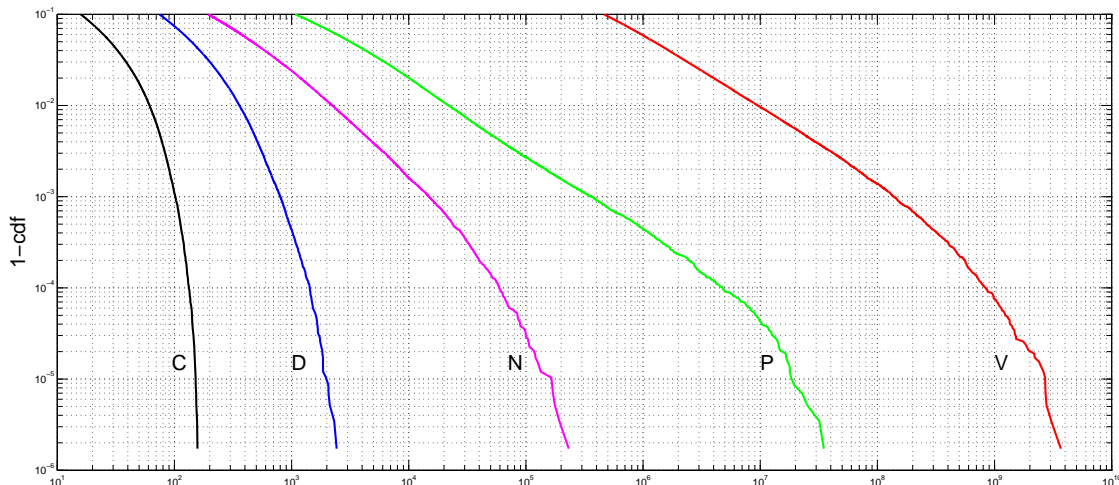


Figure 4: Tail distribution of centrality ( $C$ ), connectedness ( $D$ ) (measured in hundreds of connections), number of trades ( $N$ ), profits ( $P$ ) in Turkish Lira and trading volume ( $V$ ) in Turkish Lira.

In Table 4 we have divided the total sample into the subgroups of institutional and individual investors. The 489 institutional investors behave quite differently than the individual investors. They are on average more central and connected; the average centrality of institutional investors is 39.7 versus 5.83 for individual investors, and the average degree is 40,012 versus 3,850. Also, not surprisingly, institutional investors trade in much larger volumes. Since individual investors make up the vast majority, the summary statistics of the total investor pool are almost identical to the summary statistics of the individual investors, as is seen by comparing Tables 3 and 4. The only number that is significantly different in the two tables is average trading volume where the institutional investors, although they make up less than a per-mille of the total investor pool, increase the average trading volume by about 10% when they are included. An implication of the dominance of individual investors is that our results are not affected by whether we include or exclude institutional investors. We will therefore usually include them, but verify that the results do not change when they are excluded — for the sake of robustness.

## 4 The EIN, centrality, profits and information

### 4.1 Stability of EIN over time

For the EIN to be consistent with an information network, we would expect it to be relatively stable over time. Equivalently, for information networks to provide a meaningful concept and to be measurable, they cannot change too fast. A simple test of such stability is to divide the total time period of one year

	Individual			Institutional		
	Mean	Std. dev.	Median			
$C$	5.83	11.7	1.17	39.7	34.9	32.3
$D$	3,850	9,172	672	40,012	46,600	24,433
$N$	144	1,350	8	6,805	18,390	1,460
$P$	23.7	2.0E5	-19.0	-28,070	1.0E6	2.98
$P^e$	-31.2	1.3E5	-2.4	37,040	3.6E5	8,680
$V$	8.3E5	18.3E6	11,310	9.9E7	2.9E8	1.3E7

Table 4: Summary statistics of institutional and individual investors.

into two sub-periods of six months each, calculate EINs for both sub-periods,  $\mathcal{E}_1$  and  $\mathcal{E}_2$ , and see whether they are more similar than what they would be, if randomly generated.

Obviously, the test will depend on our assumptions about the data generating process for the EINs. The simplest null hypothesis is that these are completely random (except of course for the self-connection between an investor and himself, which is always present), i.e., that if the matrix  $\mathcal{E}_1$ , with  $N$  investors, contains  $k_1$  links, then for each pair of investors,  $i$  and  $j \neq i$ , the chance to be linked is  $\frac{k_1}{K}$ , where  $K = N(N - 1)/2$  is the total number of possible (bidirectional) links. This corresponds to a situation where the data generating process for  $\mathcal{E}_1$  was such that links were randomly added until the matrix had in total  $k_1$  elements.

We let  $y$  denote the number of overlaps between the two EINs, i.e., the number of investor pairs that are linked in both  $\mathcal{E}_1$  and  $\mathcal{E}_2$ . Given that both  $\mathcal{E}_1$  and  $\mathcal{E}_2$  are completely random (with the given data generating process), and that  $k_1 \ll K$ ,  $k_2 \ll K$ , it follows immediately that the expected number of overlaps is approximately<sup>13</sup>

$$E_{\text{Completely random}}[y] \approx \frac{k_1 k_2}{K}. \quad (7)$$

We compare the realized and expected number of overlaps, for the EINs generated with 1-minute and 5-minute time windows, in Table 5. We do this for different choices of the threshold,  $M$ , of the number of trades needed for two agents to be treated as connected in the network. We let  $M$  vary between 1-80. Clearly, the hypothesis of complete random network generating processes for the EIN's can be strongly rejected. In fact, as seen in Table 5, the likelihood of being linked is between 72.2 and 26,200 times higher than what is predicted under the hypothesis of completely random network generating processes, depending on the window length and the link threshold.

Now, obviously the EINs are not completely random; if they were, the degree distributions would be Poisson distributed, but our detailed analysis of the degree distribution in Appendix C shows that the true distribution has heavier tails. A better specified test for stability is therefore to study the number of overlaps, given the (heavy-tailed) degree distributions observed in practice. We describe the details of how such a *degree adjusted* measure is defined in Appendix A, and show that the overlap with this measure is still substantially higher than under the null: 6.09 times higher with the 5-minute window and 7.55 times higher with the 1-minute window. As also discussed in the appendix, the statistical significance of these results is huge.

<sup>13</sup>Here, the approximation is that we treat the addition of links as “draws with replacement”, whereas in practice there is no replacement, i.e., in practice the probability that a new link in  $\mathcal{E}_2$  overlaps with one in  $\mathcal{E}_1$  depends on how many links already exist in  $\mathcal{E}_2$ . The error introduced by this approximation is marginal, given that  $k_1 \ll K$  and  $k_2 \ll K$ .

$\Delta T=1$ minute					
Connection threshold	1	10	20	40	80
$k_1$	129,146,847	11,174,905	5,014,735	2,150,618	903,097
$k_2$	136,493,437	12,006,001	5,539,690	2,442,503	1,057,607
# Overlaps, $y$	11,860,359	1,347,214	659,874	314,779	148,704
$E_{\text{Completely random}}[y]$	104,750	793	165	31	6
$y/E_{\text{Completely random}}[y]$	113.2	1,690	3,997	10,084	26,200
$E_{\text{Degree adjusted}}[y]$	1,570,908				
$y/E_{\text{Degree adjusted}}[y]$	7.55				
$\Delta T=5$ minutes					
$\Delta T=5$ minutes					
Connection threshold	1	10	20	40	80
$k_1$	259,906,612	33,510,862	16,238,861	7,420,656	326,895
$k_2$	274,034,135	35,975,750	17,924,953	8,449,474	3,817,082
# Overlaps, $y$	30,556,857	4,659,221	2,400,323	1,180,224	559,647
$E_{\text{Completely random}}[y]$	423,237	7,164	1730	373	74
$y/E_{\text{Completely random}}[y]$	72.2	650	1,388	3,168	7,548
$E_{\text{Degree adjusted}}[y]$	5,017,607				
$y/E_{\text{Degree adjusted}}[y]$	6.09				

Table 5: Stability of empirical investor networks. The total number of potential connections between agents is  $K = N(N - 1)/2 = 1.68 \times 10^{11}$  (counting the relationship that investors  $i$  and  $j$  are linked as one link, i.e., not double counting bidirectional links), where  $N = 580,142$  is the number of investors.

## 4.2 Sequencing of trades

A crucial feature of the information network model is that some traders systematically trade before others. For example, in the network example introduced in Section 2 (see Figure 1), agent 2 will trade before agent 6 more often than not, since the only time agent 6 trades before agent 2 is when he gets the initial signal, whereas there are several equally probable scenarios in which agent 2 trades before agent 6. This is a distinguishing feature between the information story and several alternative stories of investing, e.g., style investing (broadly defined) and liquidity provision. Without further assumptions, two liquidity providers who trade in the same stock will tend to trade ahead of each other about half of the time each, as will two style investors using the same investment style.

To study whether there are systematic differences in the sequence in which connected investors trade, we proceed as follows. For any two investors,  $i$  and  $j$ , we observe how many times they traded in the same direction in the same stock within the time window, which we denote by  $M_{ij}$ . Of these  $M_{ij}$  trades, investor  $i$  traded before investor  $j$   $K_{ij}$  times, and investor  $j$  before  $i$   $M_{ij} - K_{ij}$  times, so the investor who went first most times did it  $\eta^{M_{ij}} \stackrel{\text{def}}{=} |K_{ij} - (M_{ij} - K_{ij})| = |2K_{ij} - M_{ij}|$  more times than the trader who went last most times.

It follows that the expected value of  $\eta^{M_{ij}}$ , given that in each trade there is a probability of  $p$  that  $i$  traded before  $j$ , and defining  $M = M_{ij}$ , is

$$E[\eta^M | p] = \sum_{k=0}^M \binom{M}{k} |2k - M| p^k (1-p)^{M-k}, \quad (8)$$

i.e., it is the expectation of the absolute value of a binomially distributed random variable. The null

hypothesis we wish to test is then whether  $p = 1/2$  in our empirical network.

We calculate the average  $\eta$  of all pairs of investors in the network,  $i$  and  $j$ , such that  $M_{ij} = M$ , for each  $1 \leq M \leq 210$ , and denote this by  $\bar{\eta}^M$ . Under the null hypothesis that  $p = 1/2$ , we then have that  $E[\bar{\eta}^M] = E[\eta^M | p = 1/2]$ . We show the results in Table 6 for various  $M$  between 5 and 210. We see that the results are inconsistent with a probability of  $p = 1/2$ . In fact, the GMM estimate (choosing  $\hat{p}$  such that  $\bar{\eta}^M = E[\eta^M | p = \hat{p}]$ ) of the probability that one investor trades before another is about  $\hat{p} = 0.75$  for large  $M$ . We note in passing that given the large number of observations, the statistical significance with which we can reject that  $p = 1/2$  is huge.

$M$	5	10	30	60	90	120	150	180	210
$E[\eta^M   p = 1/2]$	1.88	2.46	4.33	6.15	7.55	8.72	9.76	10.69	11.55
$\bar{\eta}^M$	3.69	7.15	19.54	36.85	51.36	68.02	76.77	97.38	107.0
$\hat{p}$	0.87	0.85	0.82	0.80	0.78	0.78	0.76	0.76	0.75
# observations	2.80E7	1.23E7	1.92E6	5.33E5	2.19E5	1.30E5	7.10E4	5.55E4	3.76E4

Table 6: Average  $\eta^M$ , compared with expected  $\eta^M$  given  $p = 1/2$  for various  $M$ . The probability is significantly higher than  $1/2$ . When GMM is used to match  $E[\eta^M | p] = \bar{\eta}^M$ ,  $p$  is about 0.75-0.8 for large  $M$ .

For low  $M$ , the estimated probability is even higher. However, for low  $M$  we face the issue that some orders are split and thereby executed as multiple trades, artificially boosting the observed asymmetry of trades. For example, assume that investor  $i$  submits a buy order for 200 shares in a stock and that the execution of the order is split into two trades of 100 shares at  $t$  and  $t + 2$  minutes, respectively. Now, assume that investor  $j$  submits a buy order for 100 shares that is executed in one trade at  $t + 3$  minutes. Finally, assume that these trades constitute the only source of connections between investors  $i$  and  $j$ . Then we get  $M_{ij} = K_{ij} = 2$ , and  $K_{ji} = 0$ , leading to  $\eta^{M_{ij}} = 2$ . However, in reality this was one trade, not two, and the measured asymmetry is artificial. This leads to a bias that will be more severe for low  $M$ . The issue of split trades will of course also be present for large  $M$ , but we would expect their effect to be much smaller, since with many split trades on average each investor will “benefit” the same number of times. Therefore, for the high  $M$  tests in Table 6 we expect the bias introduced by split trades to be of less importance. We address the issue of split trades in Appendix B and show that the results hold also when we only count multiple trades in the same direction within the time window once.

We also verify that trading before ones’ neighbors is related to centrality. Specifically, we define the vector  $w$ , where the  $i$ th element represents the average fraction of times investor  $i$  traded before his neighbors,  $w_i = \frac{1}{D_i - 1} \sum_j \frac{K_{ij}}{M_{ij}}$ , where  $j$  is summed over  $i$ ’s neighbors. We regress  $w$  on rescaled centrality,  $c - d$ , and verify that the two variables are positively related (t-stats of 8.2 for OLS, 4.7 for OLS with t-distributed errors, and 7.9 with iterated re-weighted robust regressions with Ramsey’s  $E$ -function, respectively).

To summarize, the nonrandom sequencing of trades support the information diffusion story of trading behavior over the style investing and liquidity provision stories.

### 4.3 Centrality and profits

The theory suggests that centrally placed investors, all else equal, are more profitable than peripheral investors. This is a novel prediction, and if it holds empirically, it lends substantial support to the information network story. Specifically, it is quite natural that the degree of an investor—being derived

from the investor’s trading behavior—is strongly related to other variables, e.g., number of trades, trading volume, and even trading profits, and it is therefore difficult to draw inferences from properties of the degree. Centrality, on the other hand, *a priori* has no such direct relation to other measures, or stories, of trading behavior—the natural interpretation is that it measures investor advantage from information diffusion. We note that centrality depends on the joint trading behavior of all investors in the market in a high-dimensional way. Our study is therefore quite different from studies that first aggregate investors into groups (e.g., institutional versus private) and then proceed by analyzing the performance of these groups.

We regress normalized profits,  $\mu$ , and normalized excess profits,  $\mu^e$ , on log-volume, number of trades, connectedness and centrality, using a 30-minute time window. To avoid influence by outliers, we truncate the data, so that investors in the bottom two percentiles and top two percentiles of connectedness are discarded. The results, using univariate regressions, are somewhat mixed as shown in Table 7, but generally support the view of a positive relation between centrality and profits. For example, the coefficients for centrality and degree are both positive for the normalized profits, suggesting that higher connectedness and centrality are associated with higher profits. Given the large sample size it is not surprising that the statistical significance of the tests are extremely high.

$\mu$						$\mu^e$					
	$c$	$d$	$c - d$	$n$	$v$		$c$	$d$	$c - d$	$n$	$v$
	0.0025					-0.00026					
		0.0023					-0.00053				
			0.016					0.012			
				0.0037					0.00069		
					0.0014					0.00014	
$t$	41	38	30	61	35	$t^e$	-2.9	-5.7	25	13.3	4.0
$\Delta\mu$	0.5%	0.4%	0.3%	0.7%	0.4%	$\Delta\mu^e$	-0.03%	-0.06%	0.3%	0.13%	0.04%

Table 7: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ), rescaled centrality ( $c - d$ ), log-trades ( $n$ ) and log-volume ( $v$ ). The  $\Delta t = 30$ -minutes window is used. The data is truncated, such that the investors in the bottom two percentiles and top two percentiles of connectedness are discarded from the data.

When excess profits are regressed, both the coefficients on centrality and degree are negative. However, the t-statistic for centrality is “only” -2.87, which may be viewed as “almost insignificant” given the size of our dataset. The coefficient of rescaled centrality ( $c - d$ ) is positive both for normalized profits and excess normalized profits, further supporting a positive relationship between centrality and profits. The bottom line in the table shows the economic significance of the results. Specifically,  $\Delta\mu$  and  $\Delta\mu^e$ , describe the predicted impact of a one standard deviation increase of each variable on normalized returns and normalized excess returns, respectively, varying between  $-0.06$  and  $0.7\%$ .

In order to identify the effect of centrality, we do multivariate regressions, where we control for trading volume, number of trades and degree. We have no reason to believe that error terms are normally distributed, so in addition to ordinary OLS, we perform an OLS regression that is robust to heavy-tailed error terms, and an iteratively re-weighted least square (using Ramsey’s E-function). The multivariate results are in line with the predictions, as shown in Table 8. The centrality coefficient comes out positive in all regressions and the results are economically significant. Specifically, a one standard deviation increase

in centrality, all else equal, implies an increase in profits per volume traded by 1.8%-2.8%, depending on the regression. Thus, an investor who traded for 10,000 TL in total during 2005 would realize profits that were 180-280 TL higher than an investor with a one standard deviation lower level of centrality, all else equal. The reverse holds for connectedness, where a one standard deviation increase leads to decreased profits per volume traded by 2.4%-3.6%.

$\mu$					$\mu^e$				
	$c$	$d$	$n$	$v$		$c$	$d$	$n$	$v$
$\beta_{OLS}$	0.0116	-0.0144	0.0080	-0.0017	$\beta_{OLS}^e$	0.0098	-0.0129	0.0038	-0.0004
$t_{OLS}$	> 20	< -20	> 20	< -20	$t_{OLS}^e$	> 20	< -20	> 20	-6.6
$\Delta\mu_{OLS}$	2.2%	-2.6%	1.5%	-0.47%	$\Delta\mu_{OLS}^e$	1.8%	-2.4%	0.71%	-0.12%
$\beta_{t-error}$	0.0148	-0.0182	0.0066	-0.0012	$\beta_{t-error}^e$	0.011	-0.0148	0.0035	-0.0004
$t_{t-error}$	13.7	-16.7	> 20	-8.4	$t_{t-error}^e$	11.8	-15.6	14.2	-3.46
$\Delta\mu_{t-error}$	2.8%	-3.4%	1.2%	-0.34%	$\Delta\mu_{t-error}^e$	2.1%	-2.8%	0.65%	-0.12%
$\beta_{Ramsey}$	0.0124	-0.0153	0.0080	-0.0017	$\beta_{Ramsey}^e$	0.0010	-0.0136	0.0038	-0.0004
$t_{Ramsey}$	> 20	< -20	> 20	< -20	$t_{Ramsey}^e$	> 20	< -20	> 20	-6.8
$\Delta\mu_{Ramsey}$	2.3%	-2.9%	1.5%	-0.47%	$\Delta\mu_{Ramsey}^e$	2.0%	-2.5%	0.72%	-0.12%

Table 8: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ) log-trades ( $n$ ) and log-volume ( $v$ ). The  $\Delta t = 30$ -minutes window is used. Coefficients are reported for OLS regression, as well as iteratively reweighed robust regressions with Ramsey’s  $E$ -function, and OLS with heavy-tailed ( $t$ -distributed) errors. The variables,  $\Delta\mu$  and  $\Delta\mu^e$  highlight the economic significance of the results by showing the change in normalized (and normalized excess) profits given a one standard deviation increase of the variable, all else equal. The data is truncated, such that the investors in the bottom two percentiles and top two percentiles of connectedness are discarded from the data.

Interestingly, the coefficients for number of trades,  $n$ , come out positive in all regressions. The economic significance is not as high as for centrality, but a one standard deviation increase in number of trades is associated with an increase of 0.65%-1.5% in profits per volume traded. Thus, measured over the whole investor universe, investors with more trades are more profitable per TL traded. The relationship is the opposite for trading volume, for which all coefficients are negative, although this is the economically least significant effect. This may suggest that sophisticated traders split their trades into multiple small trades, e.g., to avoid market impact, whereas less sophisticated investors do not. Again, all coefficients are strongly statistically significant.

The correlation between  $c$  and  $d$  is quite high (0.86 as seen in Table 3), and the coefficient for  $d$  is negative whereas the coefficient for  $c$  is positive, suggesting that it is the variation of centrality given connectedness that is important in determining profits. Our interpretation is that connectedness is associated with all types of trading, including, e.g., excess trading by noise traders, whereas centrality captures the information driven component of trading. Especially, centrality above and beyond degree is important. We measure this value of “excess” centrality by performing the same regressions as in Table 8, but regressed on  $c - d$ . The results, shown in Table 9, all come out with a positive sign, and strongly statistically significant, although the economic significance is lower for  $c - d$ . A one standard deviation increase in  $c - d$ , increases both normalized and excess profits ( $\mu$  and  $\mu^e$ ) by about 0.3%.

$\mu$				$\mu^e$			
	$c-d$	$n$	$v$		$c-d$	$n$	$v$
$\beta_{OLS}$	0.0126	0.0059	-0.0019	$\beta_{OLS}^e$	0.0109	0.0013	-0.0007
$t_{OLS}$	> 20	> 20	< -20	$t_{OLS}^e$	> 20	15.3	-10.6
$\Delta\mu_{OLS}$	0.27%	1.1%	-0.53%	$\Delta\mu_{OLS}^e$	0.23%	0.27%	-0.19%
$\beta_{t-error}$	0.0160	0.0039	-0.0014	$\beta_{t-error}^e$	0.0128	0.0006	-0.0006
$t_{t-error}$	14.9	18.0	-9.7	$t_{t-error}^e$	13.6	3.1	-5.1
$\Delta\mu_{t-error}$	0.36%	0.73%	-0.39%	$\Delta\mu_{t-error}^e$	0.27%	0.11%	-0.18%
$\beta_{Ramsey}$	0.0135	0.0058	-0.0019	$\beta_{Ramsey}^e$	0.00116	0.0013	-0.0007
$t_{Ramsey}$	> 20	> 20	< -20	$t_{Ramsey}^e$	> 20	14.2	-10.8
$\Delta\mu_{Ramsey}$	0.28%	1.1%	-0.53%	$\Delta\mu_{Ramsey}^e$	0.24%	0.25%	-0.19%

Table 9: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-rescaled centrality,  $c-d$ ,  $n$  and  $v$ . The  $\Delta t=30$ -minutes window is used. The data is truncated, such that investors in the bottom two percentiles and top two percentiles of connectedness are discarded from the data.

#### 4.4 Centrality and information

We verify that centrality is related to actual information events in that agents who trade earlier with respect to such events are more central than investors who trade later. These results provide further support for the information diffusion story over alternative explanations, e.g., liquidity provision and style investing.

Using public news outlets, we identify eleven events that can be related to large daily stock movements in 2005. We choose events for which it seems plausible that there may have been information diffusion among the population before the news made it into the news outlets, and specifically do not use earnings announcements since it could be argued that such pre-announcement date trading would be based on insider information.<sup>14</sup> The stock movements and information on related companies are given in Table 10. We have focused on medium-size companies with a couple of exceptions, namely EREGL and SAHOL, which are both quite large. The companies operate in fairly diverse business areas. All in all there were nine events that lead to positive returns and two that lead to negative returns. The information events, which were reported in public news outlets within five business days prior to and after stock movement dates, are described in Table 11. All events are stock specific, i.e., idiosyncratic for one specific firm.

We choose a time window of seven days before and after the day the event was mentioned for the first time in the news. Within this window we identify the investors who traded in the right direction (i.e., purchased the stock if the news led to positive returns and sold it if returns were negative). For each investor, we identify the time of the trade relative to the date of the information event (time 0). If an investor traded multiple times (over the time period of one event and/or across events), we use the (unweighted) average time traded for that investor. This leads to a vector,  $T$ , of trading times (in the range  $[-7, 7]$ ) for each investor who traded within the window around any of the events, in total 37,779 investors.

We regress  $T$  on the logarithms of centrality ( $c$ ), degree ( $d$ ), number of trades ( $n$ ) and volume ( $v$ ), where  $c$  and  $d$  were constructed using the 30-minute window. The results are shown in Table 12. We see that there is a strongly significant negative relation between  $c$  and  $T$ , i.e., that central investors trade

<sup>14</sup>Our view is that trading based on information diffusion is a much broader concept than trading on insider information.

<i>Ticker</i>	<i>Name</i>	<i>Main operation area</i>	<i>Movement date</i>	<i>Sign and magnitude</i>
BJKAS	Besiktas Futbol Yatirimlari	Soccer	7/26/2005	+19.4%
			12/26/2005	-33.5%
DEVA	Deva Holding	Pharmaceuticals	7/12/2005	+15.8%
DYOBY	DYO Boya	Paint and chemicals	7/14/2005	+12.25%
EREGL	Eregli Demir ve Celik Fabrikalari	Steel manufacturing	9/12/2005	+10.56%
SAHOL	Haci Omer Sabanci Holding	Multibusiness enterprise	10/4/2005	+6.99%
SEKFK	Seker Finansal Kiralama	Financial leasing	7/8/2005	-21.27%
SISE	Turkiye Sise ve Cam Fabrikalari	Glass manufacturing	7/29/2005	+7.08%
SKBNK	Sekerbank	Banking	7/12/2005	+14.55%
TEKST	Tekstilbank	Banking	8/16/2005	+12.27%
TNSAS	Tansas	Retail	8/19/2005	+12.12%

Table 10: Stocks and daily price movements used in the analysis of the relationship between centrality and information events.

earlier than peripheral investors. The result is robust to several variations. For example, very similar results are obtained — even somewhat stronger — if we move  $c$  to the left hand side of the regression and  $T$  on the right hand side, with the reversed causality interpretation that centrality is explained by investors trading early with respect to information events. It also holds in univariate regressions: The coefficient on  $c$  is negative with a t-statistic of  $-19.1$  in the OLS regression with  $c$  on the right hand side. Also, in the univariate regressions of  $T$  on  $c$ , the t-statistic in the OLS regression is  $-19.9$ , it is  $-9.3$  in the robust least squares regression, and  $-19.8$  in the reweighted iterated least square regression.

We also verify that the results are present when we use a larger set of events. We expand the number of events to 27, shown in Table 13, in total covering 67,509 investors. Some of the events (e.g., GIMA) in the expanded set were less clear-cut than in the original set, because the link between the reported news and the stock movement was (subjectively) somewhat ambiguous. For instance, it could be questioned whether the news on April 22, 2005, that GIMA (a national retail chain) was to open a branch in a mid-sized coastal town could have caused a sizable jump to its stock price, although such a jump was observed around this news event. The results for the expanded set, shown in Table 14, are still similar to those of the original set, although slightly weaker. For example the univariate OLS t-statistic is  $-16.6$ , and the multivariate OLS t-statistic is  $-8.3$ .

#### 4.5 Robustness of results

We verify that the results are robust to several variations and extensions. This also allows us to rule out several alternative explanations.

#### Out-of-sample tests

Since our dataset is restricted to only one year, we have constructed all variables using the full year of data. A potential weakness of this approach is that since profits are measured in-sample, there may potentially be a purely mechanical relationship between centrality and profits that is driving the results. To address this weakness, we divide the full year into two periods. We construct  $C$ ,  $D$ ,  $N$  and  $V$  during the first eight months, and then measure  $P$ ,  $P^e$ ,  $\mu$ , and  $\mu^e$  during the remaining four months of the year. The tests are then restricted to include only investors who traded in both sub-periods. There were 228,538 such investors.

The results, shown in Table 15 for the univariate regressions and in Table 16 for the multivariate regressions, are similar to the previous results, although the statistical significance is somewhat lower. We attribute this to the shorter time period in combination with smaller sample size. Specifically, all the centrality coefficients have the correct sign, and most of them are highly statistically significant. The main difference is that centrality coefficients for excess profits in the multivariate regression are less significant. The economic impact on profits of a one standard deviation increase in centrality varies between 0.2% and 1.8% per TL traded — which is similar to the previous tests (see Tables 7 and 8).

We also study the out-of-sample relationship between trading early on information events and centrality. In the base sample, with eleven events, all events were chosen to be in the second half of 2005, so we construct  $C$ ,  $D$ ,  $N$  and  $V$  using the first six months (and a 30-minute window). This reduces the number of traders in the sample to 32,375. The results are shown in Table 17. For the multivariate regressions, all centrality coefficients have the right sign (negative). For the robust least square regression, the coefficient is strongly significant (t-statistic -12.9), whereas the coefficients are insignificant for the OLS and reweighted least square regressions (about -1.4 for both). For the univariate regressions, all results are strongly negatively significant (t-statistics of -13.8, -9.5 and -13.7, respectively, for the three regressions). Thus, the results are also valid out-of-sample.

## Longer profit window

The argument in the Section 4.2 about sequencing of trades to rule out “style” investments as an explanation also applies to momentum “style” strategies. It is furthermore implausible that momentum plays a major role a driver behind the results in this study, given the one-month window used when defining trading profits. Momentum typically plays a role at longer horizons, see for example Jegadeesh and Titman (1993) and Jegadeesh and Titman (2001), who exclude one month returns in their construction of momentum portfolios to avoid market micro-structure issues — the results go in the wrong direction for such short horizons. Indeed, if momentum was a driving force behind our results, the results should be stronger when using a longer time horizon for profits. We face restrictions on the window length we can use, given that we only have one year of data, but we have verified that the results are not stronger when using a three-month profit window. For example, the regression of excess returns on log-rescaled centrality, using a three-month profit window, leads to a coefficient of  $\beta = 0.0069$  with a  $t$ -stat of 9.2, compared to  $\beta = 0.012$  with a  $t$ -stat of 25 when the 30-day window is used (see Table 7).

The 1-3 month profit windows also rule out market microstructure effects as a driver behind the results, since such effects play out over much shorter time horizons. The main concern would be that investors move prices when trading, regardless of whether the price impact is related to fundamental value or not. This is not an issue for trades that are closed within the time window, since our measure in this case takes market impact of both trades into account, but it may spuriously affect our results if a trader buys a stock, this moves the price for non-fundamental reasons, the trader then holds on to the stock, *and* the price does not revert to fundamentals within the profit window.

We view this as highly implausible: Since the vast majority of our traders are small retail investors, the market impact of their trades will be small. More importantly, the time horizon typically assumed for these microstructure effects — whether due to liquidity, adverse selection, or other effects — is typically way shorter than one month. The literature on trading volume and stock returns (e.g., Campbell et al. (1993) and Chan (1993)), use time windows from a few minutes up to a couple of days, and find virtually

no effect for longer horizons.<sup>15</sup> Even for the extreme case when there is a substantial synchronized demand shock of retail investors to a highly illiquid stock, no excess returns exist are observable over a 50-day horizon, as documented by Engelberg et al. (2011) in their study of the market’s reaction to Jim Cramer’s advice on the television show *Mad Money*. This is still far shorter than the 3-month profit window used in our robustness test.

### Longer time window

We have stressed that the time horizon for information diffusion that we are considering is less than a week. Such a short horizon rules out other explanations for the connection between centrality and profits, e.g., traditional style motives. With a one-week horizon for information diffusion, the time window should be chosen significantly shorter, since higher-order connections are also taken into account. So far, we have used time windows of up to 30 minutes.

To verify that our results are robust to using longer time windows, we also carry out the analysis using a time window,  $\Delta t$ , of one day. Because of computational limitations we are restricted to studying one third of the investors (about 193,000) and construct the EIN from three months of trades (instead of a year).<sup>16</sup> The results, shown in Table 18, again document a positive relationship between centrality and profits, with similar economic and statistical magnitudes as in our previous tests. Thus, the 30-minute time window adequately captures information diffusion even though it may be natural to think of the actual time period over which diffusion from one agent to another as somewhat longer. This also follows from our previous argument that even with a 30-minute window, agents who trade the same stock in the same direction on the same day have a higher order connection.

### Higher threshold for connectedness

We run the same tests as in Table 8 with a higher threshold for two investors to be identified as being connected,  $M = 10$ . The results, summarized in Table 19, are similar, although not as economically significant. With the higher threshold, a one standard deviation increase in centrality, all else equal, leads to an increase in expected profits per unit of TL traded of 0.10%, with a t-statistic of 6.89.

### Excluding institutional investors

To rule out that the results are driven by institutional investors we carry out the same tests excluding institutional investors, with virtually identical results (not reported). That is, we get very similar regression coefficients, statistical and economic significance when institutional investors are excluded as when they are included, showing that we are not capturing a difference between institutional and individual investors, but rather differences among individual investors. That the results are very similar is not surprising, given that each investor has the same weight in the regressions, and individual investors make up more than 99.9% of the investor population in our sample.

---

<sup>15</sup>The results have shown to be robust across markets and different types of stocks, see Safvenblad (2000), Berglund and Liljebloom (1998) and Huber (1997).

<sup>16</sup>Even so, the program took several days to run on UPPMAX, a high performance computer cluster, using multiple servers with up to 72GB of RAM memory.

## Conditioning on degree

To further verify that centrality above and beyond degree matters, we fix  $D$  within a limited range, and study how normalized profits are related to centrality, within this range. The results are shown in Table 20. We see that centrality is still an economically and statistically significant factor in all tests. The economic significance of a one standard deviation increase in centrality is somewhat lower than in the full test, varying between 0.3%-1.1%. This is mainly because the variation—and thereby the standard deviation—of  $c$  is much lower for any given value of  $d$ , than the unconditional standard deviation. In fact, the coefficients on  $c$  are higher than the unconditional coefficients given in Table 8. Also, the statistical significance is of course lower given the much smaller sample size for the conditional test. We note that the coefficient for  $d$  is statistically insignificant in several of the tests, which is not surprising given that the  $d$ 's are specifically chosen to have low variation. Again, the coefficient for number of trades,  $n$ , come out positive in all tests, and is especially strong for the range  $2,000 \leq D \leq 2,200$ , where a one standard deviation in log-number of trades implies a 0.75-1.6% increase in normalized trading profits.

## No neighbors in the same brokerage house

If brokers are trading on behalf of their clients, systematic sequencing of trades could mean that the brokers are prioritizing important clients rather than that the clients themselves are trading sequentially. To address this concern, we prohibit links between investors who are associated to the same brokerage house when creating the EIN. The results, shown in Table 21, are almost identical as before. Thus, such broker sequencing of trades, although potentially present, does not seem to be the mechanism behind our results.

## 5 Concluding Remarks

Our study supports a view of the stock market as a place where information is incorporated into asset prices through gradual decentralized diffusion. Information networks provide an intermediate information channel, in between the public arena where news events and prices themselves make some information available to all investors, and the completely local arena of private signals and inside information. In an information network, the degree of publicness of a signal is determined by how long it has been diffusing, in which part of the network it initially entered, and by the network's topological properties. It should not come as a surprise then that large asset price movements occur independently of public information events, that investors take on diverse portfolio positions, and that they trade extensively.

Our centrality measure, used to capture the informational advantage of an investor, is a high-dimensional object in that each investor's centrality depends on the behavior of all other investors in the market in a complex way. Our approach is therefore differentiated from other studies of investor trading behavior in the market. Centrality is one measure that takes higher order connectedness into account, but there are also other measures, e.g., second order degree. A promising topic for future research is to further understand which measures are most important for asset pricing in an information network.

Given our data, we have little to say about the true channels of information diffusion. A simplistic interpretation is that social networks lead to diffusion when individuals communicate information through their social network of friends and acquaintances. This is the view taken in several recent studies that focus on certain investor groups, e.g., in Hong et al. (2004) and Cohen et al. (2008). But information networks may also represent other channels of information diffusion, e.g., when one person overhears

a discussion about stocks in the lunch canteen, when several people read the same local newspaper, when information diffuses through Internet discussion boards, etc. Which factors determine a market's information network? Geography? Social networks? Local media? Other channels? Given datasets with more detailed information about investors in the market, further research may shed light on this important question.

## A Degree adjusted measure of the expected number of overlaps (Section 4.1)

Obviously the EINs are not completely random; if they were, the degree distributions would be Poisson distributed. However, our detailed analysis of the degree distribution in Appendix C shows that the true distribution has heavier tails. A better specified test for stability is therefore to study the number of overlaps, given the (heavy-tailed) degree distributions observed in practice.

We proceed as follows: We take the degree distribution from the EIN generated under the first six months as given, representing the true data generation process, and then compare the actual and predicted number of overlaps, given this data generating process. Specifically, given that the degree distribution of the EIN is  $D$ , and that links in the second EIN are formed such that the probability that a link added to the second network involves investor  $i$  is proportional to investor  $i$ 's degree in the first EIN (i.e., the probability is  $\frac{D_i-1}{k_1-N}$ ), the probability of an overlap, if the second EIN contains only one link, is

$$\begin{aligned} \mathbb{P}(\text{Overlap}) &= \frac{1}{2} \sum_{i=1}^N \mathbb{P}(\text{Investor } i \text{ is linked}) \times \mathbb{P}(\text{Investor } j \text{ is linked with } i | \text{Investor } i \text{ is linked}) \\ &= \frac{1}{2} \sum_{i=1}^N \frac{D_i - 1}{k_1 - N} \times \frac{D_i - 1}{N}. \end{aligned}$$

Here, the factor  $\frac{1}{2}$  arises to avoid double counting links (which are bidirectional). Finally, given that  $k_2 \ll K$ , the expected number of overlaps, when  $k_2$  links are chosen in the second EIN, is approximately  $k_2 \times \mathbb{P}(\text{Overlap})$ , leading to the following approximate formula for the expected numbers of overlaps

$$E_{\text{Degree adjusted}}[y] \approx \frac{k_2}{2k_1N} \sum_{i=1}^N (D_i - 1)^2. \quad (9)$$

We compare the expected and true number of overlaps for the 1-minute and 5-minute windows, under this assumption of a *degree adjusted* network generating process. The over-representation of overlaps, although not as high as under the assumption of a completely random data generation process, is substantial: For the 1-minute window, overlaps are 7.55 times as frequent as what would be expected under null hypothesis of a random degree adjusted network generating process. For the 5-minute window, the factor is 6.09. This is for the standard definition of degrees, with the connection threshold ( $M$ ) set to one, in which case the comparable numbers for the completely random data generating process were 113.2 and 72.2, respectively.<sup>17</sup>

We note in passing that the statistical significance of this over-representation is huge, which is why we do not provide significance levels. For example, under the first hypothesis of a completely random data generation process with a 1-minute window, the expected number of overlaps is 104,750. The standard deviation is 325, so rejection at the 5% significant level would be reached if there were 105,385 overlaps, whereas the true number of overlaps is about 11.9 million. The EINs are thus relatively stable over time, which is consistent with the intuition that they are proxies of information networks.

## B Robustness of results in Section 4.2 to split trades

To ensure that the results are robust to accounting for split trades, we create an adjusted measure that rules out splitting by “muting” subsequent trades within a time window that may lead to such bias. Specifically, we only count multiple trades within the time window (of 30-minutes) once. For example, consider a case where investor  $i$  purchased a stock at  $t$ ,  $t + 5$  minutes,  $t + 12$  minutes, and  $t + 18$  minutes, and investor  $j$  purchased the same stock at  $t + 16$  minutes. Assume further that this was the only time that investor  $i$ 's and  $j$ 's trades overlapped. Then, with the original measure,  $K_{ij} = 3$ ,  $K_{ji} = 1$  and  $M_{ij} = 4$ , representing the three times  $i$  traded before  $j$  and the one time that  $j$  traded before  $i$ . With the new measure, only one trade between  $t$  and  $t + 30$  minutes counts toward  $K_{ij}$  ( $i$ 's trade at  $t$  together with  $j$ 's trade at  $t + 16$  minutes) which leads to  $K_{ij} = 1$ . Moreover,  $j$ 's trade at  $t + 16$  minutes together with  $i$ 's trade at  $t + 18$  minutes leads to  $K_{ji} = 1$  (and  $M_{ij} = K_{ij} + K_{ji} = 2$ ) since  $i$ 's trades are only “muted” for  $K_{ij}$  trades. Clearly, this muted definition of  $K_{ij}$ ,  $K_{ji}$  and  $M_{ij}$  accounts for any bias that may be

<sup>17</sup>We do not have the degree distributions for higher connection thresholds and can therefore not calculate the expected number of overlaps when a higher connection threshold is used, under the hypothesis of degree adjusted random network generation processes.

introduced by “split” trades. In fact, it will introduce a severe bias in the opposite direction, since it will tend to even out  $K_{ij}$  and  $K_{ji}$  even when trades were not split. In other words, the bias will lead to too many observations where  $\eta^M$  is close to zero with the new definition.

We verify this bias for low  $M$ . For example, when studying  $\eta^2$  with the new definition, we have that  $\eta^2 = 1$  (corresponding to  $K_{ij} = K_{ji} = 1$ ) for 80% of the observations, whereas  $\eta^2 = 0$  (corresponding to  $K_{ij} = 2$  and  $K_{ji} = 0$ , or  $K_{ij} = 0$  and  $K_{ji} = 2$ ) for only 20%, compared with 50% each under the null hypothesis that  $p = 1/2$ . Thus, the new measure overcorrects for split trades.

To partly mitigate the effects of this overcorrection, we study the distribution of  $\eta^M$  away from 0. Specifically, we define the random variable

$$\eta_z^M = \begin{cases} \eta^M, & z \leq \eta^M \leq M, \\ 0, & 0 \leq \eta^M < z, \end{cases} \quad (10)$$

which truncates  $\eta^M$  to zero for values less than  $z$ . We then study the mean of our empirically constructed  $\eta_z^M$ ,  $\bar{\eta}_z^M$ , versus its expected value

$$E[\eta_z^M | p] = \sum_{k=z}^M \binom{M}{k} |2k - M| p^k (1-p)^k, \quad (11)$$

and test whether  $\bar{\eta}_z^M$  is significantly different from  $E[\eta_z^M | p = 1/2]$ .

Thus, to summarize, the adjusted measure mutes trades when calculating  $K_{ij}$ ,  $K_{ji}$  and  $M_{ij}$ , which leads to a different empirical distribution of  $\eta^M$ . The adjusted measure controls for split trades, but introduces a downward bias of  $\eta^M$ , through overbalancing  $K_{ij}$  and  $K_{ji}$ . This, in turns, leads us to study  $\eta_z^M$  for some  $z > 0$ —excluding these over-balanced outcomes where  $K_{ij} \approx K_{ji}$ . It is important to note that by choosing  $z > 0$ , we do *not* introduce another bias, since we adjust the expectation accordingly in (11). Rather, a choice of  $z > 0$  focuses our tests on traders who are severely unbalanced in their relationship, in that one trader went before the other many times. In Table 22 we show the tests with the new measure, for  $M$  from 30 to 210, using  $z = \frac{M}{2}$  and  $z = 20$  as cutoff points.<sup>18</sup>

In both cases, we strongly reject that  $p = \frac{1}{2}$ , although the GMM-estimates of  $\hat{p}$  is closer to  $\frac{1}{2}$  in both cases. In the first case, in which  $z = \frac{M}{2}$  is chosen,  $\hat{p}$  lies between 0.53 – 0.71 as  $M$  varies, and in the second case  $\hat{p}$  lies between 0.53 – 0.55. We do not know how much of the decrease in  $\hat{p}$  is due to mitigation of the upward bias of split trades and how much is due to the new downward bias of overcorrection. In any case, the new estimate of  $\hat{p}$  provides a lower bound on the true probabilities and, importantly, it is still higher than  $\frac{1}{2}$ .

## C Investor distributions

The distributions of  $C$ ,  $D$ ,  $N$ ,  $P$  and  $V$  are intimately related to asset pricing dynamics in that they are informative about the mechanisms that determine stock prices, and may therefore be informative about the aggregate behavior of stock markets. The ultimate goal would be to have a unified model that explains the behavior of stock markets, both in the cross section, across individual investors, and in the aggregate time series.

In an important study, Gabaix et al. (2003) provide a motivation for observed aggregate stock market dynamics. According to their arguments, Pareto distributed aggregate market returns, number of trades and trading volumes can be explained by Pareto distributed sizes of investor capital, together with additional assumptions about the mechanisms through which investors trade. Here, a vector whose elements are drawn from a distribution that satisfies  $\mathbb{P}(X \geq x) \sim x^{-\alpha}$  in the tail of the distribution is said to be *power-law* or *Pareto* distributed, with *tail exponent*  $\alpha$ .

We provide a brief summary of the arguments made in Gabaix et al. (2003). Their main result is that, (i) aggregate trading volume (both trade-by-trade, and per unit time) is Pareto distributed with tail-exponent  $\zeta_V^{AG} = 1.5$ , (ii) the distribution of aggregate number of trades per unit time is also Pareto distributed with tail exponent  $\zeta_N^{AG} = 3$ , as is (iii) the aggregate return distribution per unit time,  $\zeta_r^{AG} = 3$ . Here, the superscript *AG* represents aggregate measures over time whereas we will use *CS* for cross sectional relationships across investors. The authors verify these distributional properties empirically, and also provide an economic motivation for why they arise.

The underlying economic assumptions leading to these results are the following: For (i), it is assumed that (a) investor size is Zipf-law distributed, i.e., it is Pareto-law distributed with a tail exponent of one. Thus, given a size vector,  $S$ , where  $S_i$  is the capital of investor  $i$ ,  $S$  is power-law distributed in the tail with tail exponent  $\zeta_S = 1$ . This assumption can be empirically and theoretically motivated<sup>19</sup>; (b) The price impact of a trade on the stock’s price

<sup>18</sup>Other choices of  $z$  were also tested. The rejection of  $p = \frac{1}{2}$  holds for general combinations of  $z > 15$  and  $M > 25$ , although the estimated  $\hat{p}$  varies with the choices of  $z$  and  $M$ .

<sup>19</sup>In Gabaix (1999), a theoretical argument is made for why Zipf-law distributed sizes arise in various contexts.

is proportional to the square-root of the size of the trade, and inversely proportional to the time a trader waits to trade,  $\Delta Q_z = k \frac{\sqrt{V_z}}{T}$ . This assumption is related to the random arrival of participants in a market over time; (c) The typical size of an investor's trade is related to the investor's size according to  $V_{iz} \propto S_i^\delta$  for some constant  $\delta > 0$ . This is a technical assumption; (d) The aggregate losses that an investor makes from market impact of trades is proportional to the investor's size. This is a "survival constraint." The authors argue that if the cost for an investor is higher than what is predicted by (d), the investor will rescale to an efficient size, so investors with inefficient sizes will not exist. Therefore, the only possible outcome is that all investors must have the same cost per unit size. All these conditions are assumed to hold in the tail of the distributions, i.e., above some cutoff size  $\bar{S}$ . Moreover, they only need to hold approximately, i.e., up to higher order errors. The authors then show that assumptions (a)-(d) lead to the following tail-exponents:

$$\zeta_V^{AG} = 1.5, \quad (12)$$

which together with (b) leads to

$$\zeta_r^{AG} = 2\zeta_V^{AG} = 3. \quad (13)$$

Finally, an additional assumption (e) is made: that prices adjust linearly over time, so that if a stock is traded at  $Q^t$ , but an investor believes that the correct price is  $Q^t + M$ , then the price at time  $t'$  is expected to be  $Q^t + k(t' - t)$ , for some constant  $k$ , for  $t' \leq t + \frac{M}{k}$ . Assumption (e) implies that an investor will split his trades into "chunks" of size proportional to  $\sqrt{V}$ , which in turn implies that

$$\zeta_N^{AG} = 2\zeta_V^{AG} = 3. \quad (14)$$

Now, the previous assumptions, via the Zipf law size distribution, also have several cross sectional implications for investors' trading behavior, and potentially allow for a joint explanation of trading behavior in the cross section and in the time series. For example, it follows from (c) and (d) that the number of trades of investor  $i$ ,  $N_i$ , will be proportional to  $S_i^{1-3\delta/2}$ , so  $n_i \sim (1 - 3\delta/2) \log(S_i)$ . Further, since  $V_i = N_i \bar{V}_{iz} = N_i S_i^\delta$ , where  $\bar{V}_{iz}$  is the average size of a trade by investor  $i$ , it follows that  $\log(\bar{V}_{iz}) = v_i - n_i = \delta \log(S_i)$ . Therefore, depending on whether  $\delta$  is less than or greater than  $2/3$ , the cross sectional covariance between  $n$  and  $v - n$  will be positive or negative.

In our data,  $\text{corr}(v - n, n) = 0.29$ , when calculated for all investors, suggesting that  $\delta < 2/3$ . Similar positive correlations are obtained when only tail observations are included. For example, if only the 30,000 investors with the most number of trades are used, the correlation is 0.26. This is in line with Gabaix et al. (2003)'s results, given that  $\delta < 2/3$  for the Turkish stock market.

The relationship between investor size and trading behavior, together with the Zipf law for size distributions, also has further implications for the distributions across investors. Specifically, given that investor size is Zipf law distributed,  $\zeta_S = 1$ , and  $N_i \sim S_i^{1-3\delta/2}$ , it follows that the cross section of number of trades,  $N$ , is power law distributed with tail exponent

$$\zeta_N^{CS} = \frac{1}{1 - 3\delta/2} \quad (15)$$

Further, given that  $\bar{V}_{iz} \sim S_i^\delta$ , it follows that the cross section of aggregate trading volumes,  $V = \bar{V}_{iz} N_i \sim S_i^{1-\delta/2}$ , is of Pareto-type with tail exponent

$$\zeta_V^{CS} = \frac{1}{1 - \delta/2}. \quad (16)$$

Gabaix et al. (2003) do not take a stand on the underlying driver behind investors' trades. Within our framework, information diffusion is the underlying driver which, as discussed in Ozsoylev and Walden (2011), implies that centrally placed investors are better informed, trade more, and make higher profits. Within the specific framework of Ozsoylev and Walden (2011), profits and total volume are asymptotically linear in the information precision of each agent, here proxied by  $C$ . Thus, within that framework, cross sectional profits and trading volume should have the same tail distributions,  $\zeta_P^{CS} = \zeta_V^{CS}$ . Moreover, as long as  $C$  is a good proxy for an investor's information advantage,  $C$  should also be power-law distributed with tail exponent

$$\zeta_C^{CS} = \zeta_P^{CS} = \zeta_V^{CS}. \quad (17)$$

Of course,  $C$  is just a proxy and, the linearity of these relations depend on strong assumptions made in Ozsoylev and Walden (2011), so although we expect a positive relationship between these variables we should not be surprised to see deviations from the left equality in (17).

Are these variables Pareto distributed in practice and, if so, do they satisfy the relationships (15-17)? In

Figure 4, we show the tails of the empirical cumulative distribution functions, plotted in log-log coordinates for  $C$ ,  $D$ ,  $N$ ,  $P$  and  $V$ . In log-log coordinates, a Pareto distribution will show up as a straight line, and the slope is (minus) the tail exponent. A thinner-tailed distribution, e.g., log-normal or exponential, will show up as a concave function. We see that profits ( $P$ ) and trading volume ( $V$ ) seem to be fairly well approximated by a straight line and that the slopes seem to be similar, providing some support for (17). Centrality ( $C$ ) and degree ( $D$ ) on the other hand seem less well approximated by a straight line. The number of trades ( $N$ ) lies somewhere between, more concave than profits and volume, but less concave than centrality and degree.

We estimate how consistent the observed distributions are with Pareto-law distributed tails. We vary the tail distribution,  $\zeta$  and the cut-off point in the tail,  $N_{cutoff}$ , beyond which the a power law is assumed to hold. We use the Kolmogorov-Smirnov (KS) error, which measures the maximum difference between the true and empirical cumulative distribution functions, and choose the estimated tail exponent,  $\zeta$ , and cutoff,  $N_{cutoff}$ , to minimize the KS error.

The results are given in Table 23, in which the estimated tail exponent is also shown for various exogenously given values of the cutoff. For  $P$  and  $V$ , the errors are small (0.0074 and 0.0080 respectively) over a substantial range of investors (17,090 and 28,020 respectively). Moreover, the two estimated tail exponents of  $P$  and  $V$  are similar, in line with (17). We note that these distributions are very heavy-tailed, even heavier-tailed than the Zipf law (0.86 and 0.80 respectively). For  $N$ , the error is larger (0.0170) and the optimal tail cutoff point is lower, 6,140. For  $C$ , and  $D$ , the fit is even worse. We note that the large ranges over which the power law distribution provides a good approximation for profits and trading volumes rule out the explanation that these results arise because of a few, high-frequency trading, liquidity providers.

To understand whether these results are consistent with power law distributions, we run a goodness of fit test. We simulate draws from a true Pareto distribution and compare the Kolmogorov-Smirnov errors from these simulations with the ones obtained empirically. The results support our intuition that profits and volume are consistent with power law distributions, whereas centrality and degree are not. Further, it suggests that the distribution of number of trades is consistent with a power law. The test cannot reject that either  $N$ ,  $P$  or  $V$  are Pareto law distributed at the 5% significance level, whereas both  $C$  and  $D$  are rejected at the 1% significance level.

To compare the approximations by power laws with alternative distributions, we also estimate the tails with exponential distributions and log-normal distributions, confirming our previous results, as seen in Table 24. The power law is dominating for  $P$ , which outperforms the exponential and log-normal approximations regardless of which cutoff point is chosen. The results are almost as strong for  $V$  and  $N$ , which both have statistically significantly higher likelihood ratios over exponential and log-normal distributions for cutoffs above 1,500 investors.

For centrality,  $C$ , the results are very different. The exponential distribution provides the best fit for all but very small cutoffs, for which a log-normal distribution provides a better fit. The Pareto distribution is dominated, with high statistical significance, for all cutoff values. Similar results obtain for degree,  $D$ . In Figure 5, we show the fitted values for  $D$ ,  $N$ ,  $P$ , and  $V$  (the approximation for  $C$  is not shown since its is very similar to that of  $D$ ).

Thus,  $P$  and  $V$  and  $N$  are well approximated in the tail by power laws, and  $\zeta_P^{CS} \approx \zeta_V^{CS}$ , so (15), as well as the right equality of (17), are supported by the data. The left equality of (17), the relationship between  $\zeta_C^{CS}$  and  $\zeta_P^{CS}$ , is not supported though, since  $C$  is clearly not power law distributed in the tail. As mentioned, however, this could simply imply that the relationship between  $C$  and  $P$  is not linear.

The analysis in Gabaix et al. (2003) also predicts (16), but clearly this relationship is not consistent with our empirical results, since the estimated tail exponent  $\zeta_V^{CS}$  is less than one (it is 0.80), implying that  $\delta$  is negative. This is the case over a large range of tail cutoff points, as can be seen from Table 23 (above  $N_{cutoff}^* \approx 1,000$  the estimated tail exponent  $\zeta_V^{CS} \leq 1$ ). A negative  $\delta$  would imply that larger investors take on smaller trades, which does not seem very plausible, but strictly speaking is not impossible. More importantly though, if  $\delta < 0$ , the analysis in Gabaix et al. (2003) leading to  $\zeta_V^{AG} = 1.5$  breaks down, since it is based on the assumption that it is the largest investors who carry out the largest trades.<sup>20</sup>

Thus, there seems to be a piece missing in our understanding of equations (12-17), raising an interesting challenge for further research. Specifically, a model that reconciles the very heavy-tailed cross sectional distributions of profits and volume and the somewhat thinner-tailed distributions of number of trades with the well known aggregate distributions of trading volume and number of trades over time, would potentially lead to new insights about what determines the dynamics of prices and trading in the stock market.

<sup>20</sup>Formally, their equation (15) is not well defined when  $\delta < 0$ , since the integral is infinite.

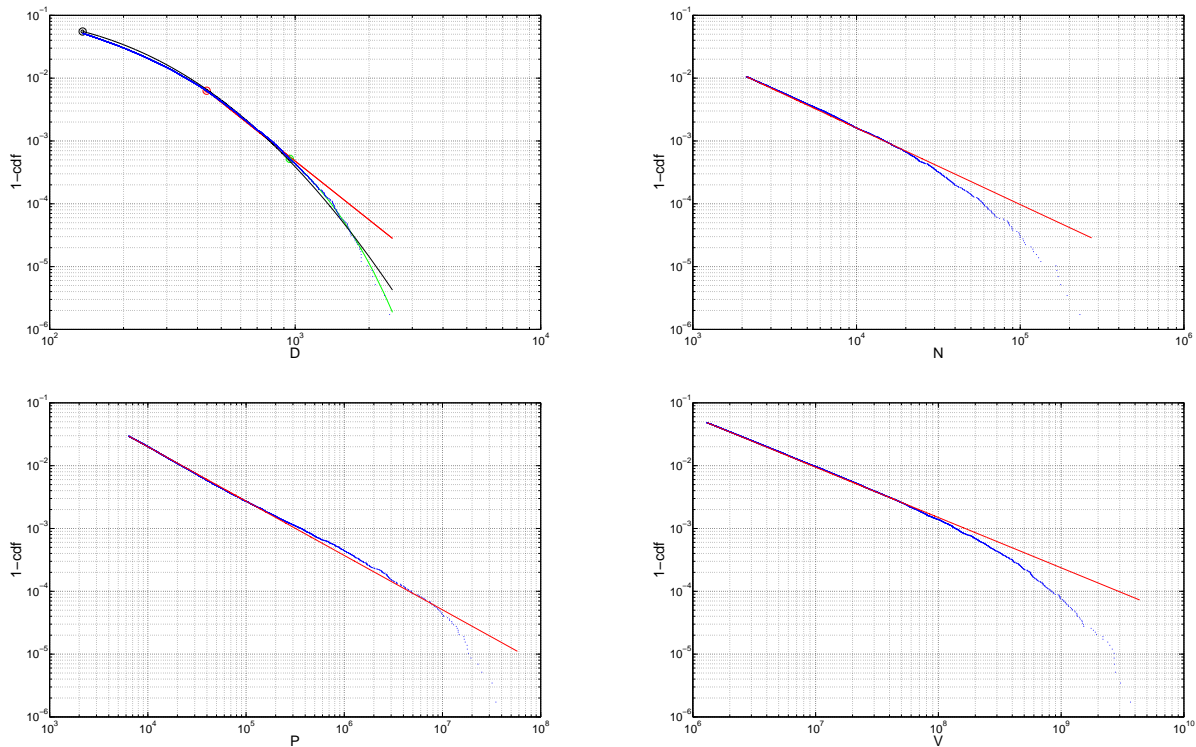


Figure 5: Estimated tail distribution of the cross section across investors of connectedness,  $D$  (upper left), number of trades,  $N$  (upper right), profits,  $P$  (lower left), and volume,  $V$  (lower right). i) Connectedness is not well approximated by single distribution: For  $N_{cutoff} = 30,000$ , one-sided log-normal distribution with  $\mu = 9.58$ ,  $\sigma^2 = 0.52$  is optimal. For  $N_{cutoff} = 3,000$ , Pareto law with tail exponent  $\alpha = 3.11$  is optimal. For  $N_{cutoff} = 300$ , exponential distribution, with  $\lambda = 3.65 \times 10^{-5}$  is optimal. ii) Number of trades is well approximated by Pareto law with  $\alpha = 1.22$  and  $N_{cutoff} = 6,140$ . iii) Profits are well approximated in tail by Pareto law with tail exponent,  $\alpha = 0.86$ ,  $N_{cutoff} = 17,090$ . iv) Volume is well approximated in tail by Pareto law with tail exponent,  $\alpha = 0.80$ ,  $N_{cutoff} = 28,020$ .

## References

- Adamic, L., C. Brunetti, J. Harris, and A. Krilenko, 2010, Trading networks, Working paper, University of Michigan.
- Barber, B., Y. T. Lee, Y. J. Liu, and T. Odean, 2009, Just how much do individual investors lose by trading?, *Review of Financial Studies* 22, 609–632.
- Barberis, N., and A. Shleifer, 2003, Style investing, *Journal of Financial Economics* 68, 161–199.
- Berglund, T., and E. Liljeblom, 1998, Market serial correlation on a small security market: A note, *Journal of Finance* 7, 1265–1274.
- Betermeier, S., T. Jansson, C. Parlour, and J. Walden, 2011, Hedging labor income risk, *Journal of Financial Economics* forthcoming.
- Bodie, Z., R. Merton, and W. Samuelson, 1992, Labor supply flexibility and portfolio choice in a life cycle model, *Journal of Economic Dynamics and Control* 16, 427–449.
- Brown, S. J., and W. N. Goetzmann, 1997, Mutual fund styles, *Journal of Financial Economics* 43, 373–399.

- Campbell, J., S. Grossman, and J. Wang, 1993, Trading volume and serial correlation in stock returns, *Quarterly Journal of Economics* 108, 905–939.
- Chan, K., 1993, Imperfect information and cross-autocorrelation among stock prices, *Journal of Finance* 48, 1211–1230.
- Chung, F., and L. Li, 2006, *Complex Graphs and Networks* (American Mathematical Society).
- Cohen, L., A. Frazzini, and C. Malloy, 2008, The small world of investing: Board connections and mutual fund returns, *Journal of Political Economy* 116, 951–979.
- Colla, P., and A. Mele, 2010, Information linkages and correlated trading, *Review of Financial Studies* 23, 203–246.
- Cutler, D. M., J. M. Poterba, and L. H. Summers, 1989, What moves stock prices, *Journal of Portfolio Management* 15, 4–12.
- Das, S. J., and J. Sisk, 2005, Financial communities, *Journal of Portfolio Management* 31, 112–123.
- Duffie, D., S. Malamud, and G. Manso, 2009, Information percolation with equilibrium search dynamics, *Econometrica* 77, 1513–1574.
- Engelberg, J., C. Sasseville, and J. Williams, 2011, Market madness: The case of mad money, *Management Science*, forthcoming.
- Fair, R. C., 2002, Events that shook the market, *Journal of Business* 75, 713–731.
- Fedyk, Y., C. Heyderdahl-Larsen, and J. Walden, 2010, Market selection in a multi-asset economy, Working paper, UC Berkeley.
- Feng, L., and M. Seasholes, 2004, Correlated trading and location, *Journal of Finance* 59, 2117–2144.
- Fracassi, C., 2009, Corporate finance policies and social networks, Working Paper, UCLA.
- Gabaix, X., 1999, Zipf’s law for cities: An explanation, *Quarterly Journal of Economics* 114, 739–767.
- Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley, 2003, A theory of power-law distributions in financial market fluctuations, *Nature* 423, 267–270.
- Grossman, S., and M. Miller, 1988, Liquidity and market structure, *Liquidity and Market Structure* 43, 617–637.
- Grossman, S., and J. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.
- Han, B., and L. Yang, 2011, Social networks, information acquisition, and asset prices, Working paper, University of Texas at Austin.
- Hellwig, M. F., 1980, On the aggregation of information in competitive markets, *Journal of Economic Theory* 22, 477–498.
- Hong, H., J. D. Kubik, and J. C. Stein, 2004, Social interaction and stock-market participation, *Journal of Finance* 49, 137–163.
- Huber, P., 1997, Stock market returns in thin markets: Evidence from the Vienna Stock Exchange, *Applied Financial Economics* 7, 493–498.
- Ivković, Z., and S. Weisbenner, 2007, Information diffusion effects in individual investors’ common stock purchases: Covet thy neighbors’ investment choices, *Review of Financial Studies* 20, 1327–1357.
- Jegadeesh, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.
- Jegadeesh, N., and S. Titman, 2001, Profitability of momentum strategies: An evaluation of alternative explanations, *Journal of Finance* 56, 699–720.

- Kyle, A. S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1336.
- Massa, M., and A. Simonov, 2006, Hedging, familiarity, and portfolio choice, *The Review of Financial Studies* 19, 633–685.
- Mayers, D., 1973, Nonmarketable assets and the determination of capital asset prices in the absence of a riskless asset, *The Journal of Business* 46, 258–267.
- Odean, T., 1999, Do investors trade too much?, *American Economic Review* 89, 1279–1298.
- Ozsoylev, H., and J. Walden, 2011, Asset pricing in large information networks, *Journal of Economic Theory* forthcoming.
- Pareek, A., 2009, Information networks: Implications for mutual fund trading behavior and stock returns, Working Paper, Yale University.
- Parlour, C., and J. Walden, 2011, General equilibrium returns to human capital and investment capital under moral hazard, *Review of Economic Studies* 78, 394–428.
- Safvenblad, P., 2000, Trading volume and autocorrelation: Empirical evidence from the Stockholm Stock Exchange, *Journal of Banking and Finance* 24, 1275–1287.
- Shiller, R. J., and J. Pound, 1989, Survey evidence on the diffusion of interest and information among investors, *Journal of Economic Behavior* 12, 47–66.
- Shive, S., 2010, An epidemic model of investor behavior, *Journal of Financial and Quantitative Analysis* 45, 169–198.
- Tetlock, P. C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 211–221.
- Tetlock, P. C., 2011, Does public financial news resolve asymmetric information, Working paper, Columbia University.

<i>Ticker</i>	<i>Movement date</i>	<i>Information event(s)</i>	<i>News source(s) and date(s)</i>
BJKAS	7/26/2005	The company is close to a deal regarding the transfer of Brazilian soccer player Ailton Goncalves da Silva playing in Schalke 04 in the German league.	Reported in the Istanbul Stock Exchange Bulletin on 7/26/2005.
	12/26/2005	There were negative news and comments regarding the company's debts in some of the daily newspapers on 12/26/2005. The company issued a statement denying these news on that day. Also, on the same day, the company announced that no profit had been generated in the previous fiscal year and therefore no dividends would be paid.	Reported in the Istanbul Stock Exchange Bulletin on 12/26/2005.
DEVA	7/12/2005	The number of Deva Holding shares in the portfolios of the emerging market funds managed by GEM Global Equities Management S.A. reached to an amount corresponding to 5.27% of the company capital.	Reported in the Istanbul Stock Exchange Bulletin on 7/13/2005.
DYOBY	7/14/2005	The company announced that it plans to construct new facilities in the Gebze Organized Industrial Zone to satisfy the increasing demand in the domestic market.	Reported in the magazine Ekonomist on 7/14/2005.
		Ahmet Yigitbasi, a senior executive of the company, told that they aim to achieve 20% revenue growth in 2005.	Reported in the magazine Tek Borsa on 7/11/2005.
EREGL	9/12/2005	In a press meeting, the management announced that the company's profit increased by 34.4% in the first half of the year compared to the same period in the previous year. The general manager of the company, Fadil Demirel, told that the valuations mentioned in the media in relation to the forthcoming privatization of the company were too low given the company's financial performance and production capacity.	Reported in the daily newspapers Aksam, Milliyet, Radikal, Sabah on 9/14/2005 and 9/15/2005.
SAHOL	10/4/2005	The company executives met Toyota officials in a four-day-long visit to Japan. Both companies declared their intention to strengthen their business partnership.	Reported in the daily newspaper Aksam on 10/7/2005.
SEKFK	7/8/2005	Rabobank, one of the leading Dutch retail banks, signed a share purchasing agreement on July 7th with Sekerbank Personel Munzam Sosyal Guvenlik and Yardimlasma Vakfi, a major shareholder of the company. The transaction price was 43% lower than the market price.	Reported in the magazine Ekonomist on 7/11/2005.
SISE	7/29/2005	The company put the world's biggest glassware oven into service in Bulgaria. The company is accelerating its investments in order to realize its vision to become the world leader in glass production.	Reported in the daily newspapers Hurriyet and Vatan on 8/1/2005.
SKBNK	7/12/2005	The negotiations regarding the sale of Sekerbank shares from Sekerbank Personel Munzam Sosyal Guvenlik and Yardimlasma Vakfi to Rabobank were completed and a share purchasing agreement was signed on 7/7/2005. Rabobank will make calls to other shareholders of Sekerbank following the realization of the share transfer. The parties agreed that, at the end of the call period, Rabobank will be the majority shareholder of Sekerbank.	Reported in the Istanbul Stock Exchange Bulletin on 7/13/2005.
TEKST	8/16/2005	In an announcement made by Fitch Ratings, Tekstilbank's good credit quality and strong risk management system were emphasized.	Reported in the Istanbul Stock Exchange Bulletin on 8/15/2005.
		It is rumored that the Hungarian OTP Bank is interested in acquiring Tekstilbank.	Reported in the magazine Ekonomist on 8/15/2005.
TNSAS	8/19/2005	Koc Holding is interested in acquiring Tansas and signed an exclusivity agreement with Dogus Holding, the current majority shareholder of the company. Koc aims to be the largest retailer in the country with this acquisition.	Reported in the Istanbul Stock Exchange Bulletin and the daily newspapers Radikal, Vatan between 8/17/2005 and 8/20/2005.
		The company's first half-year profit is 1.9 million YTL. The company made a loss of 30.2 million YTL in the first half of the previous year. Net sales of the company increased by 19%.	Reported in the magazine Para on 8/19/2005.

Table 11: Information events reported in public news outlets within 7 business days before and after the stock movement dates.

$T$	$c$	$d$	$n$	$v$
$\beta_{OLS}$	-2.5	2.6	-0.28	-0.08
$t_{OLS}$	-19.2	20.8	-10.7	-5.6
$t_{t-error}$	-9.8	9.7	-2.6	-4.0
$t_{Ramsey}$	-19.1	20.7	-10.7	-5.6

Table 12: Trading time,  $T$ , log-centrality,  $c$ , log-degree,  $d$ , log-number of trades,  $n$ , and log-volume,  $v$ . Average trading time is negatively related to centrality, i.e., central investors trade earlier with respect to information events than peripheral investors.

<i>Ticker</i>	<i>Movement date</i>	<i>Ticker</i>	<i>Movement date</i>	<i>Ticker</i>	<i>Movement date</i>
AELS	3/9/2005	DOHOL	1/14/2005	NTHOL	4/7/2005
GIMA	4/22/2005	KARTN	1/25/2005	TEBNK	1/12/2005
KRDMA	5/17/2005	TSKB	1/14/2005	TUPRS	3/18/2005
BJKAS	7/26/2005	BJKAS	12/26/2005	SEKFK	7/8/2005
ULKER	2/14/2005	AKGRT	5/2/2005	VESTL	2/14/2005
BFREN	2/5/2005	BFREN	1/25/2005	DYOBY	7/14/2005
BOYNR	1/18/2005	ECILC	3/14/2005	TEKST	8/16/2005
TNSAS	8/19/2005	DEVA	7/12/2005	EREGL	9/12/2005
SAHOL	10/4/2005	SISE	7/29/2005	SKBNK	7/12/2005

Table 13: Expanded set of information events, reported in public news outlets within 7 business days before and after the stock movement dates.

$T$	$c$	$c$	$d$	$n$	$v$
$\beta_{OLS}$	-0.19	-0.92	1.14	-0.20	-0.15
$t_{OLS}$	-16.6	-8.3	10.9	-8.9	-12.6
$t_{t-error}$	-12.3	-5.8	6.7	-3.3	-7.7
$t_{Ramsey}$	-16.5	-6.6	6.4	-0.55	-0.55

Table 14: Trading time regressed on log-centrality for extended set of information events (24 events). Average trading time is negatively related to centrality, both in univariate ( $T$  on  $c$ ) and in multivariate ( $T$  on  $c, d, n$  and  $v$ ) regressions.

$\mu$						$\mu^e$					
	$c$	$d$	$c-d$	$n$	$v$		$c$	$d$	$c-d$	$n$	$v$
	0.0030						0.0008				
		0.0028						0.0008			
			0.011						0.0018		
				0.035						0.0012	
					0.0018						0.0005
$t$	> 20	> 20	9.8	> 20	> 20	$t^e$	9.9	9.6	1.6	12.4	7.7
$\Delta\mu$	0.5%	0.5%	0.2%	0.6%	0.4%	$\Delta\mu^e$	0.2%	0.1%	0.03%	0.2%	0.13%

Table 15: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ), rescaled centrality ( $c-d$ ), log-trades ( $n$ ) and log-volume ( $v$ ). The  $\Delta t = 30$ -minutes window is used. Tests are out-of-sample, so that profits are calculated using the final four months of the year, whereas all other variables are constructed using the first eight months.

$\mu$					$\mu^e$				
	$c$	$d$	$n$	$v$		$c$	$d$	$n$	$v$
$\beta_{OLS}$	0.0110	-0.0124	0.0057	-0.0009	$\beta_{OLS}^e$	0.0010	-0.0022	0.0029	-0.0006
$t_{OLS}$	10.7	-12.1	> 20	-7.3	$t_{OLS}^e$	1.2	-2.5	13.8	-5.3
$\Delta\mu_{OLS}$	1.8%	-2.0%	1.0%	-0.23%	$\Delta\mu_{OLS}^e$	0.2%	-0.4%	0.5%	-0.2%
$\beta_{t-error}$	0.0126	-0.0146	0.0053	-0.0007	$\beta_{t-error}^e$	0.0021	-0.0037	0.0024	-0.0002
$t_{t-error}$	6.1	-7.1	10.8	-2.7	$t_{t-error}^e$	1.2	-2.6	5.8	-1.0
$\Delta\mu_{t-error}$	2.1%	-2.4%	0.9%	-0.17%	$\Delta\mu_{t-error}^e$	0.3%	-0.6%	0.4%	-0.06%
$\beta_{Ramsey}$	0.0113	-0.0127	0.0058	-0.0009	$\beta_{Ramsey}^e$	0.0012	-0.0024	0.0029	-0.0006
$t_{Ramsey}$	10.9	-12.3	> 20	< -7.1	$t_{Ramsey}^e$	1.3	< -2.7	13.6	-5.2
$\Delta\mu_{Ramsey}$	1.8%	-2.1%	1.0%	-0.23%	$\Delta\mu_{Ramsey}^e$	0.2%	-0.4%	0.5%	-0.1%

Table 16: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ), log-trades ( $n$ ) and log-volume ( $v$ ). The  $\Delta t = 30$ -minutes window is used. Coefficients are reported for OLS regression, as well as iteratively reweighed robust regressions with Ramsey's  $E$ -function, and OLS with heavy-tailed ( $t$ -distributed) errors. Tests are out-of-sample, so that profits are calculated using the final four months of the year, whereas all other variables are constructed using the first eight months.

$T$	$c$	$d$	$n$	$v$
$\beta_{OLS}$	-0.13	-0.03	0.11	-0.15
$t_{OLS}$	-13.8	-1.4	3.4	-5.6
$t_{t-error}$	-9.5	-12.9	9.5	-2.8
$t_{Ramsey}$	-13.7	-1.4	3.4	-5.5

Table 17: Trading time regressed on log-centrality ( $c$ ), log-degree ( $d$ ), log-number of trades ( $n$ ) and log-volume ( $v$ ). All variables, except  $T$ , are constructed using first six months of data.

$\mu$					$\mu^e$				
	$c$	$d$	$n$	$v$		$c$	$d$	$n$	$v$
$\beta_{OLS}$	0.065	-0.063	0.0078	0.00002	$\beta_{OLS}^e$	0.020	-0.020	0.003	0.000006
$t_{OLS}$	> 20	< -20	> 20	-1.5	$t_{OLS}^e$	15.7	-15.5	17.8	0.55
$\Delta\mu_{OLS}$	7.7%	-7.6%	1.5%	-0.04%	$\Delta\mu_{OLS}^e$	2.4%	-2.4%	0.5%	0.01%
$\beta_{t-error}$	0.056	-0.055	-0.00001	-0.00002	$\beta_{t-error}^e$	0.016	-0.016	0.0014	0.00006
$t_{t-error}$	> 20	< -20	15.3	-0.68	$t_{t-error}^e$	6.1	-6.3	4.3	0.26
$\Delta\mu_{t-error}$	6.6%	-6.6%	1.0%	-0.04%	$\Delta\mu_{t-error}^e$	1.9%	-2.0%	0.3%	0.01%
$\beta_{Ramsey}$	0.065	-0.063	0.0076	-0.00002	$\beta_{Ramsey}^e$	0.020	-0.020	0.0029	0.000002
$t_{Ramsey}$	> 20	< -20	> 20	-1.7	$t_{Ramsey}^e$	15.3	-15.7	17.5	0.23
$\Delta\mu_{Ramsey}$	7.6%	-7.6%	1.5%	-0.04%	$\Delta\mu_{Ramsey}^e$	2.3%	-2.4%	0.5%	0.006%

$\mu$						$\mu^e$					
	$c$	$d$	$c-d$	$n$	$v$		$c$	$d$	$c-d$	$n$	$v$
	0.0091						0.0028				
		0.0083						0.0025			
			0.0028						0.0009		
				0.0071						0.0023	
					0.0054						0.0046
$t$	> 20	> 20	> 20	> 20	> 20	$t^e$	> 20	> 20	7.7	> 20	> 20
$\Delta\mu$	1.1%	1.0%	0.3%	1.4%	1.0%	$\Delta\mu^e$	0.3%	0.3%	0.01%	0.5%	0.4%

Table 18: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ) log-trades ( $n$ ) and log-volume ( $v$ ). A time window of  $\Delta t = 1$  day is used. The sample is restricted to one third of the investors (193,000), and centrality and degree is calculated using the first three months of trades.

$\mu$				
	$c$	$d$	$n$	$v$
$\beta_{OLS}$	0.00042	-0.0013	0.0071	-0.0017
$t_{OLS}$	6.9	-8.4	> 20	< -20
$\Delta\mu_{OLS}$	0.10%	-0.28%	1.3%	-0.49%

Table 19: Normalized profits,  $\mu$ , log-degree ( $d$ ) log-centrality ( $c$ ), log-number of trades ( $n$ ) and log-volume ( $v$ ) with threshold  $M = 10$  and a time window of  $\Delta t = 15$ -minutes. The data is truncated, such that the investors in the bottom two percentiles and top two percentiles of connectedness are discarded from the data.

<i>D</i> -range		250-275	500-550	4,000-4,400
<i>c</i>	$\beta_{OLS}$	0.025	0.054	0.049
	$t_{OLS}$	5.63	16.0	13.7
	$\Delta\mu_{OLS}$	0.59%	1.1%	0.70%
	$\beta_{t-error}$	0.036	0.049	0.020
	$t_{t-error}$	4.0	7.2	2.7
	$\Delta\mu_{t-error}$	0.83%	1.0%	0.3%
	$\beta_{Ramsey}$	0.026	0.053	0.049
	$t_{Ramsey}$	5.83	15.7	13.7
	$\Delta\mu_{Ramsey}$	0.62%	1.0%	0.69%
<i>d</i>	$\beta_{OLS}$	-0.0085	-0.064	-0.045
	$t_{OLS}$	-0.21	-5.4	-3.0
	$\Delta\mu_{OLS}$	-0.02%	-0.34%	-0.12%
	$\beta_{t-error}$	-0.018	-0.055	-0.026
	$t_{t-error}$	-0.24	-2.3	-0.84
	$\Delta\mu_{t-error}$	-0.05%	-0.29%	-0.07%
	$\beta_{Ramsey}$	-0.0085	-0.064	-0.044
	$t_{Ramsey}$	-0.22	-5.3	-3.0
	$\Delta\mu_{Ramsey}$	-0.02%	-0.34%	-0.12%
<i>n</i>	$\beta_{OLS}$	0.0073	0.0033	0.016
	$t_{OLS}$	5.54	4.2	> 20
	$\Delta\mu_{OLS}$	0.67%	0.33%	1.6%
	$\beta_{t-error}$	0.059	0.0047	0.0075
	$t_{t-error}$	2.21	2.9	5.6
	$\Delta\mu_{t-error}$	0.54%	0.46%	0.75%
	$\beta_{Ramsey}$	0.073	0.0036	0.016
	$t_{Ramsey}$	5.45	4.5	> 20
	$\Delta\mu_{Ramsey}$	0.67%	0.36%	1.6%
<i>v</i>	$\beta_{OLS}$	-0.0013	-0.00009	-0.0017
	$t_{OLS}$	-2.1	-0.22	-4.9
	$\Delta\mu_{OLS}$	-0.2%	-0.17%	-0.26%
	$\beta_{t-error}$	-0.0011	-0.0006	-0.0016
	$t_{t-error}$	-0.91	-0.79	-0.33
	$\Delta\mu_{t-error}$	-0.2%	-0.03%	-0.04%
	$\beta_{Ramsey}$	-0.0013	-0.0002	-0.0002
	$t_{Ramsey}$	-2.1	-0.48	-4.9
	$\Delta\mu_{Ramsey}$	-0.2%	-0.12%	-0.27%

Table 20: Normalized profits,  $\mu$ , for subsample of investor network with degree between 250 – 275, 500 – 600, and 4,000 – 4,400. The window length is  $\Delta T=30$  minutes.

$\mu$					$\mu^e$				
	$c$	$d$	$n$	$v$		$c$	$d$	$n$	$v$
$\beta_{OLS}$	0.0107	-0.0131	0.0079	-0.0017	$\beta_{OLS}^e$	0.0103	-0.0133	0.0039	-0.0004
$t_{OLS}$	> 20	< -20	> 20	< -20	$t_{OLS}^e$	> 20	< -20	> 20	-6.5
$\Delta\mu_{OLS}$	2.2%	-2.6%	1.5%	-0.47%	$\Delta\mu_{OLS}^e$	2.1%	-2.6%	0.72%	-0.12%
$\beta_{t-error}$	0.0124	-0.0154	0.0066	-0.0013	$\beta_{t-error}^e$	0.0099	-0.0134	0.0035	-0.0005
$t_{t-error}$	12.9	-15.5	> 20	-8.6	$t_{t-error}^e$	11.9	-15.5	14.3	-3.6
$\Delta\mu_{t-error}$	2.5%	-3.1%	1.2%	-0.35%	$\Delta\mu_{t-error}^e$	2.0%	-2.7%	0.66%	-0.13%
$\beta_{Ramsey}$	0.0113	-0.0138	0.0079	-0.0017	$\beta_{Ramsey}^e$	0.0011	-0.0137	0.0039	-0.0004
$t_{Ramsey}$	> 20	< -20	> 20	< -20	$t_{Ramsey}^e$	> 20	< -20	> 20	-6.8
$\Delta\mu_{Ramsey}$	2.3%	-2.7%	1.5%	-0.47%	$\Delta\mu_{Ramsey}^e$	2.1%	-2.7%	0.73%	-0.13%

Table 21: Normalized profits,  $\mu$ , and excess profits,  $\mu^e$ , regressed on log-centrality ( $c$ ), log-degree ( $d$ ) log-trades ( $n$ ) and log-volume ( $v$ ). The  $\Delta t = 30$ -minutes window is used. Coefficients are reported for OLS regression, as well as iteratively reweighed robust regressions with Ramsey’s  $E$ -function, and OLS with heavy-tailed ( $t$ -distributed) errors. The EIN is constructed so that no links between agents of the same brokerage house are permitted.

$M$	30	60	90	120	150	180	210
$E[\eta_{M/2}^M   p = 1/2]$	0.0087	0.0042	5.81E-5	2.29E-6	2.94E-8	1.09E-9	1.36E-11
$\bar{\eta}^M$	0.11	0.0042	0.0056	0.0089	0.066	0.14	0.50
$\hat{p}$	0.53	0.56	0.59	0.62	0.63	0.65	0.71
$t$	7.1	14	78	368	1,550	1.10E4	2.34E5
$E[\eta_{20}^M   p = 1/2]$	0.0066	0.290	1.00	1.93	2.93	3.903	4.89
$\bar{\eta}^M$	0.020	0.54	1.96	3.73	4.86	8.09	9.92
$\hat{p}$	0.55	0.54	0.54	0.53	0.53	0.54	0.54
$t$	12.7	13.1	14.8	13.4	8.4	11.6	10.5
# observations	1.23E5	1.78E4	5,369	2,372	1,240	681	423

Table 22: Upper panel: Average  $\eta_{M/2}^M$ , compared with expected  $\eta_{M/2}^M$  given  $p = 1/2$  for various  $M$ . Here,  $M$  and  $K$  are defined using “muted” trades which do not count multiple trades within the same time window. The probability is significantly higher than  $1/2$ . When GMM is used to match  $E[\eta_{M/2}^M | p] = \bar{\eta}^M$ ,  $p$  lies between  $0.53 - 0.71$ , depending on  $M$ . Lower panel: Same calculations for  $\eta_{20}^M$ . Again, the probability is significantly different from  $1/2$ .

	$N_{cutoff}$		300	1,000	3,000	10,000	30,000
$C$	$N_{cutoff}^*$	1,920					
	$\zeta_C^{CS}$	5.97	9.80	6.74	5.28	3.45	1.63
	KS error	0.047	0.082	0.077	0.057	0.074	0.091
$D$	$N_{cutoff}^*$	3,660					
	$\zeta_D^{CS}$	3.11	4.28	3.57	3.16	2.42	1.41
	KS error	0.029	0.049	0.052	0.033	0.052	0.074
$N$	$N_{cutoff}^*$	6,140					
	$\zeta_N^{CS}$	1.22	1.81	1.41	1.28	1.17	0.93
	KS error	0.0170	0.066	0.060	0.033	0.019	0.040
$P$	$N_{cutoff}^*$	17,090					
	$\zeta_P^{CS}$	0.86	0.98	0.82	0.82	0.87	0.74
	KS error	0.0074	0.053	0.033	0.012	0.011	0.039
$V$	$N_{cutoff}^*$	28,020					
	$\zeta_V^{CS}$	0.80	1.38	1.07	0.92	0.83	0.76
	KS error	0.0080	0.077	0.060	0.037	0.018	0.021

Table 23: Estimated tail-exponent,  $\zeta^{CS}$ , for degree ( $D$ ), centrality ( $C$ ), number of trades ( $N$ ), profits ( $P$ ) and Volume ( $V$ ). For degree and centrality, a 15-minute window length was used. Kolmogorov-Smirnoff (KS) statistic used to choose optimal cutoff point and maximum-likelihood estimator to choose  $\alpha$ , given cut-off point.

<i>C</i>							
$N_{cutoff}$		$N_{cutoff}^*$	300	1,000	3,000	10,000	30,000
Pareto	lr	-2.84E3	-3.71E2	-1.46E3	-4.588E3	-1.67E4	-8.65E4
	<i>t</i>	6.21	5.65	6.67	5.63	13.4	14.5
Exponential	lr	<b>-2.81E3</b>	-3.66E2	<b>-1.44E3</b>	<b>-4.51E3</b>	<b>-1.63E4</b>	<b>-8.27E4</b>
	<i>t</i>		2.5				17.7
Log-normal	lr	-2.82E3	<b>-3.65E2</b>	-1.43E3	-4.52E3	-1.63E4	-8.31E4
	<i>t</i>	4.14		0.95	5.73	6.22	

<i>D</i>							
$N_{cutoff}$		$N_{cutoff}^*$	300	1,000	3,000	10,000	30,000
Pareto	lr	<b>-3.980E4</b>	-3.338E3	-1.112E4	<b>-3.276E4</b>	-1.075E5	-5.22E5
	<i>t</i>		0.77	0.15			> 20
Exponential	lr	-3.982E4	<b>-3.337E3</b>	<b>-1.111E4</b>	-3.277E4	-1.076E5	-5.21E5
	<i>t</i>	3.37			3.22	4.82	> 20
Log-normal	lr	-3.988E4	-3.338E3	-1.112E4	-3.281E4	<b>-1.074E5</b>	<b>-5.19E5</b>
	<i>t</i>	15.6	1.80	3.68	1.61	10.0	

<i>N</i>							
$N_{cutoff}$		$N_{cutoff}^*$	300	1,000	3,000	10,000	30,000
Pareto	lr	<b>-5.70E4</b>	<b>-3.317E3</b>	-1.0534E4	<b>-2.94E4</b>	<b>-8.91E4</b>	<b>-3.814E5</b>
	<i>t</i>			0.68			
Exponential	lr	-5.87E4	-3.323E3	-1.061E4	-3.00E4	-9.24E4	-4.05E5
	<i>t</i>	14.5	1.46	5.26	9.21	18.4	> 20
Log-normal	lr	-5.72E4	-3.319E3	<b>-1.0527E4</b>	-2.95E4	-8.94E4	-3.818E5
	<i>t</i>	7.54	0.36		3.38	10.0	6.05

<i>P</i>							
$N_{cutoff}$		$N_{cutoff}^*$	300	1,000	3,000	10,000	30,000
Pareto	lr	<b>-1.89E5</b>	<b>-4.72E3</b>	<b>-1.449E4</b>	<b>-3.94E4</b>	<b>-1.17E5</b>	<b>-4.94E5</b>
	<i>t</i>						
Exponential	lr	-2.13E5	-4.79E3	-1.50E4	-4.22E4	-1.30E5	-
	<i>t</i>	> 20	3.7	8.5	14.4	20.0	> 20
Log-normal	lr	-1.90E5	-4.72E3	-1.451E4	-3.96E4	-1.18E5	-4.96E5
	<i>t</i>	> 20	0.07	1.78	8.68	19.6	13.8

<i>V</i>							
$N_{cutoff}$		$N_{cutoff}^*$	300	1,000	3,000	10,000	30,000
Pareto	lr	<b>-4.63E5</b>	-6.25E3	-2.00E4	<b>-5.71E4</b>	<b>-1.778E5</b>	<b>-7.93E5</b>
	<i>t</i>		1.08	0.86			
Exponential	lr	-4.92E5	-6.26E3	-2.02E4	-5.83E4	-1.85E5	-8.5E5
	<i>t</i>	2.92	3.35	6.98	12.17	> 20	> 20
Log-normal	lr	-4.65E5	<b>-6.24E3</b>	<b>-2.00E4</b>	-5.72E4	-1.781E5	-7.95E5
	<i>t</i>	15.6			1.62	10.0	> 20

Table 24: *C*-panel: Likelihood ratio test for centrality distribution, *C*, comparing Pareto, exponential and one-sided log-normal distribution in tails, with different cutoffs,  $N_{cutoff}$ . Exponential distribution dominates for  $N_{cutoff} \geq 500$ . Log-normal distribution dominates for  $N_{cutoff} < 500$ . *D*-panel: Likelihood ratio test for degree distribution *D* in tails, with different cutoffs,  $N_{cutoff}$ . Pareto distribution provides best fit for cutoff points 1,  $500 \leq N_{cutoff} \leq 600$ . Below, Exponential distribution dominates and above log-normal distribution dominates. *N*-panel: Likelihood ratio test for distribution of number of trades *N* in tails, with different cutoffs,  $N_{cutoff}$ . Pareto distribution provides best fit for cutoff points above  $N_{cutoff} = 1, 500$  and below  $N_{cutoff} = 500$ . Between, for  $500 \leq N_{cutoff} \leq 1, 500$ , Log-normal distribution provides best fit. *P*-panel: Likelihood ratio test for profit distributions *P* in tails, with different cutoffs,  $N_{cutoff}$ . Pareto distribution provides best fit for all cutoff points. *V*-panel: Likelihood ratio test for distribution of trading volume *V* in tails, with different cutoffs,  $N_{cutoff}$ . Pareto distribution provides best fit for cutoff points above  $N_{cutoff} = 1, 500$ .