Investor Networks in the Stock Market*

Han Ozsoylev† Johan Walden‡ M. Deniz Yavuz§ Recep Bildik¶

April 22, 2011

Abstract

We study investor networks in the stock market, through the lens of information network theory. We use a unique account level dataset of all trades on the Istanbul Stock Exchange in 2005, to identify traders who are similar in their trading behavior as linked in an empirical investor network (EIN). This empirical investor network is consistent with several predictions from the theory of information networks. The EIN is relatively stable over time, some investors systematically trade before their neighbors in the network, centrally placed investors earn higher profits, and the cross sectional distributions of profits and trading volume are heavy-tailed with similar tail exponents. We also identify several theoretical challenges for future research.

Keywords: *Information networks, heterogeneous investors, portfolio choice, power law.*

*__Working paper: Comments are welcome.__ We thank seminar participants at the Swedish Institute for Financial Research (SIFR), the Berkeley finance lunch seminar, Johns Hopkins University and University of Minnesota. 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. We are grateful to Jennifer Conrad, Nicolae Garleanu, Simon Gervais, and Raghu Rau for valuable suggestions. Finally, we thank Gil Shallom for developing the initial code.

†Said 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.

‡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.

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

¶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 sources of investor heterogeneity have been identified 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, Lee, Liu, and Odean (2009); Fedyk, Heyderdahl-Larsen, and Walden (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.[1] 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 (Ozsoylev and Walden (2010)). The underlying assumption of this literature is that agents receive information signals about stocks and that these signals diffuse through an information network. Some agents receive signals earlier than others, depending on their position in the network. Agents who are centrally placed tend to receive signals early, whereas agents who are in the periphery tend to receive them later—perhaps not until the signals are completely incorporated into prices and are thereby public. As a result, the trading behavior and profitability of agents are determined by their position in the network, and the dynamics of aggregate asset prices depend on general topological properties of the network. Specifically, agents who are more centrally placed in the network have higher trading profits.

Identifying the underlying information network for 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 can 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 Trading Network (EIN), and in a stylized model we use simulations to show

[1] There is also a large literature that explains heterogeneous portfolio holdings with hedging motives (see, e.g., Mayers (1973); Bodie, Merton, and Samuelson (1992); Massa and Simonov (2006); Parlour and Walden (2011), Betermeier, Jansson, Parlour, and Walden (2011)). For example, investors who differ in the human capital risk they are facing will this risk in capital markets and thereby choose different investment portfolios. Such considerations will 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 will also induce trading, but also typically at a low frequency of rebalancing. We include such trading motives in our broad definition of "investment styles."

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, to test the predictions of the theory. We first verify that the EIN is fairly stable over time, since if the EIN proxies for a true information network it should not vary drastically from one period to another. 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 two EINs are similar: the likelihood that two investors are linked in the second period, given that they are linked in the first, is between 7.5-113 times higher than what would be the case if the two networks were randomly generated, depending on the specification and test used. 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 over alternative explanations, e.g., liquidity provision and style investing, providing our second contribution.

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 0.2-2.8% 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 made on average 20-280 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 study the cross sectional distribution of profits, number of trades, trading volume, number of connections and centrality. Using Kolmogorov-Smirnov tests, we find that profits and trading volume are severely heavy-tailed, consistent with power-law distributions, whereas the distributions of connectedness and centrality are thinner tailed, better approximated by log-normal or exponential distributions. We relate these cross sectional results with those reported in Gabaix, Gopikrishnan, Plerou, and Stanley (2003) for the time series behavior of market returns and trading volume, and show that the two sets of results together raise interesting challenges for future theoretical research.

We believe that this line of research has the potential to shed light on several fundamental questions about stock markets. For example, it well-known that large aggregate stock market movements rarely are accompanied by public news events (see, Cutler, Poterba, and Summers (1989), and Fair (2002)). Such movements may suggest that information "shocks" occur at a

more heterogeneous level than the public arena, and that an important information aggregation mechanism occurs at a more local level. There may also be further implications. As discussed, e.g., in Gabaix, Gopikrishnan, Plerou, and Stanley (2003), the distributions of returns, trading volume and number of trades over time are heavy-tailed, implying that extreme events like stock-market crashes occur relatively frequently. A better understanding of the structure of information diffusion in the market may help in our understanding of these distributional properties of aggregate stock markets.

Our paper belongs to the literature on 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: they attribute more than a quarter of the correlation between households' stock purchases and stock purchases made by their neighbors to word-of-mouth communication. Hong, Kubik, and Stein (2004) provide further evidence that fund managers' portfolio choices are influenced by word-of-mouth communication. Cohen, Frazzini, and Malloy (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)). Our study contributes to this 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 investor behavior and trading profits.

Information diffusion may occur through social networks, but we stress that we view the concept of information networks more broadly, given that other channels of diffusion also exist (e.g., Internet discussion boards and newsletters). Our dataset does not allow us to distinguish between these different sources of information diffusion. For a further discussion about general information diffusion in stock markets, see Ozsoylev and Walden (2010).

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. In Section 3, we describe the data and the methodology used to create the EIN, 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 cross sectional distributions. Concluding remarks are made in Section 5.

2 Model

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

Let us for simplicity assume a network structure according to Figure 1, in which there are $N_I = 21$ agents in an information network. Each node (circle) in the network represents a 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. We assume that 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. 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.

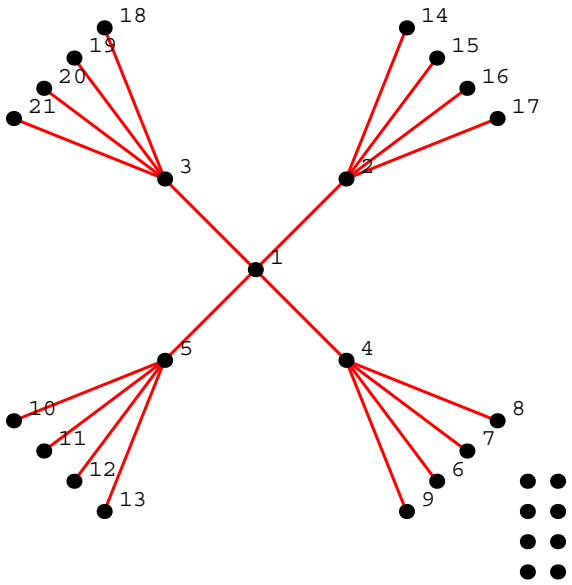

Figure 1: *Information network of 21 agents in a market. Each agent is represented by a node (a filled circle). An edge (line) between two agents represents that these are connected in the network, i.e., that they are neighbors. In addition, there is a large number of liquidity traders, represented by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

Trading occurs at discrete times, $t = 0, 1, 2, \ldots$. At each point in time, one of the N_I agents in the network receives a valuable signal about stocks in the market, and trades on this information. Thus, at time t, agent n_t receives a signal and trades. The expected profits from this trade is $\pi > 0$. We do not specify further what types of trades the signal generates. 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 trades. For simplicity, we assume that there is a noise trader, who is willing to take the opposite position in the trade (with expected profits of $-\pi$), whereas agents in the information network only trade when they receive information.[2]

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 think of a situation, 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 the day after agent n_t received the signal, in which case he shares his information.

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 public channels. 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 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 information 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

[2]These assumptions are, of course extremely stylized, but as shown in Ozsoylev and Walden (2010) 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.

(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

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

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

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 (2010).

There is a large literature on how to measure centrality (see, e.g., Chung and Li (2006) and references therein). We will use the notion of *eigenvector centrality*, but to define this measure we need to introduce some notation. A general network of N agents can be represented by a neighborhood 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.[3] 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 a $N \times N$ neighborhood matrix, where we use the convention that the first N_I agents are the ones in the information network, and the remaining N_U are the noise traders (each of which is only connected to himself). The market's

[3]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 ith row and jth column of the matrix $\mathcal{E}$, or, if there can be no confusion, we write it as $\mathcal{E}_{ij}$.

neighborhood matrix can therefore be decomposed into

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

where $\mathbf{I}$ is a $N_U \times N_U$ identity matrix, representing that noise traders are connected to themselves, but have no neighbors in the information network, $\mathbf{0}$ is a $N_I \times N_U$ matrix of zeros representing that there are no connections between agents in the information network and noise traders, and $\mathcal{E}^I$ is a $N_I \times N_I$ matrix representing the connections among agents in the information network. For example, the network 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, i.e., elements for which $\mathcal{E}_{ij} = 1$. The upper left part of the matrix represents the agents in the information 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.

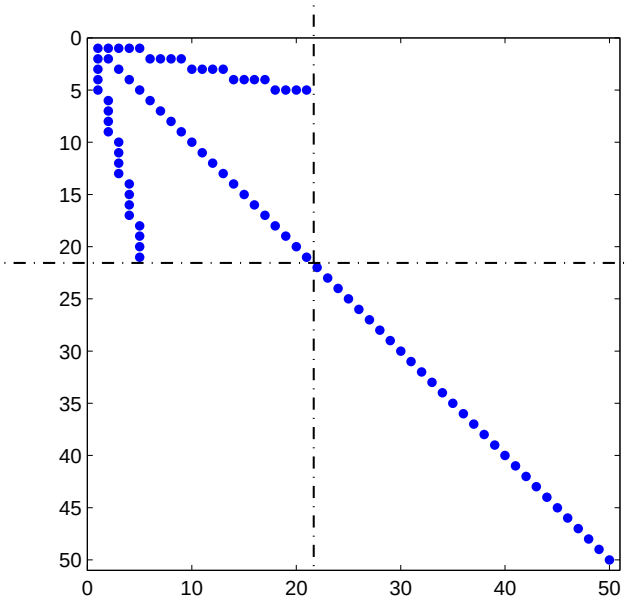

Figure 2: *The network matrix for the network with $N_I = 21$ agents in the information network and $N_U = 29$ noise traders.*

The neighborhood matrix, $\mathcal{E}$ contains all relevant information about the network. For example,

a vector with the degree of agents can be constructed by summing all the elements in each row of the neighborhood matrix,

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

where $\mathbf{1}$ is a vector of ones. We can also calculate the m-order degree vectors, where the ith element is

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

Here, $\mathcal{E}^m$ denotes the mth power of the matrix $\mathcal{E}$ using standard matrix algebra, 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 $\mathbf{A}_{ij} = 0$.

Similarly, a vector C where the ith 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 investors who are connected to her, we derive

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

The proportionality constant, λ, is an eigenvalue of $\mathcal{E}$ and C is the corresponding eigenvector. The eigenvector corresponding to the largest eigenvalue is the centrality vector.[4] We note that (5) determines C uniquely, down to an arbitrary scaling constant. For large matrices, an efficient way of solving (5) is by using *power iterations*.[5]

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

[4]For an $N \times N$ matrix, $\mathbf{A}$, the set of eigenvalues contains the roots of an Nth 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.

[5]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 a few iterations.

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. 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 reasons. These 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 these traders are not central in the information network.

There are also purely technical advantage in working with centrality: to calculate the mth degree vector, $\mathcal{E}^m$ needs to 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 λ^1 using the power method, only matrix by vector calculations using $\mathcal{E}$ are needed, and convergence typically occurs within less than twenty iterations. We will therefore mainly work with C, even though we have also calculated D^2 for comparison in some of our tests.

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, which 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 travel 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.

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

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, so we need to estimate it from observed trades. Within the previous model, we can generate an Empirical Investor Network (EIN), based on the key observation that agents who are connected in the network will tend trade in similar stocks in the same direction at similar points in time. Thus, given observed trades by agents in the market, we identify two agents as being linked, if they are either both buying or both selling the same stocks within the time period $[t, t+1]$, for some t.

We do this for a slightly more general information diffusion model than the one just described. Specifically, we assume that the agent who gets the initial signal, with probability q_1 shares with each of his neighbors, and that the events of sharing with different neighbors are independent, so that an agent can independently "have lunch" with two or more of his neighbors between t and $t+1$. 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.

A realized EIN is shown in Figure 3. We see that the general structure of the true network is

identified correctly, although not every link is correct. For example, in the upper left part of the EIN that 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 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.

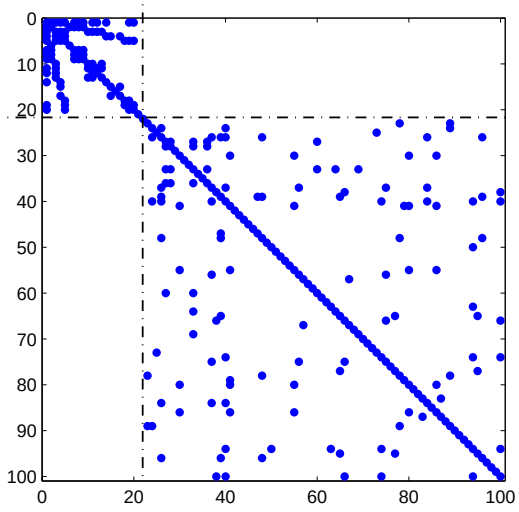

Figure 3: *The empirical information network, EIN, generated in a simulation, when the true network has* $N = 100$ *traders.*

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 in this process. 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 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 $Corr(C, C_{EIN}) = 0.64$. Moreover, so is the 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, $Corr(C_{EIN}/D_{EIN}, P) = 0.55$, 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 although the described mechanism for information diffusion is quite specific, a broader interpretation of what constitutes an information network, as discussed in Ozsoylev and Walden (2010), is also possible. In a general information network, agents differ on when they act on information signals. This could be because they receive signals at different points time, for example because of diffusion through a social network as in our stylized model in this paper, but also because of other diffusion mechanisms, e.g., that some agents read information in the local news paper a day after other agents see it on TV. It may even be that some agents take more time to process complex public signals than others. These different stories — which our data will not be able to distinguish between — have the common property that information is gradually incorporated into agents' trading behavior and asset prices.

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 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 use the relationship between connectedness and trades, described in the theory section, to back out the EIN from the trading data. Specifically, we choose a time window, ΔT, and a threshold,

M. If two investors traded in the same stock in the same direction (either both being on the buy side or on the sell side of the trades) within the time period ΔT, at least M times, then they are defined to be connected in the EIN, $\mathcal{E}_{ij}^{\Delta t,M} = 1$. We note that our definition of EIN is different from the one taken in Adamic, Brunetti, Harris, and Krilenko (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, in line with the previous section, all links in the network are bidirectional, i.e., if investor i is linked to j, then j is also linked to i. Thus, we do not take into account which investor traded first when creating the EIN. We mainly focus on the case where the threshold for a connection is $M = 1$, but we will also show that similar results arise when we choose a higher threshold, $M = 10$.

We vary the length of the window between one minute up half an hour. The results are similar for different window lengths. This is quite unsurprising, since two investors who are directly linked when a window length of ΔT is chosen, are typically also indirectly connected (at a higher degree than one) with a window length $\Delta T' < \Delta T$.[6] 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 half an hour, 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 up to about a week, taking higher distance effects into account. For computational reasons, we limit the maximum window length to half an hour, and for several analyses we allow ourselves to use shorter window lengths, since it turns out that the structure of the EIN is similar across window lengths (see Section 4.3).

In Table 2, we provide summary statistics for the EIN, using different window lengths. 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 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

[6] 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.

ΔT	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.

a good proxy is given by identifying traders with extremely high trading frequencies, which we will control for.

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 (see Section 4.3). 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 ith 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) 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, Lee, Liu, and Odean (2009), but focusing on individual investors' trades rather than on investor groups. Briefly, we define a window length, ΔT_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 (#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.[7] 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 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 into account trading volumes. We divide these profits

[7]Here, we use the ISE 100 index as a proxy for the market.

by the total trading volume, to get normalized profit vectors, μ and μ^e, i.e.,

$$\mu_i = \frac{P_i}{V_i}, \qquad \text{and} \qquad \mu_i^e = \frac{P_i^e}{V_i}, \tag{6}$$

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 15-minute window for the EIN, $\Delta T = 15$ 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 will verify that the distributions are indeed heavy-tailed in Section 4.4; (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.

To verify that our measures of information diffusion at higher distances, C and D^2, contain additional information beyond what is provided by connectedness, we show the distribution of centrality and second-order degree, given that agents have M links, $D_i = M$, for some different M's. Indeed, there is considerable dispersion of the higher order degrees, even when conditioned on a specific number of connections.

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 surprising, institutional investors trade in much larger volumes.

Finally, since the individuals investors make up the vast majority of investors in the market, 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 permille 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 future 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

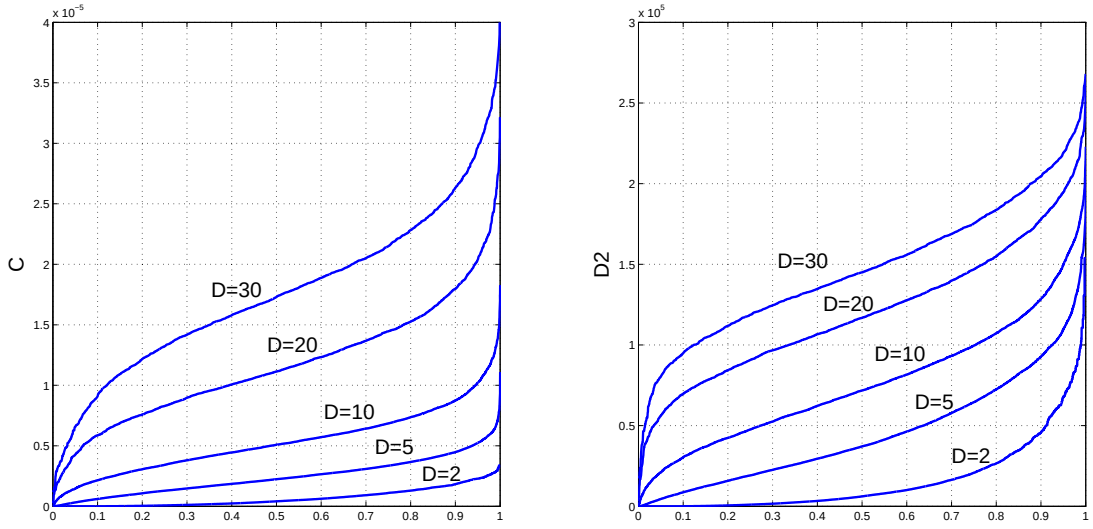

Figure 4: *Distribution of centrality, C_i, (left) and of second degree, D_i^2, (right) given first degree, $D_i = M$, $M = 2, 5, 10, 20, 30$.*

excluded — for the sake of robustness.

4 The EIN, centrality, profits and distributions

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 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

				Corr					
	Mean	Std. dev.	Median	C	D	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
				Corr					
	Mean	Std. dev.	Median	c	d	n	μ	μ^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			
μ	-0.014	0.085	-0.038	0.052	0.055	0.083	1		
μ^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 (μ, μ^e) and log-volume (v).

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 $k_1 << K$, $k_2 << K$, it follows immediately that the expected number of overlaps is approximately[8]

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

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

[8]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 $k_1 << K$ and $k_2 << K$.

	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.

be Poisson distributed. However, our detailed analysis of the degree distribution in Section 4.4 will show 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 $k_2 << K$, 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

$$E_{\text{Degree adjusted}}[y] \approx \frac{k_2}{2k_1N}\sum_{i=1}^{N}(D_i-1)^2. \tag{8}$$

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

ΔT=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				ΔT=5 minutes
ΔT=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.

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.[9]

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 324, 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 relative stable over time, which is consistent with the intuition that they are proxies of information networks.

4.2 Sequencing of trades

A crucial feature of the information diffusion model is that some traders will systematically trade before others. For example, in the network model introduced in Section 2, 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

[9] 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.

that i traded before j, and defining $M = M_{ij}$, is

$$E\left[\eta^M \middle| p\right] = \sum_{k=0}^{M} \left(\begin{array}{c} M \\ k \end{array} \right) \Big|2k - M\Big| p^k (1-p)^k, \tag{9}$$

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 η 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.

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.

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 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 η^M is close to zero with the new definition.

We verify this bias for low M. For example, when studying η^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 η^M away from 0. Specifically, we define the random variable

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

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

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

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

Thus, to summarize, the adjusted measure mutes trades when calculating K_{ij}, K_{ji} and M_{ij}, which leads a different empirical distribution of η^M. The adjusted measure controls for split trades, but introduces a downward bias of η, through overbalancing K_{ij} and K_{ji}. This, in turns, leads us to study η_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 $z > 0$ focuses our tests on traders who are severely unbalanced in their relationship in that one trader went before the other many times. In Table 7 we show the tests with the new measure, for M from 30 to 210, using $z = M/2$ and $z = 20$ as cutoff points.[10]

In both cases, we strongly reject that $p = 1/2$, although the GMM-estimates of $\hat{p}$ is closer to

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

M	5	10	30	60	90	120	150	180	210
$E[\eta^M \mid p = 1/2]$	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 η^M, compared with expected η^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.

1/2 in both cases. In the first case, in which $z = M/2$ 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 bounds on the true probabilities, which is still higher than 1/2.

We also verify that trading before ones' neighbors is related to centrality. Specifically, we define the vector w, where the ith 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).

We thus find support for the information diffusion story of trading behavior over the style investing and liquidity provision stories.

4.3 Centrality and Profits

The theory suggests that more centrally placed investors, all else equal, are more profitable than less centrally placed 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 disentangle trading motives from 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 regress normalized profits, μ, and normalized excess profits, μ^e, on log-volume, connectedness and centrality, using a 30-minute time window. To avoid influence by outliers, we truncate the

M	30	60	90	120	150	180	210
$E[\eta^M_{M/2}\|p=1/2]$	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
$E[\eta^M_{20}\|p=1/2]$	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 7: Upper panel: Average $\eta^M_{M/2}$, compared with expected $\eta^M_{M/2}$ 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 $E[\eta^M_{M/2}|p] = \bar{\eta}^M$, p lies between $0.53 - 0.71$, depending on M. Lower panel: Same calculations for $\eta^M_2 0$. Again, the probability is significantly different from 1/2.

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 8, 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.

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 iterated re-weighted least square (using Ramsey's E-function). The multivariate results are consistent, as shown in Table 9. 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 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%.

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 in 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. Given the large sample size, it is not surprising that 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 9, but regressed on $c - d$. The results, shown in Table 10, all come out with the a positive sign, and strongly statistically significant, although the economic significance is lower for $c - d$. The effect on μ and μ^e of a one standard deviation increase in $c - d$ is an increase in trading profits by between $0.23\% - 0.36\%$, depending on which regression is used (OLS, OLS with heavy-tailed error terms, or iterative re-weighted least square).

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 region. The results are shown in Table 11. 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, 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 9. 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.

We have also performed the same tests, but excluding the institutional investors, with virtually identical results. 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 differences between 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.

μ						μ^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 8: Normalized profits, μ, and excess profits, μ^e, regressed on log-centrality (c), log-degree (d), rescaled centrality ($c-d$), log-trades (n) and log-volume (v). The Δt =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.

We also run the same tests as in Table 9 with a higher threshold for two investors to be identified as being connected, $M = 10$. The results, summarized in Table 12, 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.

We note that the argument in the previous section 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 because the results go in the wrong direction for such short horizons. Indeed, if momentum was a driving force behind

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

Table 9: Normalized profits, μ, and excess profits, μ^e, regressed on log-centrality (c), log-degree (d) log-trades (n) and log-volume (v). The Δt =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.

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 8). It therefore seems like momentum is not a driving force behind the positive relationship between profits and centrality we find in this study.

Eigenvector centrality is just one method of estimating higher order connectedness and thereby information precision. As argued in the theory section, other functions of higher order degrees may also work, e.g., second order degree. We add log-second order degree, d^2 to the regression, where $d_i^2 = \log(D_i^2)$ for investor i. Computational limitations force us to use the EIN calculated with a 1-minute window length when calculating D^2. The results of the previous tests with shorter windows are qualitatively the same as when using window a window length of 30 minutes, although the economic significance is not as strong.

The results are shown in Table 13, both with and without d^2. Interestingly, d^2 dominates c when included, although c captures most of the influence on profits of higher order connections when d^2 is left out. The main advantage of using c is that it is substantially easier to calculate

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

Table 10: Normalized profits, μ, and excess profits, μ^e, regressed on log-rescaled centrality, $c-d$, n and v. The Δt =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.

numerically. As mentioned, the computational issues of calculating d^2 forced us to restrict the window length to one minute, for which the centrality results are weaker than when longer window lengths are used. It is an open question which measure is best with a longer window length. More generally, an interesting question for future research is to further explore which type of higher-order connectedness—centrality, higher-order degree, or some other measure—best captures the value of an investor's position in the information network, while still being computationally efficient.

4.4 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, Gopikrishnan, Plerou, and Stanley (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* α.

We provide a brief summary of the arguments made in Gabaix, Gopikrishnan, Plerou, and Stanley (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[11]; (b) The price impact of a trade on the stock's price 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:

$$\zeta_V^{AG} = 1.5, \tag{12}$$

which together with (b) leads to

$$\zeta_r^{AG} = 2\zeta_V^{AG} = 3. \tag{13}$$

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

[11] In Gabaix (1999), a theoretical argument is made for why Zipf-law distributed sizes arise in various contexts.

that

$$\zeta_N^{AG} = 2\zeta_V^{AG} = 3. \tag{14}$$

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 δ 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, $corr(v-n, n) = 0.29$, 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, Gopikrishnan, Plerou, and Stanley (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 $N_i \sim S^{1-3\delta/2}$, it follows that the cross section of number of trades, N, is power law distributed with tail exponent

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

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

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

Gabaix, Gopikrishnan, Plerou, and Stanley (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 (2010), implies that centrally placed investors are better informed, trade more, and make higher profits. Within the specific framework of Ozsoylev and Walden (2010), 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

$$\zeta_C^{CS} = \zeta_P^{CS} = \zeta_V^{CS}. \tag{17}$$

Of course, C is just a proxy and, the linearity of these relations depend on strong assumptions made in Ozsoylev and Walden (2010), 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 5, 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.

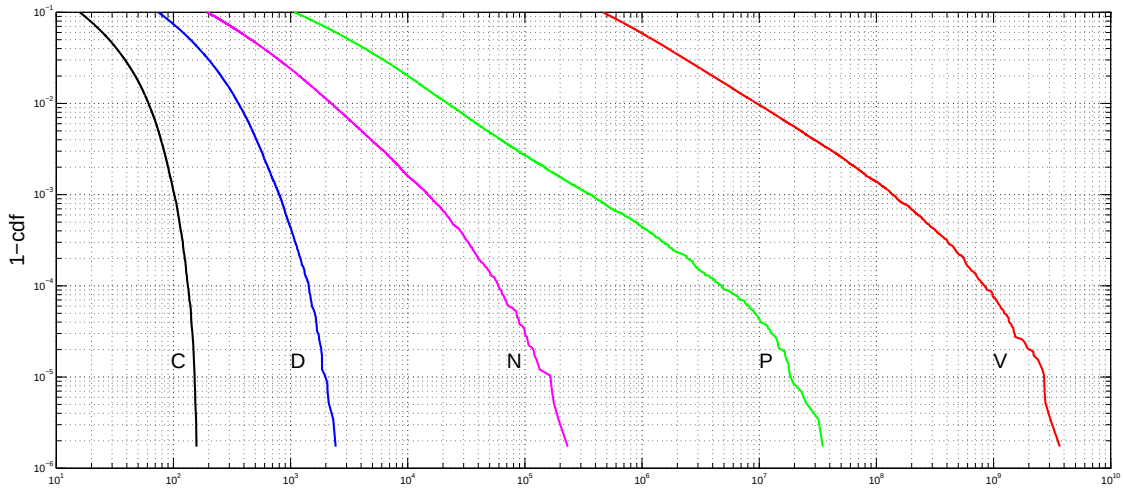

Figure 5: *Tail distribution of centrality (C), connectedness (D) (measured in hundreds of connections), number of trades (N), profits (P) in TL and trading volume (V) in TL.*

We estimate how consistent the observed distributions are with Pareto-law distributed tails. We vary the tail distribution, ζ 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, ζ, and cutoff, N_{cutoff}, to minimize the KS error.

The results are given in Table 14, 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 15. 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 6, 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 ζ_C^{CS} and ζ_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.

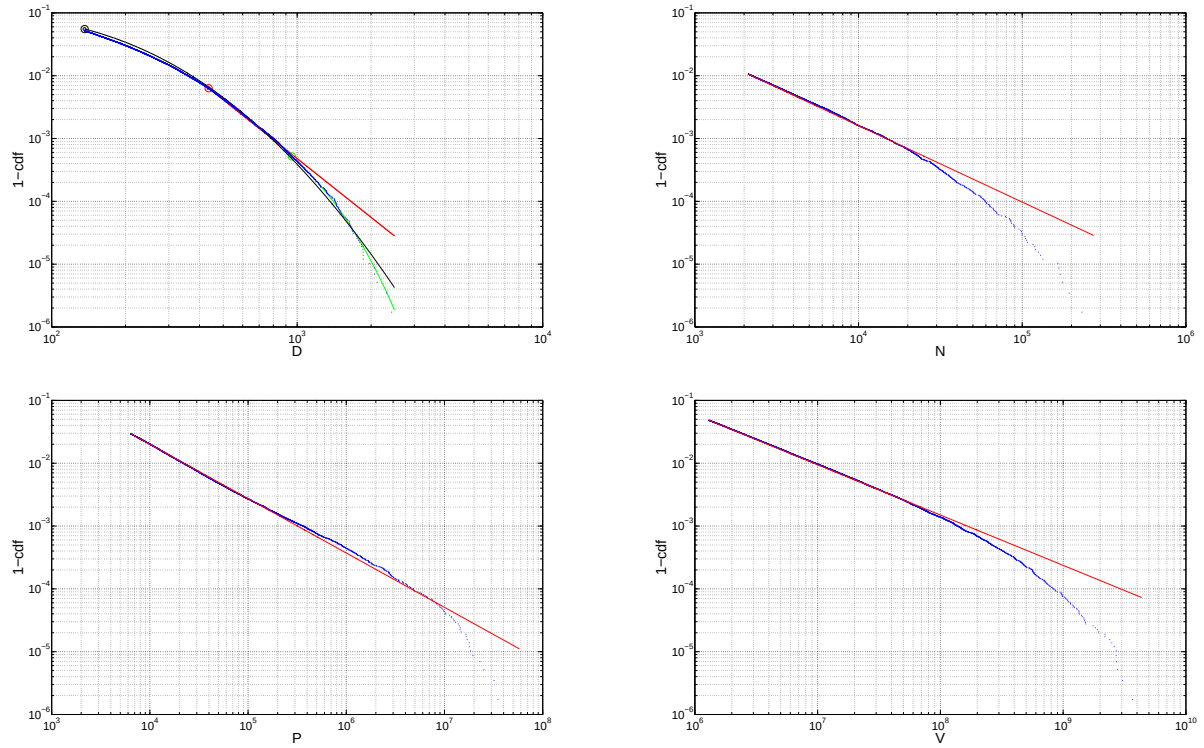

Figure 6: *Estimated tail distribution of the cross section across investors of connectedness, D (upper left), number of trades, N (upper right), profits, P (lower left), and volume, V (lower right). i) Connectedness is not well approximated by single distribution: For* $N_{cutoff} = 30,000$*, one-sided 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,* $\alpha = 0.80$*,* $N_{cutoff} = 28,020$*.*

The analysis in Gabaix, Gopikrishnan, Plerou, and Stanley (2003) also predicts (16), but clearly this relationship is not consistent with our empirical results, since the estimated tail exponent ζ_V^{CS} is less than one (it is 0.80), implying that δ is negative. This is the case over a large range of tail cutoff points, as can be seen from Table 14 (above $N^*_{cutoff} \approx 1,000$ the estimated tail exponent $\zeta_V^{CS} \leq 1$). A negative δ 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, Gopikrishnan, Plerou, and Stanley (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.[12]

[12]Formally, their equation (15) is not well defined when $\delta < 0$, since the integral is infinite.

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.

5 Concluding Remarks

Our study supports a view of the stock market as a place where information is incorporated into asset prices through the gradual diffusion in decentralized information networks. These networks provide an intermediate 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 the 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 the network's topological properties. It should not come as a surprise then that large asset price movements occur independently of public events, that investors take on diverse portfolio positions, and that they trade extensively.

How important is information diffusion for asset pricing? This question may be addressed by focusing in on specific time periods during which there were large stock movements, and studying the behavior of central versus peripheral investors during such time periods. Such an event focused study is outside of the scope of this paper, but we view it as a promising topic for future research.

Given our data, we have little to say about the true channels of information diffusion. A simple interpretation is that social networks lead to diffusion when individuals communicate information through their social network to friends and acquaintances. This is the view taken in several recent studies that focus on certain investor groups, e.g., in Hong, Kubik, and Stein (2004) and Cohen, Frazzini, and Malloy (2008). More generally, information diffusion may work through other channels, e.g., when one person overhears a discussion about stocks in the lunch canteen, when several people read the same local newspaper or attend the same professional networking events, in Internet discussion boards, etc. Which factors determine the market's real information network? Geography? Social networks? Local news channels? Other channels? Given datasets more with detailed information about investors in the market, further research may shed light on this question.

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.

Betermeier, S., T. Jansson, C. Parlour, and J. Walden (2011): "Hedging Labor Income Risk," Working paper, University of California at Berkeley.

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.

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.

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.

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.

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.

Hellwig, M. F. (1980): "On the aggregation of information in competitive markets," *Journal of Ecnomic 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.

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.

——— (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(2), 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(2), 258–267.

Odean, T. (1999): "Do investors trade too much?," *American Economic Review*, 89, 1279–1298.

Ozsoylev, H., and J. Walden (2010): "Asset Pricing in Large Information Networks," UC Berkeley.

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.

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.

	D-range	250-275	500-550	4,000-4,400
c	β_{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%
	β_{Ramsey}	0.026	0.053	0.049
	t_{Ramsey}	5.83	15.7	13.7
	$\Delta\mu_{Ramsey}$	0.62%	1.0%	0.69%
d	β_{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%
	β_{Ramsey}	-0.0085	-0.064	-0.044
	t_{Ramsey}	-0.22	-5.3	-3.0
	$\Delta\mu_{Ramsey}$	-0.02%	-0.34%	-0.12%
n	β_{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%
	β_{Ramsey}	0.073	0.0036	0.016
	t_{Ramsey}	5.45	4.5	> 20
	$\Delta\mu_{Ramsey}$	0.67%	0.36%	1.6%
v	β_{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%
	β_{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 11: Normalized profits, μ, for subsample of investor network with degree between $250-275$, $500-600$, and $4,000-4,400$. The window length is ΔT=30 minutes.

μ				
	c	d	n	v
β_{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 12: Normalized profits, μ, log-degree (d) log-centrality (c), log-number of trades (n) and log-volume (v) with threshold $M = 10$ and a time window of Δt =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.

Excluding d^2	c	d^2	d	n	v
β_{OLS}	0.00010		-0.0013	0.0070	-0.0018
t_{OLS}	15.9		-13.6	> 20	< -20
$\Delta\mu_{OLS}$	0.19%		-0.29%	1.3%	-0.50%
Including d^2	c	d^2	d	n	v
β_{OLS}	0.00002	0.0013	-0.0019	0.0073	-0.0018
t_{OLS}	1.93	9.3	-16.5	> 20	< -20
$\Delta\mu_{OLS}$	0.04%	0.25%	-0.43%	1.3%	-0.51%

Table 13: Normalized profits, μ, regressed on c, d, n, and v, and on c, d, d^2, n, and v. In both cases a 1-minute window length was used.

	N_{cutoff}		300	1,000	3,000	10,000	30,000
	N^*_{cutoff}	1,920					
C	ζ^{CS}_C	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
	N^*_{cutoff}	3,660					
D	ζ^{CS}_D	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^*_{cutoff}	6,140					
N	ζ^{CS}_N	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
	N^*_{cutoff}	17,090					
P	ζ^{CS}_P	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
	N^*_{cutoff}	28,020					
V	ζ^{CS}_V	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 14: Estimated tail-exponent, ζ^{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 α, given cut-off point.

C

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
	t	6.21	5.65	6.67	5.63	13.4	14.5
Exponential	lr	**-2.81E3**	-3.66E2	**-1.44E3**	**-4.51E3**	**-1.63E4**	**-8.27E4**
	t		2.5				17.7
Log-normal	lr	-2.82E3	**-3.65E2**	-1.43E3	-4.52E3	-1.63E4	-8.31E4
	t	4.14		0.95	5.73	6.22	

D

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

N

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

P

N_{cutoff}		N^*_{cutoff}	300	1,000	3,000	10,000	30,000
Pareto	lr	**-1.89E5**	**-4.72E3**	**-1.449E4**	**-3.94E4**	**-1.17E5**	**-4.94E5**
	t						
Exponential	lr	-2.13E5	-4.79E3	-1.50E4	-4.22E4	-1.30E5	-
	t	> 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
	t	> 20	0.07	1.78	8.68	19.6	13.8

V

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

Table 15: 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 $N_{cutoff} = 1,500$.