RICE UNIVERSITY

Measuring Information in Financial Markets

by

Edwin Hu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Jefferson Duarte, Chair
Gerald D. Hines Associate Professor of Real Estate Finance

Alexander W. Butler
Professor of Finance

Kerry E. Back
J. Howard Creekmore Professor of Finance

Robin C. Sickles
Reginald Henry Hargrove Chair of Economics

Houston, Texas
April, 2016
ABSTRACT

Measuring Information in Financial Markets

by

Edwin Hu

The measurement of information in financial markets is fundamental to our understanding of investor heterogeneity and the efficiency of stock prices.

In the first chapter, I introduce Weighted Trading Correlation Network (WTCN) estimation to uncover hidden information linkages between investors based on observed pairwise trading correlations. I use daily-level institutional manager trades to compute network snapshots. I show that WTCNs have two topological features characteristic of social networks—an approximate power-law degree distribution and positive assortative mixing. These results are consistent with information transfers occurring through social interactions.

In the second chapter, I develop Information Diffusion Centrality—a measure of network centrality based on the idea that investors exchange information bilaterally through one-on-one interactions in proportion to the strength of their connections to one another. I compare Information Diffusion Centrality to two measures of centrality based on different diffusion mechanisms in their ability to predict institutional trading performance. The first, Degree Centrality, supposes that investors are infor-
mationally linked, but there are no network flows. The second, Eigenvector Centrali-
ity, corresponds to a world where information spreads epidemically through simple
contagions. I show that only Information Diffusion Centrality predicts higher perfor-
ance, whereas Degree and Eigenvector Centrality predict zero or lower performance.
These results may explain why valuable information tends to remain localized—i.e.
why there is persistent information heterogeneity among investors—despite the fact
that social networks facilitate the rapid transmission of ideas.

In the third chapter, my co-authors and I examine whether the Easley and O’Hara
(1987) PIN model’s recently documented failure to identify private information arises
from the model’s inability to describe the data or from the model’s reliance on order
flows alone. We find that the PIN model mistakenly identifies private information
from turnover because it is unable to describe the order flow data. We propose a
model that addresses this shortcoming but also depends on order flow alone. We find
that the extended model does not perform as well as the Odders-White and Ready
(2008) model, which relies on both returns and order flow.
Contents

Abstract ii
List of Illustrations v
List of Tables vi

1 Network Analysis of Trading Correlations 1

1.1 Related Literature and Challenges 3

1.1.1 Computational Challenges 3
1.1.2 Network Dynamics 4
1.1.3 Network Formation 5
1.1.4 Data Availability 6

1.2 Weighted Trading Correlation Network Estimation 6

1.2.1 Weighted Networks 7
1.2.2 Estimating Weighted Trading Correlation Networks 8
1.2.3 Network Statistics 11

1.3 Data 12

1.4 Network Analysis 14

1.4.1 Network Summary Statistics 14
1.4.2 Degree Distribution 15
1.4.3 Clustering 18

1.5 Conclusion 19
2 Information Diffusion in Institutional Investor Networks 21

2.1 Background and Related Literature 28

2.2 Framework 29

2.2.1 Networks Primer 30

2.2.2 Estimating Weighted Trading Correlation Networks 32

2.2.3 Information Diffusion Centrality 32

2.2.4 Trading Performance 36

2.3 Data 39

2.3.1 Fund Summary Statistics 41

2.3.2 Event Summary Statistics 46

2.4 Centrality and Trading Performance 48

2.5 Evidence from Merger Announcements 53

2.6 Ruling Out Alternative Hypotheses 57

2.6.1 M&A Advisor “Tipping” 57

2.6.2 Expertise versus Information Diffusion 60

2.6.3 (No) Timing Sudden Director and Manager Deaths 61

2.7 Conclusion 65

3 What does the PIN model identify as private information? 67

3.1 Data 75

3.2 Why does PIN fail? 78

3.2.1 Description of the PIN model 78

3.2.2 How does the PIN model identify private information? 87

3.2.3 Explaining PIN anomalies with turnover 99

3.3 Does order flow alone reveal private information? 105
3.3.1 Extending the PIN model ........................................... 106
3.3.2 The OWR model ....................................................... 117
3.3.3 A horse race between EPIN and OWR models ................. 124
3.4 Conclusion ............................................................... 136

A WTCN and IDC ......................................................... 139
A.1 Information Sharing Matrix ........................................... 139
A.2 Eigenvector centralities ................................................. 140
A.3 ANCerno Data ............................................................ 142
A.4 Propensity Score Matching ............................................. 144

B What does the PIN model identify as private information? .... 146
B.1 The DY model ............................................................ 146
  B.1.1 The DY model ....................................................... 146
  B.1.2 Estimation of the DY model ..................................... 148
  B.1.3 $CPIE_{DY}$ .......................................................... 150
  B.1.4 How does the DY model identify private information? ... 150
B.2 Estimating Order Flow, $r_{o,i,t}$ and $r_{d,i,t}$ ....................... 153
B.3 Details about the PIN model ........................................... 154
  B.3.1 PIN Likelihood ...................................................... 154
  B.3.2 Maximum likelihood procedure .................................. 155
  B.3.3 Computing $CPIE_{PIN}$ ............................................. 155
  B.3.4 $CPIE_{PIN}$ of M&A targets around announcements .......... 156
B.4 Details about the EPIN model .......................................... 158
  B.4.1 Negative binomial distribution in EPIN model ............... 158
B.5 Details about the OWR model .................................................. 158
  B.5.1 OWR Likelihood .............................................................. 158
  B.5.2 How does the OWR model identify private information? . . . 160
B.6 Tables and Figures .............................................................. 162
Illustrations

1.1 Triangle Network. ......................................................... 8
1.2 Weighted Trading Correlation Network with an active
    liquidity trader. .............................................. 10
1.3 Degree Histogram. ..................................................... 15
1.4 Degree Distribution. .................................................. 17
1.5 Degree-Dependent Clustering. ....................................... 19

2.1 Random Walks on a Network. ........................................ 31
2.2 Information Diffusion vs. Eigenvector Centrality. ............. 37
2.3 Centrality 1999Q1–2011Q3. ......................................... 45
2.4 Weighted Trading Correlation Network (2002Q2). ............. 47

3.1 PIN Tree. ................................................................ 81
3.2 PIN Parameters. ......................................................... 86
3.3 XOM EO. ................................................................ 93
3.4 Breakdown of the PIN Model. ....................................... 97
3.5 Earnings Announcements - PIN. .................................... 101
3.6 EPIN Tree. ............................................................... 108
3.7 XOM EPIN. ............................................................. 113
3.8 OWR Tree. ............................................................... 119
3.9 Earnings Announcements .............................................. 126
3.10 Earnings Announcements - EPIN Decomposition ......... 128
3.11 Earnings Announcements - OWR Decomposition ........ 130
3.12 Opportunistic Insider Trades ...................................... 133

B.1 DY Tree ................................................................. 168
B.2 XOM DY ................................................................. 169
B.3 Breakdown of the DY Model ........................................ 169
B.4 M&A Targets - PIN .................................................... 170
Tables

1.1 Network Summary Statistics. ........................................ 14

2.1 Fund Summary Statistics 1999Q1–2011Q3. ......................... 43
2.2 Event Summary Statistics. ........................................... 49
2.3 Interim Trading Performance and Centrality. ..................... 52
2.4 Round-trip Trading around M&A Announcements — Targets. 56
2.5 Tipping around M&A Announcements — Targets. ............... 59
2.6 Round-trip Trading around New Product Announcements. ... 62
2.7 Round-trip Trading around Sudden Director & Manager
   Deaths. ........................................................................ 64

3.1 Summary Statistics. ....................................................... 79
3.2 PIN Parameter Estimates. ............................................... 85
3.3 PIN Model Regressions. .................................................. 89
   PIN Model Regressions. Continued. .................................. 90
3.4 PIN Regressions Around Earnings Announcements. .......... 103
3.5 EPIN Parameter Estimates. ............................................ 111
3.6 EPIN Model Regressions. ............................................... 115
   EPIN Model Regressions. Continued. ............................... 116
3.7 OWR Parameter Estimates. ............................................. 123
3.8 Return Reversals. ................................................. 136

A.1 ANCerno Data (1999Q1–2012Q1). ........................... 144
A.2 Normalized Differences. ........................................ 145

B.1 DY Estimates. ...................................................... 163
B.2 DY Model Regressions. .......................................... 164
B.2 DY Model Regressions. Continued. ....................... 165
B.3 OWR Model Regressions. ................................. 166
B.3 OWR Model Regressions. Continued. ................... 167
Chapter 1

Network Analysis of Trading Correlations

Recent theory suggests that investor heterogeneity is a natural consequence of the underlying information network structure.¹ The information networks literature is based on the idea that investors have overlapping information sets—i.e. that they are “informationally linked” (e.g. Colla and Mele, 2010). However, information linkages are unobserved, and must be estimated from trading data. The basic consequence of information linkages is that traders who are informationally “linked” will have a tendency to trade the same stocks in the same direction on the same days. This paper describes Weighted Trading Correlation Network (WTCN) estimation as a way to empirically proxy for the underlying information network.

The primary contribution of this paper—and indeed a key motivation for the method—is to introduce a weighted network approach to the literature which may be used to study network flows. In sociology and computer science it is well-known that information diffuses through “weak ties” (Granovetter, 1973; Bakshy, Rosenn, Marlow, and Adamic, 2012). However, the strength of information linkages has received little attention in economics and finance. Most network analysis focuses on un-weighted networks (e.g. Ozsoylev, Walden, Yavuz, and Bildik (2014)) which as-

¹For example: Colla and Mele (2010), Ozsoylev and Walden (2011), Han and Yang (2013), and Walden (2014)
sumes that linkages are either present or absent, and ignores the strength of the connections. Trading correlation edge-weights are a natural way of measuring the “strength” of the information linkages between investors or the degree to which investors’ information sets overlap.

I compute WTCNs using daily trading data from institutional managers from 1999–2011. I verify that WTCNs are different from randomly generated networks in two ways. First, WTCNs have a power-law degree-distribution, characterized by a heavy right tail, whereas randomly generated networks have Poisson degree distribution which do not have a heavy right tail. Hence the degree distribution of WTCNs are consistent with the scaling laws of social and economic networks.\(^2\) Second, WTCNs are highly clustered and demonstrate positive assortative mixing—a feature found only in social networks (Newman and Park, 2003). In contrast, randomly generated networks have low—asymptotically zero—clustering. Together, these results suggest that WTCNs capture a genuine information network, which constitutes the second contribution of this paper.

The remainder of this paper is as follows. Section 2.1 reviews the challenges facing the small but growing information and trading networks literatures. Section 2.2 describes some network preliminaries and the WTCN estimation. Section 2.3 describes the institutional trading data. Section 2.4 conducts network analysis and simulation of WTCNs. Section 2.5 concludes.

\(^2\)See for example, Jackson and Rogers (2007), Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012), Barabási and Albert (1999), and Gabaix (2009).
1.1 Related Literature and Challenges

The theory and econometrics of networks in economics and finance is still in its infancy (Imbens, 2010). As such, there are many limitations and open questions. In this section I highlight four major challenges: estimation, dynamics, formation, and data.

1.1.1 Computational Challenges

The most salient challenge in the estimation of information networks is computational. In most real-world markets there can be upwards of thousands or millions of investors trading at any given time. The number of possible linkages grows at \( O(N^2) \).\(^3\) The rapid growth in complexity creates a unique challenge for the literature. As Gomez-Rodriguez, Leskovec, and Krause (2012) points out, even with a few thousand nodes, exact maximum-likelihood estimation is not feasible and approximations are necessary. Ozsoylev, Walden, Yavuz, and Bildik (2014) develops a simple and scalable approximate estimator of un-weighted Empirical Investor Networks (EINs) and estimates information networks using overlapping time-stamped trades on the Istanbul Stock Exchange for all (mostly individual) participants. For the empirical application of WTCNs, this paper estimates weighted networks using daily pairwise trading correlations of institutional investors. The correlation approach to weighted network estimation is inspired by the computational biology literature, where similar approaches have found widespread success in estimating large gene co-expression \( ^3N(N-1)/2 \) assuming a symmetric adjacency matrix.
networks (See Horvath, 2011).

1.1.2 Network Dynamics

A second challenge is in understanding network dynamics. One of the primary motivations behind a using network approach to study information flows is that—in theory—in information flows “often in predictable paths, along the network” (Cohen, Frazzini, and Malloy, 2008). The basic intuition is that information “shocks” arrive to the network and “propagate” along information linkages to “central” investors. Network centrality embeds network flows and network structure such that, given a fixed network, central investors should have superior access to information compared to peripheral investors. Information asymmetry and heterogeneity arise naturally in the network framework. Despite the simple intuition, there is very little theoretical or empirical guidance on information flows through networks. This section briefly surveys two mechanisms found in the literature.

The first mechanism of information diffusion supposes that information spreads quickly in an “epidemic” (one-to-many) fashion. Walden (2014) derives trading profits as a function of a measure of centrality closely related to the well-known Katz centrality measure. Katz centrality, as a member of the eigenvector family of centrality measures, has the natural interpretation of shock propagation in an epidemic or contagion model. For example, eigenvector centrality is used in production and banking networks to summarize how small shocks are propagated and amplified in economic networks (Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi, 2012; Acemoglu, Ozdaglar, and Tahbaz-Salehi, 2015). In an epidemic model information propagates
quickly through “central hubs” to the whole network in a similar manner. Ozsoylev, Walden, Yavuz, and Bildik (2014) and Rantala (2015) find evidence consistent with the epidemic diffusion of information among individual investors.

The second mechanism of information diffusion supposes that information percolates slowly via bilateral (one-to-one) information exchanges. Stein (2008) and Cujean (2016) model bilateral information exchanges between institutional investors and show that the key feature of this alternative mechanism is that information tends to “cluster” and does not spread far. Hu (2015) develops a measure of centrality based on the intuition of bilateral information exchanges and shows that it predicts trading performance for central institutional investors. Moreover, eigenvector centrality fails to predict institutional trading performance. Taken together these results suggest that the mechanisms of information diffusion for institutions are different from those of individuals.

1.1.3 Network Formation

A third challenge is in understanding how and why networks form. On the theory side there is little discussion of endogenous networking apart from Stein (2008), which proposes a dynamic game with incentive-compatible “conversations.” Other models including Colla and Mele (2010), Ozsoylev and Walden (2011), Han and Yang

---

4One interesting similarity is in the assumptions about price impact and incentive compatibility. Walden (2014), Ozsoylev and Walden (2011), and most of the information networks literature assume that investors are "small" and have no price impact and assume that information is shared exogenously. Cujean (2016) models information percolation among fund managers in a search style model where investors never meet twice and therefore cannot benefit from lying, but still assumes small investors and exogenous information sharing as in Duffie and Manso (2007).
(2013), and Walden (2014) take information networks as given and discuss the trading and asset pricing implications. On the empirical side papers that study network formation using econometric choice models typically study relatively small networks, and require detailed data on participants (nodes) and relationships (edges). While information network formation is a very interesting economic problem, it is outside the scope of this paper given the data and computational limitations.

1.1.4 Data Availability

The fourth challenge is in data availability. The information and trading networks literature is currently limited by database availability. To my knowledge, there are only a handful of datasets containing actual investor trades, and almost zero datasets available to researchers containing investor identifiers. However, research in the near-future may be able to use newly constructed audit-trail data to test and develop new theories. For example, SEC Rule 613 proposed in 2010, requires the creation of a comprehensive consolidated audit trail database to record all trading activity—from generation through routing, modification, cancellation, or execution—in US National Market Securities. Title VII of the Dodd-Frank Act provides a comprehensive framework for the regulation of over-the-counter swaps by the SEC and the CFTC, which requires data on security-based swaps and non-security-based swaps (e.g. energy, agriculture, and other “commodity” swaps) respectively. For example, Loon and

\[\text{Footnote 5}:\] Hochberg, Lindsey, and Westerfield (2015) study VC co-investment networks in a Tobit framework. In their setting there are roughly 1,000 participants per year and the connections are well defined. Chandrasekhar (2015) surveys the literature on network formation using choice models.
Zhong (2016) examines the effect of Dodd-Frank on OTC costs and liquidity using real-time CDS trades from a CFTC-registered Swap Data Repository.

1.2 Weighted Trading Correlation Network Estimation

I begin with a brief introduction to networks in Section 2.2.4 before describing the empirical estimation of Weighted Trading Correlation Networks in Section 2.2.1. I then describe two key network statistics: degree centrality, and clustering in Section 2.2.1.

1.2.1 Weighted Networks

Most network analysis focuses on collections of nodes joined by edges in a binary fashion. These networks are typically represented by (0,1) “adjacency” matrices:

\[
A_{ij} = \begin{cases} 
1 & \text{if } i \text{ is connected to } j \\
0 & \text{otherwise}
\end{cases}
\]  

(1.1)

However, many settings, especially social and economic ones, include inherently “stronger” links and “weaker” links (Granovetter, 1973). Both strong and weak links are important, and focusing on only binary adjacency matrices ignores potentially important data. Therefore, we can also represent networks in terms of “weighted” adjacency matrices, \(A_{ij} = w_{ij}\), where \(w_{ij}\) is the weight or strength of the connection between \(i\) and \(j\).

For example, we can think of a simple triangle network with three nodes A, B,
and C represented in geometric and adjacency matrix form below in Figure 1.1. The connection between A and B is strongest with a weight of 1, followed by the link between B and C with a weight of 0.5. The connection between A and C has the lowest weight of 0.25. In Figure 1.1 Panel A (below), edge thickness corresponds to the strength of the connections. The same data is encoded in the off-diagonal terms of the adjacency matrix in Panel B.

Figure 1.1: **Triangle Network.** A network is defined as a collection of nodes (●) and edges (─). Panel A shows the geometric representation of the network and Panel B shows the (weighted) adjacency matrix representation. The adjacency matrix is symmetric, and the edge weights are on the off-diagonal entries.

(a) Geometric Representation

(b) Adjacency Matrix

\[
A = \begin{bmatrix}
0 & 1 & 0.25 \\
1 & 0 & 0.5 \\
0.25 & 0.5 & 0
\end{bmatrix}
\]

1.2.2 Estimating Weighted Trading Correlation Networks

We often do not know the strength of the linkages between investors and hence the adjacency matrix must be estimated from trading data. I estimate the strength of the connections using pairwise trading correlations. The intuition is that investors with more overlapping information should have stronger links and higher trading correlations.
An investor’s buying activity of $n$ stocks over $T$ periods can be represented by a $(0,1)$ $nT$-vector stacked by stock, $\mathbf{b}_i = [b_{i1}, b_{i2}, b_{i3}, \ldots, b_{in}]'$. $\mathbf{b}_i$ is a $T$-vector which takes on values of one if if investor $i$ bought stock $n$ in period $t$ and is zero otherwise. The buy correlation between investor $i$ and $j$ is the Pearson correlation, $\rho_{ij}^b = \frac{\text{cov}(\mathbf{b}_i, \mathbf{b}_j)}{\sigma_{\mathbf{b}_i} \sigma_{\mathbf{b}_j}}$, and the sell correlation, $\rho_{ij}^s$, is defined analogously for the vector of selling activity.

The edge weight between investors is computed by taking the average of buy and sell trading correlations: $\hat{w}_{ij} = \frac{\rho_{ij}^b + \rho_{ij}^s}{2}$. I call the weighted adjacency matrix with elements $\hat{A}_{ij} \in [0,1]$ the Weighted Trading Correlation Network (WTCN).\(^6\) The binary (un-weighted) version of the WTCN has the interpretation where investors are linked if they have non-negative trading correlations. The minimum correlation threshold of zero is tunable, drawing a parallel to the Ozsoylev, Walden, Yavuz, and Bildik (2014) EIN parameter $M$ which represents the minimum number of overlapping trades needed to define a connection.

Weighting based on trading correlations has the added benefit of mitigating the influence of active liquidity traders on network estimates. For example, a related un-weighted procedure, Empirical Investor Networks (Ozsoylev, Walden, Yavuz, and Bildik, 2014), assigns links in a binary fashion based on the presence of overlapping trades. If two investors trade the same stock on the same side in the same period then they are connected. However, in the presence of active liquidity traders this methodology tends to identify many non-information linkages and can assign undue

---

\(^6\)Weighted networks based on pairwise correlations are common in the genomics and systems biology literature (see, Horvath, 2011, for example).
centrality to un-informed investors.

As a hypothetical example suppose the true information network is the star network in Figure 1.2 (below) of four investors who all share information with one central investor. The five informed investors are represented by dark nodes (●). Now suppose that these five investors only trade based on shared information. As a result, they will have perfectly correlated trades as indicated by the dark lines connecting them. There is also one very active liquidity trader (○) who does not have or share any information but trades frequently. Because the liquidity trader has overlapping trades with all five informed investors we observe that the liquidity trader is part of the network. However, because the liquidity trader trades in periods when the informed traders are not trading, the overall trading correlation with the informed investors is low, and the liquidity trader’s centrality is also low. If we ignore the trading correlations and connect investors based only on the existence of overlapping trades the liquidity trader appears to be most well-connected, with five connections when the bona fide central informed investor has only four connections. Trading correlation weighting correctly relegates active liquidity traders to the periphery of information networks.

1.2.3 Network Statistics

Degree Centrality and Power Laws

The simplest measure of network centrality is Degree Centrality, which is defined as the sum of a node’s direct connections. The Degree Centrality for node $i$ is defined
Figure 1.2: **Weighted Trading Correlation Network with an active liquidity trader.** The “true” star network consists of five informed investors (●) connected to one another via black lines (→). Adding an active liquidity trader (○) results in many un-informative linkages to the informed investors. Using Weighted Trading Correlation Network estimation attenuates the strength of the un-informed connections (→). Node size corresponds to the sum of the weighted edges such that the liquidity trader is “unimportant” and on the periphery.

as the row sum of the adjacency matrix $A$:

$$d_i = \sum_j A_{ij}. \quad (1.2)$$

The definition of degree centrality does not depend on whether the edges (the elements of $A$) are weighted or un-weighted.

Degree centrality one of the key network statistics used to describe the topological structure of a network. Most real-world networks are known to have power-law degree distributions (Barabási and Albert, 1999), where the proportion of nodes with a degree $k$ higher than a large $x$ is proportional to $1/x^\gamma$ where $\gamma$ is the power-law
exponent:

\[ P(k > x) = \frac{c}{x^\gamma} \]  

(1.3)

for some constant \( c \). Power-law distributions (e.g. the Pareto distribution) have many unique properties, one of which is a characteristic heavy right tail which distinguishes them from the more familiar exponential family of distributions (e.g. the Normal distribution). Power-law distributions appear in many interesting settings in economics and finance (e.g. income and wealth inequality (the Pareto principle), the size distributions of cities and firms (Zipf’s law), etc., see Gabaix, 2009, for a survey). Gabaix, Gopikrishnan, Plerou, and Stanley (2006) argues that the power law in the distribution of institutional investors—i.e. the presence of large traders—can explain the large fluctuations in volume, volatility, and returns. The forces behind and implications of power-laws in social and economic networks are still largely unexplored.

**Clustering**

Another important property of real-world networks is known as clustering. Clustering in the context of networks is defined in terms of triangles. A triangle is formed if a triplet of nodes \( t, u, \) and \( v \) is closed, such that \( t \) is connected to \( u \) and \( v \), and \( u \) is also connected to \( v \).

The clustering coefficient for node \( i \) is computed as: \( c_i = \frac{T(i)}{k_i(k_i - 1)} \) which is the fraction of observed triangles to the fraction of possible triangles. In weighted networks
the clustering can be computed as:

$$c_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k}(\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}$$

(1.4)

where $\hat{w}$ is the edge weight normalized by the maximum edge weight. It is clear that the weighted version of clustering is equivalent to the un-weighted version when the weights are replaced with binary adjacency matrix elements.

### 1.3 Data

To estimate WTCNs I use daily trading data from institutional managers from a proprietary database provided by ANcerno Ltd. (a.k.a. Abel/Noser Solutions Ltd.). ANcerno records the date, direction (buy or sell), and shares of all trades made by their clients from 1999Q1 to 2011Q3. My sample stops in 2011Q3 because ANcerno removed identifiers in 2012, preventing me from tracking subsequent trading activity. Furthermore, 2011Q4 and 2012Q1 have only one-third of the number of funds compared to 2011Q3 which results in unreliable network estimates.

I restrict my sample to all trades made in the US, in US currency, of common stock listed on the NYSE, AMEX, and NASDAQ exchanges, for which I am able to match the ANcerno provided point-in-time CUSIP to a CRSP PERMNO. The 799 managers in my sample trade a total of 8,555 unique common stocks. Overall, the investors make 141.85 million trades of 1.11 trillion shares valued at USD 34.26 trillion dollars. Puckett and Yan (2011) estimate that the ANcerno institutions account for approximately 8% of the total dollar value of CRSP trading volume.
between 1999 and 2005.

Because ANcerno timestamps are incomplete (see Anand, Irvine, Puckett, and Venkataraman, 2013) I estimate Weighted Trading Correlation Networks based on daily-level trading activity. To mitigate concerns that overlapping trades are coincidental, I restrict my sample of overlapping trades to “time-sensitive” overlapping trades. To identify time-sensitive trades, I exploit the fact that ANcerno tracks the number of days over which a stock was traded as part of an order ticket. I define time-sensitive trades as trades of stocks executed within a single day. If a fund manager receives short-lived information he or she would have a strong incentive to trade immediately, versus splitting up the order execution over multiple days, consequently sacrificing price impact for immediacy. On average, time-sensitive trades make up 62% of order executions. The resulting quarterly network snapshots contain on average 356 managers, with a minimum of 312 in 2011Q3 and a maximum of 392 in 2002Q1.

1.4 Network Analysis

In this section I show that WTCNs demonstrate two topological features characteristic of social networks. In addition, I show that throwing away the data contained in the edge weights can lead to very different network statistics.
1.4.1 Network Summary Statistics

Table 1.1 summarizes the degree and clustering for the 18,295 manager-quarter observations in the sample from 1999Q1–2011Q3.

Table 1.1: **Network Summary Statistics.** This table summarizes the weighted and un-weighted degree and clustering for each of the 19,295 manager-quarter observations in the WTCN snapshots from 1999Q1–2011Q3. Weighted estimates are based on WTCN estimation, and un-weighted estimates are based on the [0,1] binary “projections” described in Section 2.2.1. Degree and clustering are described in Section 2.2.1.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>18,295</td>
<td>95.98</td>
<td>70.06</td>
<td>40.00</td>
<td>78.00</td>
<td>137.00</td>
<td>0.86</td>
<td>-0.13</td>
</tr>
<tr>
<td>k_w</td>
<td>18,295</td>
<td>3.12</td>
<td>1.40</td>
<td>2.26</td>
<td>2.65</td>
<td>3.36</td>
<td>2.73</td>
<td>9.38</td>
</tr>
<tr>
<td>c</td>
<td>18,295</td>
<td>0.62</td>
<td>0.19</td>
<td>0.48</td>
<td>0.62</td>
<td>0.77</td>
<td>-0.26</td>
<td>-0.24</td>
</tr>
<tr>
<td>c_w</td>
<td>18,295</td>
<td>0.006</td>
<td>0.003</td>
<td>0.003</td>
<td>0.005</td>
<td>0.008</td>
<td>2.236</td>
<td>47.878</td>
</tr>
</tbody>
</table>

The average un-weighted degree $k$ is 95, and is somewhat right skewed with a negative kurtosis. The average weighted degree $k_w$ is lower at 3.12. The weighted degree is right skewed with high kurtosis, which is typical of many real-world networks. The average clustering $c$ is 62%, and the average weighted clustering is 60 basis points. Weighted degrees and clustering are lower because edge weights take continuous values between [0,1], whereas the un-weighted version treats all edges as 1.

1.4.2 Degree Distribution

The presence of a power-law in the degree distribution is a good test of whether the estimated networks are genuine information networks or just “noise.” It is well-known
that in random networks where edges are formed independently—i.e. in the Erdős and Rényi (1959) (ER) model—the degree distribution is Poisson, which by nature of being in the exponential family of distributions, displays no heavy-tail.\textsuperscript{7} Figure 1.3 Panels A and B plot histograms of the degree distributions for the quarterly WTCNs snapshots from 1999Q1–2011Q3 in un-weighted and weighted forms.

\textbf{Figure 1.3 : Degree Histogram.}

At a first glance both types of degree distributions display a long right tail, which may be indicative of a power law.

To see whether the estimated networks follow a power law I conduct three tests. First I plot cumulative histogram or rank/frequency versions of the same data on a log-log scale following Newman (2005) and Gabaix (2009). In a log-log cumulative histogram plot power law distributions appear as straight lines where the slope

\textsuperscript{7}Asymptotically, assuming a constant expected degree. Otherwise in a finite network the degree distribution is binomially distributed.
corresponds to the power law exponent ($-\gamma$). In contrast, exponential family distributions appear as downward curving lines. Therefore a simple way to detect a power law is to plot the actual data against fitted power-law and exponential distributions. Second, I compute the Kolmogorov-Smirnov distance $D$, which measures the maximum vertical distance between the cumulative empirical degree distribution and the fitted power-law distribution. A distance $D = 0$ indicates a perfect fit, and larger values indicate poor fit. Third, I compute the likelihood ratio $R$ between the power law fit and the exponential fit. A positive likelihood ratio indicates that the empirical distribution is closer to a power-law than an exponential, and a negative ratio indicates the opposite. Figure 1.4 presents the results of the three tests.

Figure 1.4: **Degree Distribution.** This figure plots the degree distribution on a log-log scale. The solid blue line represents the empirical degree distribution. The dotted blue (green) line represents the maximum-likelihood fitted power-law (exponential) distribution. $\gamma$ is the estimated power law tail parameter. $D$ is the Kolmogorov-Smirnov Distance between the empirical and power-law distribution. $R$ is the likelihood ratio between the power-law and exponential distributions. Degree estimates are pooled from quarterly network snapshots based on trading data from 1999Q1–2011Q3.

(a) Unweighted

(b) Weighted
The empirical degree distributions are plotted in blue. In each panel I fit the degree distribution to a power-law distribution using maximum likelihood and plot the fitted distribution as the dashed blue line. The fitted exponential distributions are represented by the dotted green lines. In both panels the empirical distribution is a downward curving line which is typical of finite sample network estimates. However, in Panel A, corresponding to the unweighted degree estimates, it is clear that the empirical cumulative density does not follow a straight line, and is more curved than even the exponential distribution. The weighted degree estimates in Panel B fall somewhere between the two distributions. The Kolmogorov-Smirnov distance is $D = 0.122$ for the unweighted degree estimates, which is nearly ten times larger than that of the weighted degree estimates, $D = 0.014$. The likelihood ratio for the unweighted degree estimates is large and negative ($R = -44$), which clearly favors the exponential distribution. The likelihood ratio for the weighted degree estimates is large and positive ($R = 8$), which favors the power-law distribution.

Taken together, these tests provide evidence that Weighted Trading Correlation Networks are characterized by a power-law degree distribution, consistent with a genuine information network. Furthermore, throwing away the data contained in the edge-weights results in a network topology that is indistinguishable from noise.

### 1.4.3 Clustering

Dorogovtsev (2004) describes a class of “correlated” networks where clustering is a function of degree ($c(k)$). In ER random graphs—which belong to the broader class
of “uncorrelated” networks—the clustering coefficient for each node is a constant.\(^8\) Most real-world networks demonstrate degree-dependent clustering, reflecting some underlying correlation structure. Figure 1.5 plots the average clustering as a function of the degree for both unweighted and weighted network estimates.

Figure 1.5: **Degree-Dependent Clustering.** This figure plots the average clustering coefficient for nodes with degree \(k\). Clustering and degree estimates are pooled from quarterly network snapshots based on trading data from 1999Q1–2011Q3.

(a) Unweighted

(b) Weighted

---

Figure 1.5 Panel A shows a hump-shaped relationship in the degree dependent clustering. Panel B shows a monotonic relationship between clustering and degree for the weighted network estimates. The interpretation is that nodes with a large number of connections tend to cluster with one another. This phenomenon is called assortative mixing, and is a feature unique to social networks (Newman and Park, 2003). For example, in board networks “big-shot” directors who sit on many boards

---

\(^8\)The "configuration model" of Bender and Canfield (1978) is a random graph model with a given degree sequence (e.g. power-law) that is uncorrelated.
tend to sit on boards with other “big-shot” directors (Davis and Greve, 1997). These results suggest that information networks may be embedded within social networks, consistent with information transfers through social interactions (e.g. Cohen, Frazzini, and Malloy, 2008, 2010; Pool, Stoffman, and Yonker, 2015).

1.5 Conclusion

This paper develops Weighted Trading Correlation Network estimation. The topological properties of Weighted Trading Correlation Networks based on daily institutional manager trades—a power-law degree distribution and positive assortative mixing—are consistent with “small-world” social networks and theoretical descriptions of information networks (e.g. Walden, 2014). These results suggest a deeper connection between social and information networks, which would be consistent with information transfers through social interactions (Cohen, Frazzini, and Malloy, 2008, 2010; Pool, Stoffman, and Yonker, 2015). Moreover, the structure of the network means that there are many short chains between investors, which in theory facilitate the rapid diffusion of information (Kleinberg, 2000).
Chapter 2

Information Diffusion in Institutional Investor Networks

Many authors in sociology and economics argue that the way in which people communicate affects the diffusion of information (e.g. Granovetter, 1973, 1985, 2005; Ellison and Fudenberg, 1995; Shiller, 1995). Moreover, a recent theoretical literature demonstrates that the structure of social or information networks shapes the efficiency of financial markets.\(^1\) Yet there is surprisingly little evidence that investors share valuable information with one another. One important exception, Shiller and Pound (1989) finds in a survey of 30 institutional investors that the majority purchased stocks based on conversations with other investment professionals. The result is surprising because, as Stein (2008) points out, institutional investors have strong incentives to compete for—rather than share—information.\(^2\) Given the implications for stock market efficiency, as well as for our understanding of investor behavior, it is fair to say that the empirical case for information sharing is underdeveloped.

In this paper I examine the hypothesis that institutional investors share valu-

\(^1\)Colla and Mele (2010), Ozsoylev and Walden (2011), Han and Yang (2013), and Walden (2014)

\(^2\)As for why informed investors would share information, Stein (2008) argues that in a repeated game it is optimal to share truthfully so that new information may be produced and shared in return. Also, it may be valuable for a fund manager to be able take a position and credibly communicate this to other managers so that their subsequent trades will move the stock price in the right direction.
able information *through the grapevine*—e.g., via word-of-mouth communication—by studying network flows and trading performance using more than a decade of order-level institutional trades. My empirical approach exploits the fact that if investors share valuable information, then network information flows generate predictable information asymmetry between “central” and “peripheral” investors which should be reflected in their trading performance. To provide large-sample evidence that investors share valuable information, I proceed in three steps.

First, I estimate the structure of information networks connecting investors to one another based on pairwise trading correlations. As a motivating example, we can think of investors sharing information through an online messaging platform. If two investors frequently chat with one another and share information about specific stocks, then we should expect their trades to have a positive contemporaneous correlation. A *Weighted Trading Correlation Network*, represented by nodes (investors) and edges (trading correlations), is an estimate of the underlying network of conversations between investors.\(^3\)

Second, I develop a new measure, *Information Diffusion Centrality*, to measure the expected information advantage provided by an investor’s position in a network. Information Diffusion Centrality is based on a simple model interweaving the dynamics of bilateral information sharing with network structure. The main prediction

\(^3\) Instant Bloomberg, founded in 1982, is often described as an private network for investors to share and develop investment research ideas—predating current online social network platforms and even the widespread use of e-mail (see Edgecliffe-Johnson, Andrew; Alloway, Tracy; and Philip Stafford. “Instant Bloomberg not going to fade away yet” *Financial Times* 21 May 2013). In addition, Saavedra, Hagerty, and Uzzi (2011) shows that simultaneous trading corresponds with instant messaging activity of 66 day traders.
of the model is that central investors have a higher probability of receiving novel information earlier, whereas peripheral investors tend to have less valuable commonly-held information. Intuitively, when trading opportunities are short-lived, information heard second-hand is more valuable than information heard third-hand—especially when other investors already have the information.

Third, I measure trading performance based on actual trades, since an investor’s information advantage is reflected in the optimal timing of purchases and sales of individual stocks. Implied trading performance based on quarterly changes in holdings is more common in the finance literature, but using changes in holdings assumes that all trades occur at the end of the quarter, which ignores the timing of the majority of intraquarter trades. In my main analysis, I compute quarterly interim trading performance using all trades made by each investor in a given quarter following Puckett and Yan (2011).

I lack a good instrument for centrality that would allow me to cleanly identify the effects on trading performance. As a result, I rely on predictive regressions to measure the expected information advantage from being central. Centrality predicts subsequent trading performance if it measures an investor’s information advantage, and equally importantly, if an investor’s position in the network is stable over time. Stein (2008) shows that if information sharing is mutually beneficial, then there exists a sustainable “conversational equilibrium.” I verify that my network estimates are highly persistent, which is consistent with such an equilibrium. I also take considerable care to account for the fact that centrality is not randomly assigned by propensity-score matching central investors to peripheral investors.
While my main analysis is most naturally motivated by access to information, it does not rule out the possibility that central investors’ information advantage is due to superior information processing. To address this interpretation, I examine a setting in which access to information is directly related to trading performance—merger and acquisition announcements. For example, Ahern (2015) examines 239 instances of illegal insider trading around mergers and acquisitions and finds evidence that individuals in “tipping chains” located closer to the original source of information earn higher returns. Specifically, I examine round-trip trading performance—e.g. buys and subsequent sells—in target stock around merger announcements to verify that centrality measures access to information. Because not all investors trade around every event, I account for potential sample selection issues using a Heckman (1976) selection model.

I conduct three additional empirical tests to address concerns that centrality proxies for common sources of information, product market expertise, or spuriously predicts trading performance.

First, Jegadeesh and Tang (2011) find that institutions connected to brokerage houses that serve as merger advisors are net buyers of target stocks and trade profitably ahead of merger announcements. I name-match brokers to merger advisors and estimate investors’ connections to advisors based on the share volume executed by each brokerage house in the previous quarter. I use merger advisor connections to control for common sources of information regarding merger announcements.

---

4Round-trip trading is consistent with investors trading on short-lived information—investors may want to reverse their positions in order to lock in gains or take advantage of overreactions (Hirshleifer, Subrahmanyam, and Titman, 1994; Brunnermeier, 2005).
Second, I examine round-trip trading around new product announcements where industry or product market expertise is likely to predict superior trading performance (e.g. Kacperczyk, Sialm, and Zheng, 2005). I construct a measure of investors’ expertise in announcing firms’ products based on the concentration of their trades in the previous year in the stocks of firms with similar product descriptions. I use the concentration of trades to control for product market expertise.

Third, as a falsification test, I examine sudden deaths of board directors and key executives to see if centrality spuriously predicts trading performance when there is no information. Sudden deaths are used as unpredictable shocks to corporate governance (Nguyen and Nielsen, 2010), hence there should be no returns to being central in this setting.

My results are as follows. In predictive regressions of interim trading performance on Information Diffusion Centrality, I find that central investors—those with above median Information Diffusion Centrality—have 0.32% (0.48%) higher principal-weighted (equal-weighted) average abnormal trading performance in the following quarter. To put this estimate in perspective, the median investor’s principal-weighted (equal-weighted) average trading performance in my sample from 1999Q1 to 2011Q3 is 0.35% (0.47%). Becoming central would nearly double the median investor’s expected trading performance. The results hold when controlling for measures of trad-

---


6 I identify sudden deaths by searching through news article headlines for keywords such as death of, demise of, and passing of, and using the article text I filter out deaths of individuals over the age of 65, deaths related to cancer, suicide, disease, illness, or complications related to prior conditions.
ing activity and past performance, and are robust in a propensity score matched sample of central and peripheral investors.

Central investors also trade profitably ahead of merger announcements. Central investors have 1.50% (1.53%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. To put this estimate in perspective, the median investor has a 1.27% (1.24%) principal-weighted (equal-weighted) trading performance around merger announcements. Becoming central would more than double the median investor’s expected trading performance. Again, the results are robust in a matched sample, and in addition the results are nearly identical when I jointly estimate the propensity to trade around mergers along with the returns to being central using a Heckman (1976) selection model.

Common sources of information due to merger and acquisition (M&A) advisor connections do not explain the returns to being central. Investors connected to M&A advisors have 0.91% (0.87%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. Controlling for M&A advisor connections, central investors still have 1.40% (1.44%) higher principal-weighted (equal-weighted) trading performance.

Centrality is distinct from expertise, and both predict round-trip trading performance around new product announcements. In a multiple regression, experts have 0.51% (0.49%) higher principal-weighted (equal-weighted) average abnormal trading performance, and central investors have 0.66% (0.66%) higher trading performance. Omitting expertise, central investors have 0.65% (0.65%) higher principal-weighted (equal-weighted) round-trip trading performance.
Finally, when there is no information to share, there is no information advantage to being central. Central investors’ trading performance around sudden deaths, relative to peripheral investors, is statistically indistinguishable from zero.

This paper makes two contributions. First, I find evidence that institutional investors regularly share valuable information. Information Diffusion Centrality predicts trading performance, and is also very persistent, which suggests that the consistent outperformance observed in the literature, e.g. Kacperczyk, Sialm, and Zheng (2008) and Puckett and Yan (2011), is due to a conversational equilibrium in the spirit of Stein (2008). The second contribution is methodological. Most network analysis in economics and finance assumes that connections are binary, and ignores network flows. Binary network representations ignore the fact that novel information often flows through “weak ties” (Bakshy, Rosenn, Marlow, and Adamic, 2012). Weighted Trading Correlation Networks represent the strength of connections using observable trading activity. The challenge of modelling network flows is more subtle. The same network structure can embed many different dynamics—moreover, each measure of centrality implicitly assumes a particular network flow. For example, existing measures of centrality in the finance literature—borrowed from the sociology literature—are measures of the diffusion of influence (e.g. Bonacich, 1972a,b). I show that using measures of influence as proxies for information diffusion can lead to the opposite empirical conclusions.\footnote{Borgatti (2005) shows via simulation that measures of centrality give the “wrong” answer when used as proxies for other flows.} Information Diffusion Centrality is the first theoretically justifiable measure of centrality in the finance literature, to my knowledge,
to model bilateral information sharing.

Overall, my results support the view that the way in which economic agents converse—i.e., the structure and dynamics of social and information networks—has important implications for financial markets. Shiller (1995) argues that many behavioral biases may be explained by patterns of conversations. Hirshleifer (2015) argues that the next step in behavioral finance is to study social finance—in other words, to incorporate theories of sociology into our understanding of financial markets. Therefore herding, home bias, and underreaction—three of the most well-documented behavioral biases—are arguably rational given the localization of information embedded within the community structure of networks.

The remainder of the paper is organized as follows. Section 2.1 reviews the literature. Section 2.2 describes the network framework. Section 2.3 describes the institutional trading data. Section 2.4 demonstrates that central investors have superior trading performance on average, and Section 2.5 shows that central investors have superior trading performance around merger announcements. Section 2.6 examines alternative hypotheses. Section 2.7 concludes.

2.1 Background and Related Literature

The empirical literature on “community effects” is often interpreted as a evidence of word-of-mouth communication. Community effects appear to be an important factor in individuals’ stock market participation (Brown, Ivković, Smith, and Weisbenner, 2008), stock purchasing decisions (Ivković and Weisbenner, 2007), and local portfolio
performance (Ivković and Weisbenner, 2005). Institutional investors who live in the same city and same neighborhood also have similar holdings (Hong, Kubik, and Stein, 2005; Pool, Stoffman, and Yonker, 2015). Pool, Stoffman, and Yonker (2015) also shows that “neighborhood” portfolios earn abnormal returns which suggests that there is an information advantage to being local, consistent with Coval and Moskowitz (1999, 2001).

With several significant differences, my paper is most closely related to Ozsoylev, Walden, Yavuz, and Bildik (2014) (OWYB). First, OWYB estimates networks based on overlapping trades, and use an ad-hoc measure called Rescaled Centrality to show that more central investors earn higher returns. I estimate networks based on correlated trading, and I develop Information Diffusion Centrality based on a model of bilateral information sharing. Second, I extend their analysis to rule out alternative hypotheses that could explain the returns to being central including superior information processing, common sources of information, expertise, and spurious return predictability. Third, OWYB use data from the Istanbul Stock Exchange during 2005 which is comprised of 99.9% individual investors, whereas I focus on institutional investors in the US from 1999Q1–2011Q3. It is not surprising that individuals rely on word of mouth if they are uninformed investors. But institutions are informed, have price impact, and compete for information. Furthermore institutions make up the majority of US stock trading volume (Boehmer and Kelley, 2009), hence word of mouth among institutions is likely to have a larger economic impact.

---

8E.g. Hendershott, Livdan, and Schürhoff (2015); Sias and Starks (1997); Akins, Ng, and Verdi (2011)
2.2 Framework

I begin with a brief introduction to networks in Section 2.2.1, and describe the empirical estimation of Weighted Trading Correlation Networks in Section 2.2.2, before defining Information Diffusion Centrality in Section 2.2.3. I describe the measures of abnormal trading performance in Section 2.2.4.

2.2.1 Networks Primer

Weighted Networks

Most network analysis focuses on collections of nodes joined by edges in a binary fashion. These networks are typically represented by (0,1) “adjacency” matrices:

\[
A_{ij} = \begin{cases} 
1 & \text{if } i \text{ is connected to } j \\
0 & \text{otherwise}
\end{cases}.
\] (2.1)

However, many settings, especially social and economic ones, include inherently “stronger” links and “weaker” links (Granovetter, 1973). Both strong and weak links are important, and focusing on only binary adjacency matrices ignores potentially important data. Therefore, we can also represent networks in terms of “weighted” adjacency matrices, \( A_{ij} = w_{ij} \), where \( w_{ij} \) is the weight or strength of the connection between \( i \) and \( j \).
Degree Centrality and Random Walks on Networks

The simplest measure of network centrality is Degree Centrality which is the sum of a node’s direct connections. The Degree Centrality for node $i$ is defined as the row sum of the adjacency matrix $A$:

$$d_i = \sum_j A_{ij}.$$  \hspace{1cm} (2.2)

The diagonal matrix, $D_{ii} = d_i$, is often used to define a random walk on a network $W = D^{-1}A$. The typical off-diagonal element, $W_{ij} = \frac{1}{d_i}$, describes the probability of a random walker jumping from node $i$ to node $j$. Note that the probability of a jump from $i$ to any of its immediate neighbors is identical, and the row-sums of $W$ are one. A symmetric random walk is defined as $W = D^{-1/2}AD^{-1/2}$. The typical off-diagonal element of the symmetric random walk matrix is $W_{ij} = \frac{w_{ij}}{\sqrt{d_i} \sqrt{d_j}}$. Unlike the “standard” random walk, the probability of a jump from $i$ to $j$ is the same as the probability of a jump from $j$ to $i$.

As a simple example, consider an un-weighted network of five nodes A–E in Figure 2.1 below. A standard random walker (Panel A) from A jumps to B with probability $\frac{1}{4}$, and with probability one from B to A. The symmetric random walker jumps from A to B, and vice versa, with probability $\frac{1}{2}$. As we will see in Section 2.2.3, the symmetric random walk matrix will be useful in describing the dynamics of bilateral information sharing.
Figure 2.1: **Random Walks on a Network.** The standard random walk in Panel A assumes that a random walker from A jumps with equal probability $\frac{1}{4}$ to any one of the nodes B–E, while a random walker from B jumps with probability 1 to A. The symmetric random walk in Panel B assumes that a random walker from A jumps to B with the same probability that a random walker from B jumps to A, which in this example is $\frac{1}{2}$.

(a) Standard Random Walk  
(b) Symmetric Random Walk

2.2.2 Estimating Weighted Trading Correlation Networks

We often do not know the strength of the connections between investors and hence the adjacency matrix must be estimated from trading data. I estimate the strength of the connections between investors using pairwise trading correlations. More frequent information sharing is reflected in higher trading correlations and stronger links.

An investor’s buying activity of $n$ stocks over $T$ periods can be represented by a $(0,1)$ $nT$-vector stacked by stock, $b_i = [b_{i1}, b_{i2}, b_{i3}, \ldots, b_{in}]'$. $b_{in}$ is a $T$-vector which takes on values of one if if investor $i$ bought stock $n$ in period $t$ and is zero otherwise. The buy correlation between investor $i$ and $j$ is the Pearson correlation, $\rho_{ij}^b = \frac{\text{cov}(b_i, b_j)}{\sigma_b_i \sigma_b_j}$, and the sell correlation, $\rho_{ij}^s$, is defined analogously for the vector of selling activity.

I estimate the weight of the connections between investors by taking the average of
buy and sell trading correlations: \( \hat{w}_{ij} = \rho_{ij}^b + \rho_{ij}^s \). I call the resulting adjacency matrix with elements \( \hat{A}_{ij} \in [0,1] \) the Weighted Trading Correlation Network (WTCN).\(^9\)

### 2.2.3 Information Diffusion Centrality

The ideal measure of an investors’ information advantage should capture not only how much information an investor receives, but how delayed the information is as it diffuses through the network. Many measures of centrality exist in the finance literature but the question of what measure is “correct” is difficult to answer empirically because most measures of centrality are highly correlated with one another (see for example Valente, Coronges, Lakon, and Costenbader, 2008).\(^\text{10}\) Therefore, the best way to distinguish between measures of centrality is to understand the implicit assumptions that each makes about network flows. For a more detailed discussion of centrality and network flows see Borgatti (2005).

Degree Centrality (Equation 2.2) measures flows that involve no indirect connections. In the context of information networks, Degree corresponds to information linkages generated by correlated signals. In practice, information linkages may be due to correlated signals or information sharing, both of which would generate correlated trading (Colla and Mele, 2010). If information linkages are primarily due to

---

\(^9\)Weighted networks based on pairwise correlations are common in the genomics and systems biology literature (see, Horvath, 2011, for example).

correlated signals, then investors with high Degree, i.e. many signals, should have an information advantage. However if information sharing is important, then Degree Centrality will be unable to capture information “heard through the grapevine” as it ignores access to information through indirect links.

Eigenvector Centrality measures flows of influence or importance (Bonacich, 1972a,b). Eigenvector Centrality is defined recursively in vector form:

$$\varepsilon = \frac{1}{\lambda} A\varepsilon.$$  \hspace{1cm} (2.3)

Bonacich and Katz Centrality are generalizations of Eigenvector Centrality, all of which define investors’ centrality based on connections to other central investors. Eigenvector-like measures are useful because they take into account indirect links and are easy to compute. Ozsoylev, Walden, Yavuz, and Bildik (2014) use Eigenvector Centrality as a proxy for information diffusion, although it is more appropriate as a measure of social influence on opinion formation in the context of DeMarzo, Vayanos, and Zwiebel (2003). In short, if I persuade an influencer to share my opinion, then I am important.

I define Information Diffusion Centrality as:\textsuperscript{11}

$$c = S_i p.$$  \hspace{1cm} (2.4)

\textsuperscript{11}It is worth mentioning that Information Diffusion Centrality is distinct from the Diffusion Centrality defined from Banerjee, Chandrasekhar, Duflo, and Jackson (2013) which is essentially a finite sum version of Katz Centrality. In Appendix A.2 I show that Katz Centrality can be derived as a special case of a proxy for Information Diffusion Centrality, hence Diffusion Centrality can be derived in a similar fashion.
The heart of Information Diffusion Centrality is the symmetric information sharing matrix $S_t = e^{-(t-W)t}$, where $I$ is the identity matrix and $W$ is the symmetric random walk matrix. The information sharing matrix describes the dynamics of continuous bilateral information sharing based on a stylized model of “social learning” in which agents take averages of the differences between their signals and the signals of their neighbors (see Appendix A.1). Information Diffusion Centrality is parameterized by $t$ and $p$. $t > 0$ describes the time scale of information diffusion, such that when $t \to \infty$ information diffuses completely. $p$ represents the probability that an investor receives new information. In empirical applications, I assume that $t = 1$ and $p$ is uniform, which assumes information arrives with equal probability to each investor.\footnote{In other words, agents only update based on new or different information. According to Kahneman (2011), “our mind has a useful capability to focus on whatever is odd, different or unusual.”}

Information Diffusion Centrality can be interpreted as a measure of the expected discounted information advantage as information diffuses through the network following the paths of symmetric random walks on the network. To see this interpretation more clearly, we can rewrite Information Diffusion Centrality as an exponential sum of symmetric random walks $W$:

$$c = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} W^k p,$$  \hspace{1cm} (2.5)

\footnote{The assumption of $t = 1$ is innocuous. Assuming $p \propto d$, i.e. better connected investors receive information from outside the network more frequently, makes Information Diffusion Centrality more correlated with Degree and Eigenvector, which actually reduces its ability to predict trading performance. Hence, the uniform assumption seems to be a better description of the arrival of information to the network.}
which follows from the definition of a matrix exponential. The $ij$th element of the $k$-order symmetric random walk matrix $W^k$ describes the probability of a random walker jumping from node $i$ to node $j$ in $k$ hops. The $ith$ element of $W^k p$ is the sum of the probabilities of information arriving at node $\forall j \neq i$ and diffusing to node $i$ in $k$ hops. The contribution to centrality from all $k$-order random walks is “discounted” by $\frac{t^k}{k!}$. Naturally, information heard second-hand is more valuable than information heard third-hand, especially when opportunities for profitable trading are short-lived.

The symmetric random walk interpretation implies that connections to low degree nodes can actually provide more centrality than high degree nodes. If $i$’s neighbor $j$ has few connections, i.e. low degree $d_j$, then the probability of a random walk jumping from $j$ to $i$ is relatively high. Eigenvector Centrality, by its recursive construction, predicts the opposite—connections to higher degree nodes defines importance. Combining the discounting and symmetric random walk features, we can interpret Information Diffusion Centrality as discounting commonly-held information in favor of novel information.

To visualize the difference between Eigenvector Centrality and Information Diffusion Centrality we can look at a simple example with 13 nodes in Figure 2.2. The most central node $C$ has six connections, with access to novel information from five unique connections and one connection to the cluster on the bottom left. The cluster is fully connected such that any information that arrives to any of the seven nodes is shared with all seven members. As a result, any information in the fully connected cluster is commonly held. Node size is proportional to the number of con-
Connections and node color correlates with Information Diffusion Centrality on the left and Eigenvector Centrality on the right. Information Diffusion Centrality correctly identifies C as the most central node because it has access to novel information from five nodes, while nodes in the fully connected cluster have low centrality despite each node having many connections. According to Eigenvector Centrality, C is peripheral while all of the members of the fully connected cluster are central because each of them is connected to other well-connected nodes.

Figure 2.2: Information Diffusion vs. Eigenvector Centrality. The central node C has six connections, five unique connections, and one to a fully connected cluster of seven nodes. Node size is proportional to the number of connections and node color is proportional to Information Diffusion Centrality on the left and Eigenvector Centrality on the right. Information Diffusion Centrality identifies C as the most central node, whereas the nodes in the fully connected cluster are not central despite the fact that each of them has many connections. Eigenvector Centrality identifies C as being peripheral because most of its connections are not well-connected.
2.2.4 Trading Performance

In the main analysis I compute quarterly interim trading performance by tracking the abnormal performance of all stocks bought and sold by a fund from the execution date, using the execution price, until the end of the quarter. In the event-study portion of the paper I compute round-trip trading performance around specific events. For example, a round-trip trade could consist of a fund buying target stock prior to the public announcement of a merger and selling the same stock following the announcement.

Trading performance computed using actual trades is distinct from implied performance computed using quarterly changes in portfolio holdings. Kacperczyk, Sialm, and Zheng (2008) find a large and persistent return gap between reported fund performance and the return on a portfolio based on changes in holdings. The authors argue that the return gap can be explained by fund managers’ informational advantage in the optimal timing of purchases and sales of individual stocks. Puckett and Yan (2011) use actual trades to show that there exists persistent trading performance among the top performers, whereas using implied quarterly changes produces no persistent performance. The discrepancy arises because changes in holdings lack the exact timing and execution prices of actual trades. Following Puckett and Yan (2011) I compute a fund’s quarterly interim trading performance as the difference between the weighted-average of buy and sell trading performances.

Puckett and Yan (2011) also show that changes in quarterly holdings do not capture round-trip trades. Round-trip trades are important if private information is
short-lived and trading opportunities dissipate quickly. Investors trading on short-lived information may want to reverse their positions in order to lock in gains or exploit overreactions (Hirshleifer, Subrahmanyam, and Titman, 1994; Brunnermeier, 2005). To measure round-trip trading around an event I restrict my sample to trades made 60 days before and 30 days after the announcement date. I define a round-trip trade as a buy (sell) trade in the [-60,-1] window followed by a subsequent sell (buy) in the [0,30] window. A fund’s round-trip trading performance for a particular event is computed as the principal- or equal-weighted average of all signed holding period returns using the actual execution prices and volume traded. I compute a fund’s quarterly round-trip trading performance as the average round-trip trading performance across events that occur in a given quarter in which a fund trades.

For both interim and round-trip trading performance I adjust prices and share volume for stock splits. For interim trading performance I also cumulate dividends over the interim holding period and include them in the return calculations. I also compute “excess” or abnormal trading performance using Daniel, Grinblatt, Titman, and Wermers 1997 (DGTW) size, book-to-market, and momentum characteristic-matched value weighted portfolio returns computed over the corresponding (interim or round-trip) holding period and adjusted for delistings.

2.3 Data

To estimate an investor’s centrality and trading performance I use high-frequency institutional trading data from a proprietary database provided by ANcerno Ltd.
(a.k.a. Abel/Noser Solutions Ltd.). The ANcerno data allows me to observe the exact date, price, direction (buy or sell), and shares traded for all funds in the database from 1999Q1 to 2011Q3. My sample stops in 2011Q3 because ANcerno removed fund identifiers in 2012, preventing me from tracking subsequent trading activity. Furthermore, 2011Q4 and 2012Q1 have only one-third of the number of funds compared to 2011Q3 which results in unreliable network estimates.

I restrict my sample to all trades made in the US, in US currency, of common stock listed on the NYSE, AMEX, and NASDAQ exchanges, for which I am able to identify the ANcerno fund, and for which I am able to match the ANCerno provided point-in-time CUSIP to a CRSP PERMNO. The 10,355 funds in my sample trade a total of 8,555 unique common stocks. Overall, the funds make 141.85 million trades of 1.11 trillion shares valued at USD 34.26 trillion dollars. Puckett and Yan (2011) estimate that the ANcerno institutions account for approximately 8% of the total dollar value of CRSP trading volume between 1999 and 2005. Table A.1 in Appendix A.3 provides yearly summary statistics on the trading activity in my sample.

Because ANcerno timestamps are incomplete (see Anand, Irvine, Puckett, and Venkataraman, 2013) I estimate Weighted Trading Correlation Networks based on daily-level trading activity. To mitigate concerns that overlapping trades are coincidental, I restrict my sample of overlapping trades to “time-sensitive” overlapping trades. To identify time-sensitive trades, I exploit the fact that ANcerno tracks the number of days over which a stock was traded as part of an order ticket. I define time-sensitive trades as trades of stocks executed within a single day. If a fund manager receives short-lived information he or she would have a strong incentive to trade
immediately, versus splitting up the order execution over multiple days, consequently sacrificing price impact for immediacy. On average, time-sensitive trades make up 62% of order executions. The resulting quarterly network snapshots contain on average 1,192 funds, with a minimum of 780 in 2011Q3 and a maximum of 1,639 in 2002Q1.

In the event-study portion of the paper I use merger and acquisition (M&A) announcements, new product announcements, and news announcements of the sudden deaths of board directors and key managers (CEO, CFO, VPs, etc.). The data on mergers comes from the Thomson Reuters SDC Platinum database (SDC). I use SDC’s detailed information on M&As to identify announcement dates, the identities of the target firms, and the names of the merger advisors.

The data on new product announcements and sudden deaths comes from the S&P Capital IQ (CapIQ) Key Developments database (codes 41; 16, 101, and 102 respectively). The CapIQ data includes headlines and full-text articles of over 100 types of significant corporate events. To identify sudden deaths I search through headlines and full-text articles for keywords such as death of, demise of, and passing of. In addition, I filter out deaths of individuals over the age of 65, deaths related to cancer, suicide, disease, illness, or complications as reported in the articles.

2.3.1 Fund Summary Statistics

Table 2.1 Panel A summarizes the main variables of interest. The performance measures are the principal-weighted average abnormal interim trading performance (PW) and the equal-weighted performance (EW). The measures of centrality are Informa-
tion Diffusion Centrality (Centrality), defined in Eq. 2.4, Eigenvector Centrality (Eigenvector) defined in Eq. 2.3 and Degree Centrality (Degree) defined in Eq. 2.2. I also report the average share volume (Volume), and number of trades (\# Trades).

The average fund in my sample has a principal-weighted (equal-weighted) average abnormal interim trading performance of 0.52% (0.65%). For comparison, Puckett and Yan (2011) report that the average fund between 1999 and 2005 has a principal-weighted average abnormal interim trading performance of 0.57%. The average fund in my sample, over the same period, has a principal-weighted abnormal interim trading performance of 0.56%.

Because I am only concerned with a fund's relative centrality within a quarter, I normalize Information Diffusion and Eigenvector Centrality, without loss of generality, such that a fund's centrality can be interpreted as the percentage of the “total centrality” in a given quarter. Normalizing also facilitates comparing the two centrality measures. The median fund has an Information Diffusion Centrality of 2.9% and an Eigenvector Centrality of 0.07%.

The median fund makes 171 trades (Trades), exchanges 669,819 million shares (Volume), and has a Degree of four. The Degree is much smaller than the number of trades because trading correlations tend to be low. The average edge weight is about 2.6%. If every trade made by the median fund overlaps with one fund then the Degree is roughly $171 \times 0.026 \approx 4.5$.

Table 2.1 Panel B reports Spearman correlations for the same variables and includes autocorrelations on the diagonal. The two performance measures have a 77% correlation with one another, and are negatively correlated with the measures trading
activity and Eigenvector and Degree Centrality. Information Diffusion Centrality has a positive correlation of 21 (36) basis points for principal-weighted (equal-weighted) average abnormal trading performance. The centrality measures are positively correlated with one another and the trading measures, and are also highly autocorrelated. Degree and Information Diffusion Centrality have autocorrelations of 91%, which is consistent with the notion that the networks are stable over time and that there exists a sustainable “conversational equilibrium” (Stein, 2008).

Table 2.1 Panel A shows that many of the variables are heavy-tailed with high kurtosis, and the centrality and trading measures, apart from Information Diffusion Centrality, are positively skewed. Most real-world networks are characterized by power-law degree distributions (Barabási and Albert, 1999), and as a result most centrality measures have heavy right tails. Heavy tails tend to bias certain statistics which is why, one, I focus on the medians instead of means, two, I use Spearman correlations instead of Pearson correlations, and three, I take logs of the centrality and trading measures before using them in subsequent regressions. Information Diffusion Centrality and trading performance are slightly negatively skewed. The negative skew is due to the trading correlation edge weighting, which mitigates the influence of active liquidity traders which would otherwise have many connections and have high centrality.

To visualize the effect of edge weighting more clearly, Figure 2.3 plots the distribution of Information Diffusion Centrality over the sample period at the fund-quarter level using both unweighted (−) and trading correlation weighted edges (→). The solid lines are the median values of Information Diffusion Centrality unweighted and
Table 2.1: **Fund Summary Statistics 1999Q1–2011Q3.** Interim trading performance is measured quarterly as the difference between a fund’s principal- or equal-weighted average excess buy return minus its average excess sell return. Returns are computed as the simple return from the actual execution price to the end of quarter closing price including dividends and are also DGTW-adjusted over the same holding period. The principal weight is computed using the actual price × volume traded. Prices and volume are adjusted for stock splits, and benchmark returns are adjusted for delistings. Centrality is Information Diffusion Centrality defined in Eq. (2.4), and Eigenvector is Eigenvector Centrality defined in Eq. (2.3). Both measures are computed based on quarterly trades and reported as percentages of the “total” centrality. Degree is the sum of a fund’s weighted edges (trading correlations) in a given quarter. Volume refers to fund-quarter share volume traded, and # of Trades refers to the actual number of transactions. Correlations in Panel B are Spearman correlations and the diagonal entries are autocorrelations.

(a) Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Med.</th>
<th>Std</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW</td>
<td>54,796</td>
<td>0.52%</td>
<td>0.32%</td>
<td>6.3%</td>
<td>−0.01</td>
<td>5.55</td>
</tr>
<tr>
<td>EW</td>
<td>54,796</td>
<td>0.65%</td>
<td>0.46%</td>
<td>5.9%</td>
<td>−0.19</td>
<td>6.00</td>
</tr>
<tr>
<td>Centrality</td>
<td>54,796</td>
<td>2.9%</td>
<td>2.9%</td>
<td>6.0%</td>
<td>−1.86</td>
<td>9.68</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>54,796</td>
<td>0.75%</td>
<td>0.07%</td>
<td>2.9%</td>
<td>7.27</td>
<td>59.55</td>
</tr>
<tr>
<td>Degree</td>
<td>54,796</td>
<td>4.89</td>
<td>4.09</td>
<td>3.04</td>
<td>2.19</td>
<td>7.09</td>
</tr>
<tr>
<td>Volume</td>
<td>54,796</td>
<td>19.03M</td>
<td>669,819</td>
<td>188.40M</td>
<td>33.53</td>
<td>2,287.08</td>
</tr>
<tr>
<td># of Trades</td>
<td>54,796</td>
<td>2,401</td>
<td>171</td>
<td>20,802</td>
<td>25.15</td>
<td>959.94</td>
</tr>
</tbody>
</table>

(b) Correlations

<table>
<thead>
<tr>
<th></th>
<th>PW</th>
<th>EW</th>
<th>Cent</th>
<th>Eig</th>
<th>Deg</th>
<th>Vol</th>
<th>Trades</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW</td>
<td>0.048</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>EW</td>
<td>0.77</td>
<td>0.065</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Cent</td>
<td>0.0021</td>
<td>0.0036</td>
<td>0.91</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Eig</td>
<td>-0.051</td>
<td>-0.070</td>
<td>0.52</td>
<td>0.72</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Deg</td>
<td>-0.026</td>
<td>-0.021</td>
<td>0.23</td>
<td>0.55</td>
<td>0.91</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Vol</td>
<td>-0.030</td>
<td>-0.023</td>
<td>0.14</td>
<td>0.20</td>
<td>0.21</td>
<td>0.84</td>
<td>.</td>
</tr>
<tr>
<td>Trades</td>
<td>-0.00048</td>
<td>0.014</td>
<td>0.31</td>
<td>0.25</td>
<td>0.37</td>
<td>0.66</td>
<td>0.87</td>
</tr>
</tbody>
</table>
weighted, which are highly correlated over time. The shaded regions show the distribution between the 5 and 95 percentiles of Information Diffusion Centrality. The lighter shaded region corresponds to unweighted Information Diffusion Centrality, which is heavily right-skewed and has high variance much like Degree and Eigenvector Centrality. In contrast, the darker shaded region, which corresponds to trading-correlation weighted Information Diffusion Centrality, is not right-skewed and has a noticeably smaller variance.

Figure 2.4 plots an example of a Weighted Trading Correlation Network in 2002Q2. Nodes are arranged using a force-directed layout with edge weights corresponding to the estimated trading correlations. Adjacent nodes are “attracted” to one another, while all nodes have “repulsive” forces. Lighter edges indicate lower trading correlations and darker edges indicate higher trading correlations. Node size corresponds to the number of connections (Degree) and node color corresponds to Information Diffusion Centrality. The most central node (●) using the force-directed layout appears on the right side because it has many distinct connections which provide it access to novel information.

2.3.2 Event Summary Statistics

Table 2.2 Panel A reports the number of M&As, new product announcements, and director and manager deaths, as well as the average number of funds trading in the median event per year. On average there are 695 mergers, 7,749 new product

\[\text{\textsuperscript{14}}\text{For example, the Pearson correlation between unweighted Information Diffusion, Eigenvector, and Degree Centrality is around 90\% which is mostly driven by the outliers in the right tail.}\]
Figure 2.3: **Centrality 1999Q1–2011Q3.** This Figure plots the distribution of Information Diffusion Centrality (Eq. 2.4) over the sample period. The solid lines represent the median, and the shaded regions represent the distribution between the 5 and 95 percentiles of Information Diffusion Centrality. The Figure includes Information Diffusion Centrality based on unweighted (—), and weighted (—) edges.
Figure 2.4: **Weighted Trading Correlation Network (2002Q2).** Nodes represent funds and edges represent trading correlations. Nodes are placed using a force-directed layout with edge weights corresponding to trading correlations. Lighter edges indicate lower trading correlations and darker edges indicate higher trading correlations. Node size is proportional to the number of edges, and node color is proportional to Information Diffusion Centrality. Lighter colored nodes (○) are “central” while darker colored nodes (●) are “peripheral.”
announcements, and 26 deaths per year for which at one fund in my sample makes a round-trip trade. On average, 13 funds trade around the typical merger in a given year, 21 trade around new product announcements, and 8 trade around director and manager deaths. The CapIQ data starts in 2001, which is why the sample of new products and deaths is sparse in the early sample, and the trades data end in 2011Q3 which is why 2011 has fewer events.

Panel B includes average cumulative abnormal returns (CARs) from -30 to +30 based on a one-factor CAPM model as well as average principal-weighted (PW) and equal-weighted (EW) average abnormal round-trip trading performance. The mean CAR for target firms around merger announcements is 6.49%, the mean CAR around new product announcements is -0.74%, and the mean CAR around director and manager deaths is 2.41%. The mean principal-weighted (equal-weighted) average abnormal round-trip trading performance around merger announcements is 1.29% (1.27%) and is statistically different from zero, consistent with the observation that there are information leakages prior to merger announcements (see Betton, Eckbo, and Thorburn, 2008). The mean principal-weighted (equal-weighted) average trading performance around new product announcements is -0.03% (-0.05%) and is statistically significant but economically small. The mean principal-weighted (equal-weighted) average trading performance around deaths is -0.10% (-0.14%) and is statistically indistinguishable from zero, consistent with the notion that sudden deaths are unpredictable (e.g. Nguyen and Nielsen, 2010).

---

The one-factor model is estimated over a year of data (252 days) with a 30-day gap between the estimation window and the event window. Other factor models produce similar results since
Table 2.2: **Event Summary Statistics.** Panel A shows the number of merger announcements, new product announcements, and director and manager deaths per year, as well as the average number of funds trading around the median event. Panel B shows the average cumulative abnormal returns (CARs) and round-trip trading performance around events. Merger announcements come from the Thomson Reuters SDC Platinum database and director and manager sudden deaths and new product announcements come from the S&P Capital IQ Key Developments Database (event types 101, 102, 16 for deaths and 41 for new products). CARs are measured from [-30, +30] using a one-factor CAPM model estimated over a 252-day window with a 30-day gap between the estimation window and the event window. Round-trip trading performance is measured as the principal- or equal-weighted average of all signed simple excess returns of buy (sell) trades initiated between [-60, -1] and reversed between [0, +30]. *t*-statistics are based on un-adjusted standard errors.

**(a) Distribution of Events**

<table>
<thead>
<tr>
<th>Year</th>
<th>Mergers # Events</th>
<th>Mergers # Funds</th>
<th>New Products # Events</th>
<th>New Products # Funds</th>
<th>Deaths # Events</th>
<th>Deaths # Funds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>824</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2000</td>
<td>986</td>
<td>7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2001</td>
<td>668</td>
<td>7</td>
<td>125</td>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2002</td>
<td>597</td>
<td>11</td>
<td>7178</td>
<td>16</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>2003</td>
<td>688</td>
<td>11</td>
<td>7384</td>
<td>18</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>2004</td>
<td>743</td>
<td>15</td>
<td>8835</td>
<td>19</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>2005</td>
<td>803</td>
<td>12</td>
<td>9533</td>
<td>16</td>
<td>37</td>
<td>5</td>
</tr>
<tr>
<td>2006</td>
<td>817</td>
<td>13</td>
<td>9168</td>
<td>20</td>
<td>35</td>
<td>10</td>
</tr>
<tr>
<td>2007</td>
<td>829</td>
<td>15</td>
<td>9084</td>
<td>23</td>
<td>42</td>
<td>7</td>
</tr>
<tr>
<td>2008</td>
<td>697</td>
<td>14</td>
<td>9860</td>
<td>25</td>
<td>48</td>
<td>11</td>
</tr>
<tr>
<td>2009</td>
<td>554</td>
<td>16</td>
<td>9174</td>
<td>29</td>
<td>39</td>
<td>14</td>
</tr>
<tr>
<td>2010</td>
<td>507</td>
<td>16</td>
<td>8865</td>
<td>27</td>
<td>36</td>
<td>10</td>
</tr>
<tr>
<td>2011</td>
<td>324</td>
<td>20</td>
<td>6036</td>
<td>23</td>
<td>17</td>
<td>6</td>
</tr>
</tbody>
</table>

**(b) Event Returns**

<table>
<thead>
<tr>
<th>Event</th>
<th># Events</th>
<th># Trades</th>
<th>CAR</th>
<th>t</th>
<th>PW</th>
<th>t</th>
<th>EW</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>M&amp;A Targets</td>
<td>9,242</td>
<td>171,968</td>
<td>6.49%</td>
<td>19.52</td>
<td>1.29%</td>
<td>32.97</td>
<td>1.27%</td>
<td>34.34</td>
</tr>
<tr>
<td>New Products</td>
<td>84,465</td>
<td>2,770,048</td>
<td>-0.74%</td>
<td>-8.50</td>
<td>-0.03%</td>
<td>-4.91</td>
<td>-0.05%</td>
<td>-8.78</td>
</tr>
<tr>
<td>Deaths</td>
<td>219</td>
<td>4,274</td>
<td>2.41%</td>
<td>1.18</td>
<td>-0.10%</td>
<td>-0.55</td>
<td>-0.14%</td>
<td>-0.90</td>
</tr>
</tbody>
</table>
2.4 Centrality and Trading Performance

If valuable information diffuses through the network, then there should be predictable information asymmetry between investors due to the structure of the network. Central investors, in particular, should consistently outperform peripheral investors.

As formal test of this hypothesis, I estimate predictive regressions of the form:

$$ r_{i,t} = \alpha_t + c_{i,t-1}\beta_0 + X_{i,t-1}\beta_1 + \epsilon_{i,t}. $$

\( r_{i,t} \) is the average abnormal interim trading performance (in logs) computed using DGTW-adjusted returns at the fund-quarter level, denoted by the subscripts \( i \) and \( t \) respectively. I include a quarter fixed effect, \( \alpha_t \), because I am interested in the cross-sectional variation in trading performance. \( c_{i,t-1} \) is the vector of centralities including Information Diffusion, Degree, and Eigenvector Centrality. I say that a fund is central if it has above median centrality in a given quarter. Finally, \( X_{i,t-1} \) is the vector of controls: trading volume, the number of trades, and trading performance all from the previous quarter and in logs.

Table 2.3 presents the main results of the paper. Panel A (B) of Table 2.3 shows OLS estimates from regressions of principal-weighted (equal-weighted) average abnormal interim trading performance on measures of centrality. In Column 1 of Panel A (B), funds with above median Information Diffusion Centrality have, on average, 0.32% (0.48%) higher principal-weighted (equal-weighted) average abnormal trading performance in the following quarter. The effect is statistically different from zero at most of the return is due to the announcement effect.
the one percent level, based on standard errors clustered by fund and quarter. For comparison, the median fund in my sample has a 0.35% (0.47%) principal-weighted (equal-weighted) average abnormal interim trading performance. In other words, becoming central would nearly double the median fund’s expected trading performance.

In contrast, Columns 2 and 3 of Table 2.3 show that funds with above median Degree and Eigenvector Centrality have lower trading performance. Central funds according to Degree Centrality have 0.11% (0.04%) lower trading performance, and central funds according to Eigenvector Centrality have 0.37% (0.42%) lower principal-weighted (equal-weighted) average abnormal interim trading performance. The parameter estimates for Degree are statistically indistinguishable from zero, and the estimates for Eigenvector are large and negative.

Column 4 reports estimates for a multiple regression specification with all three measures of centrality and controls for last quarter’s trading volume, number of trades, and interim trading performance. The results remain unchanged, central funds according to Information Diffusion Centrality have 0.41% (0.52%) higher trading performance, and central funds according to Degree and Eigenvector Centrality have lower trading performance.

Column 5 repeats the multiple regression from Column 4 within a sample of central funds matched to peripheral funds based on Degree, Eigenvector, and the same set of controls. I require that the difference in propensity scores does not exceed 0.1 basis points in absolute value. Central funds, according to Information Diffusion Centrality, have 0.49% (0.57%) higher principal-weighted (equal-weighted)
average abnormal interim trading performance. Within the smaller sample of 10,196 propensity score matched fund-quarter observations, the only observable difference between central and peripheral funds, according to Information Diffusion Centrality, is that central funds have higher Degree Centrality.\textsuperscript{16} Funds with higher Degree tend to have lower trading performance, which biases against the finding that funds with higher Information Diffusion Centrality have higher trading performance as the two measures of centrality are positively correlated.

Overall, the results in Table 2.3 suggest that central funds according to Information Diffusion Centrality have a substantial information advantage. Using a measure of influence—e.g., Eigenvector Centrality—to proxy for information diffusion negatively predicts trading performance. Intuitively, connections to well-connected individuals provides commonly-held information which is not valuable. In contrast, Information Diffusion Centrality measures an investors’ access to novel information which is why it predicts trading performance. Degree has a zero or negative coefficient because the relationship between trading performance and Degree is concave.

\subsection{2.5 Evidence from Merger Announcements}

In this section I restrict my attention to round-trip trading around merger announcements where I expect access to information to be the primary determinant of trading performance. For instance, Ahern (2015) examines 239 instances of illegal insider trading around M&As and finds evidence that individuals closer to the original source

\textsuperscript{16}Normalized differences are reported in Table A.2 in the Appendix.
Table 2.3: **Interim Trading Performance and Centrality.** In Panel A (B) the dependent variable is the quarterly fund-level log principal-weighted (equal-weighted) average abnormal interim trading performance. Centrality, Degree, and Eigenvector are indicator variables which are one if the fund has above median Information Diffusion, Degree, or Eigenvector Centrality in the previous quarter $t - 1$. Control variables include the previous quarter’s trading volume, number of trades, and trading performance, all of which are in logs and are standardized. Column 5 contains central and peripheral funds matched on Degree and Eigenvector (standardized), and the same control variables where the difference in propensity scores does not exceed 0.1 basis points in absolute value. Coefficients and $R^2$ values are reported in percentages. $t$-statistics are reported in parentheses, with standard errors clustered by fund and quarter. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

(a) Principal-Weighted

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Centrality$_{t-1}$</strong></td>
<td>0.322***</td>
<td>0.414***</td>
<td>0.486***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.64)</td>
<td>(4.83)</td>
<td>(2.83)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Degree$_{t-1}$</strong></td>
<td>-0.109</td>
<td>-0.162</td>
<td>-0.408**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.08)</td>
<td>(-1.37)</td>
<td>(-2.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Eigenvector$_{t-1}$</strong></td>
<td>-0.365***</td>
<td>-0.378***</td>
<td>-0.238</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3.42)</td>
<td>(-3.54)</td>
<td>(-1.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Qtr FE</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>$R^2(%)$</strong></td>
<td>0.64</td>
<td>0.59</td>
<td>0.66</td>
<td>1.00</td>
<td>1.88</td>
</tr>
<tr>
<td><strong>Obs.</strong></td>
<td>54,796</td>
<td>54,796</td>
<td>54,796</td>
<td>54,782</td>
<td>11,045</td>
</tr>
</tbody>
</table>
of information earn higher returns. Hence, if central investors are well-positioned to receive information on average, then they might also have superior access to information about mergers.

To test this hypothesis I estimate the same regression as in Equation 2.6, where the dependent variable is now the average abnormal round-trip trading performance based on trades made in target stock around merger announcements. As a robustness test, I restrict my sample to a set of propensity score matched funds as in Section 2.4. Because only a subset of funds trade around mergers, I jointly estimate the propensity to trade and the returns to being central using a Heckman (1976) selection model. The model can be written:

\[
\begin{align*}
    r_{i,t} &= \alpha_t + c_{i,t-1}\beta_0 + X_{i,t-1}\beta_1 + \epsilon_{i,t}^1 \quad \text{(main regression)} \\
    Z_{i,t-1}\gamma + \epsilon_{i,t}^2 & \quad \text{(selection equation)}
\end{align*}
\]
where \( \epsilon^1 \sim N(0, \sigma), \epsilon^2 \sim N(0, 1) \), and \( \text{corr}(\epsilon^1, \epsilon^2) = \rho \). The assumption is that \( r_{i,t} \) is observed, i.e. a fund trades in mergers, if \( Z_{i,t-1} \gamma + \epsilon^2_{i,t} > 0 \). I include an indicator for whether a fund’s initial Centrality, Degree, and Eigenvector is above median as additional instruments in \( Z \).

Table 2.4 Columns 1–3 (4–6) present estimates from regressions of principal-weighted (equal-weighted) average abnormal round-trip trading performance on measures of centrality and trading activity. In Column 1 (4), central funds have 1.50% (1.53%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. To put the magnitude of the estimate in perspective, the median fund has a 1.27% (1.24%) principal-weighted (equal-weighted) average abnormal round-trip trading performance around merger announcements. Becoming central would more than double the median fund’s expected trading performance. Funds with above median Degree Centrality have 1.39% (1.47%) lower trading performance, and funds with above median Eigenvector Centrality have 3.9% (3.8%) lower trading performance.

In Column 2 (5), within the sample of central funds matched to peripheral funds, central funds have 2.46% (2.45%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. Funds with above median Degree Centrality have 1.13% (1.17%) lower trading performance, and funds with above median Eigenvector Centrality have 3.16% (3.12%) lower trading performance.

The model in Column 3 (6) controls for the propensity to trade around mergers within the matched sample. As before, central funds have 2.55% (2.58%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance.
Funds with above median Degree Centrality have 1.33% (1.42%) lower trading performance, and funds with above median Eigenvector Centrality have 2.99% (2.95%) lower trading performance.

Taken together, the above evidence suggests that central funds appear to be informed about mergers before they are publicly announced. Central funds have more than three times higher round-trip trading performance relative to the average fund within the matched sample and controlling for the likelihood of trading around mergers. Degree and Eigenvector Centrality are unable to predict superior trading performance because they do not capture the diffusion of information.

### 2.6 Ruling Out Alternative Hypotheses

In this section I show that centrality is unrelated to common sources of information, industry or product market expertise, and does not predict trading performance spuriously when there is no private information.

#### 2.6.1 M&A Advisor “Tipping”

Large broker-dealers frequently act as merger advisors to acquiror and target firms and potentially “tip off” their institutional clients to these deals. Jegadeesh and Tang (2011) find that institutions connected to merger advisors are net buyers of target stocks—they also make profitable trades in aggregate. Therefore, brokerage house connections may be an important source of common information about mergers.

I begin by name-matching ANcerno brokers to SDC merger advisors on both the
Table 2.4: **Round-trip Trading around M&A Announcements — Targets.** Merger announcements come from the Thomson Reuters SDC Platinum Database. The dependent variable is the quarterly fund-level principal-weighted (equal-weighted) average abnormal round-trip trading performance in target stock around merger announcements. Centrality, Degree, and Eigenvector are indicator variables which are one if the fund has above median Information Diffusion, Degree, or Eigenvector Centrality in the previous quarter $t - 1$. Control variables include the previous quarter’s trading volume, number of trades, and trading performance, all of which are in logs and are standardized. Columns 2 and 5 contain central and peripheral funds matched on Degree and Eigenvector (standardized), and the same control variables where the difference in propensity scores does not exceed 0.1 basis points in absolute value. In Columns 3 and 6 I include centrality indicators based on the first fund-quarter observation as additional instruments in a Heckman selection model (Equation 2.7) estimated using maximum likelihood. $\lambda$ refers to the coefficient on the inverse Mills ratio, and standard errors are clustered at the quarterly level. Coefficients and $R^2$ values are reported in percentages. OLS $t$-statistics are reported in parentheses, with standard errors clustered by fund and quarter. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>Principal Weighted</th>
<th></th>
<th>Equal Weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td><strong>Centrality$_{t-1}$</strong></td>
<td>1.495***</td>
<td>2.463***</td>
<td>2.546***</td>
<td>1.531***</td>
</tr>
<tr>
<td></td>
<td>(4.29)</td>
<td>(3.22)</td>
<td>(3.47)</td>
<td>(4.29)</td>
</tr>
<tr>
<td><strong>Degree$_{t-1}$</strong></td>
<td>-1.389**</td>
<td>-1.132</td>
<td>-1.335</td>
<td>-1.465**</td>
</tr>
<tr>
<td></td>
<td>(-2.34)</td>
<td>(-1.11)</td>
<td>(-1.29)</td>
<td>(-2.48)</td>
</tr>
<tr>
<td><strong>Eigenvector$_{t-1}$</strong></td>
<td>-3.907***</td>
<td>-3.159***</td>
<td>-2.985***</td>
<td>-3.839***</td>
</tr>
<tr>
<td></td>
<td>(-6.60)</td>
<td>(-3.04)</td>
<td>(-2.94)</td>
<td>(-6.54)</td>
</tr>
<tr>
<td>$\lambda$ (%)</td>
<td>0.34</td>
<td></td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>s.e.($\lambda$)</td>
<td>0.36</td>
<td></td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Qtr FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$R^2$ (%)</td>
<td>2.40</td>
<td>4.69</td>
<td>2.50</td>
<td>4.73</td>
</tr>
<tr>
<td>Obs.</td>
<td>21,658</td>
<td>4,171</td>
<td>11,544</td>
<td>21,658</td>
</tr>
</tbody>
</table>
target and acquiror side. In most M&A deals there are multiple advisory firms on each side, so there are potentially many sources of informative tips. In order to try and isolate the relevant connections, I compute the total shares traded by each fund with each of their brokers for each quarter. I say that a fund is connected to an M&A advisor if in the previous quarter the advisor was one of the top-quintile brokers for that fund.\textsuperscript{17} I then repeat the round-trip trading analysis with an indicator for whether a fund has any connections to an M&A advisor (Tip=1) in the previous quarter.

Table 2.5 Column 1 (4) shows that funds connected to M&A advisors have 0.91\% (0.87\%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. Central funds, those with higher than median Information Diffusion Centrality, still have 1.40\% (1.44\%) higher principal-weighted (equal-weighted) trading performance. Both estimates are statistically significant at the one percent level.

In Column 2 (5), within the sample of central funds matched to peripheral funds, funds connected to M&A advisors have 0.63\% (0.78\%) higher trading performance although the estimate is not statistically significant. Central funds have 2.45\% (2.43\%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance.

The model in Column 3 (6) controls for the propensity to trade around mergers within the matched sample. Funds connected to M&A advisors have 0.39\% (0.54\%) average abnormal round-trip trading performance.

\textsuperscript{17}Jegadeesh and Tang (2011) focus on top decile brokers at the yearly level.
higher trading performance and the estimate is not statistically significant. As before, central funds have 2.54% (2.57%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance.

Table 2.5: Tipping around M&A Announcements — Targets. A fund is connected to a brokerage house (Tip=1) if the brokerage house was one of the funds’ main brokers in the last quarter. A fund’s brokers are ranked according to share volume executed in each quarter, and top quintile brokers are identified as main brokers. Control variables include the previous quarter’s trading volume, number of trades, and trading performance, all of which are in logs and are standardized. Coefficients and $R^2$ values are reported in percentages. OLS $t$-statistics are reported in parentheses, with standard errors clustered by fund and quarter. Stars indicate significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

<table>
<thead>
<tr>
<th></th>
<th>Principal Weighted</th>
<th></th>
<th>Equal Weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Centrality$_{t-1}$</td>
<td>1.400*** (4.02)</td>
<td>2.448*** (3.20)</td>
<td>2.540*** (3.45)</td>
<td>1.440*** (4.06)</td>
</tr>
<tr>
<td>Degree$_{t-1}$</td>
<td>-1.491** (-2.52)</td>
<td>-1.219 (-1.20)</td>
<td>-1.395 (-1.37)</td>
<td>-1.562** (-2.65)</td>
</tr>
<tr>
<td>Eigenvector$_{t-1}$</td>
<td>-3.851*** (-6.56)</td>
<td>-3.106*** (-3.02)</td>
<td>-2.938*** (-2.94)</td>
<td>-3.787*** (-6.50)</td>
</tr>
<tr>
<td>Tip</td>
<td>0.906*** (3.93)</td>
<td>0.632 (0.95)</td>
<td>0.392 (0.53)</td>
<td>0.869*** (3.72)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>0.37 (0.37)</th>
<th></th>
<th>0.35 (0.37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$(%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.e.($\lambda$)</td>
<td></td>
<td>0.37 (0.37)</td>
<td></td>
<td>0.37 (0.37)</td>
</tr>
<tr>
<td>Qtr FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$R^2$(%)</td>
<td>2.48</td>
<td>4.73</td>
<td>2.57</td>
<td>4.78</td>
</tr>
<tr>
<td>Obs.</td>
<td>21,658</td>
<td>4,171</td>
<td>11,544</td>
<td>21,658</td>
</tr>
</tbody>
</table>

Controlling for brokerage house connections does not attenuate the returns to being central by much. The returns to being central in Column 1 (3) are only ten basis points lower than are reported in Table 2.4. In Columns 2–3 (5–6) the returns
to being central are only one basis point lower. These results suggest that centrality is largely unrelated to brokerage house connections.

2.6.2 Expertise versus Information Diffusion

Trading performance may also be related to expertise in a particular industry or product market (Kacperczyk, Sialm, and Zheng, 2005). I examine new product announcements as a setting in which I expect product market expertise to be distinguishable from centrality as predictors of trading performance.

To measure product market expertise I rely on the text-based network industry classifications (TNICs) provided by Hoberg and Phillips (2010, 2015). Hoberg and Phillips (2010, 2015) classify firms into TNICs industries using the product descriptions in 10-K filings. Each product announcement in my sample is assigned to an industry based on yearly GVKEY-TNIC pairs. I compute a fund’s expertise in the announcing firm’s product market as the trade-weighted average cosine similarity between the announcing firm’s product description and the product descriptions of all of the other firms in which the fund trades. The resulting measure is bounded between zero and one, where larger values indicate that a fund’s trading is concentrated in the announcing firm’s industry. Intuitively, if a fund exclusively trades in technology stocks such as AAPL, MSFT, and IBM then the fund is classified as an expert in the TNIC technology industry.\(^\text{18}\) I say that a fund is an Expert = 1 if it has above

\(^{18}\)TNIC definitions change over time as product categories change. Because I am only concerned with whether a fund trades stocks similar to the announcing firm within a given year, the time series variation in TNICs does not affect my definition of expertise. See Hoberg and Phillips (2010, 2015) for more details.
median average expertise, measured across all round-trip trades in announcing firm stock in a given quarter.

In Table 2.5 Column 1 (5), central funds, those with higher than median Information Diffusion Centrality, have 0.65% (0.65%) higher principal-weighted (equal-weighted) average abnormal round-trip trading performance. Controlling for expertise, the returns to being central appear to be higher. In Column 2 (6), Experts have 0.51% (0.49%) higher principal-weighted (equal-weighted) average trading performance, and central funds have 0.66% (0.66%) higher trading performance.

In Column 3 (7) I construct a sample of Experts matched to non-Experts.19 In the matched sample Experts have 0.75% (0.72%) higher principal-weighted (equal-weighted) average trading performance, and central funds have 0.80% (0.80%) higher trading performance.

The model in Column 4 (8) controls for the propensity to trade around product announcements within the matched sample. Experts have 0.55% (0.53%) higher principal-weighted (equal-weighted) average trading performance, and central funds have 0.75% (0.75%) higher trading performance.

The results in Table 2.6 indicate that both Centrality and Expertise are significant predictors of round-trip trading performance around new product announcements. More importantly, centrality is distinct from expertise.

19I form matches based on log standardized Information Diffusion Centrality and the same covariates reported in Section 2.4. A matched sample of central and peripheral investors produces similar magnitude estimates.
Table 2.6: **Round-trip Trading around New Product Announcements.** New product events are identified in the S&P Capital IQ Key Events Database as event type 41. Hoberg and Phillips (2010, 2015) classify firms into industries (TNICs) using the product descriptions found in 10-K filings. Each product announcement is assigned to an industry based on yearly GVKEY-TNIC pairs. A fund’s product market expertise is measured as the trade-weighted average cosine similarity between the announcing firm’s product description and the product descriptions of all of the other firms in which the fund traded stocks. A fund is an Expert = 1 if it has above median average expertise in a given quarter. Columns 3 and 7 contain Expert and non-Expert funds matched on on Degree and Eigenvector (standardized), and controls for trading activity. I require that the difference in propensity scores does not exceed 0.1 basis points in absolute value. In Columns 4 and 8 I include centrality indicators based on the first fund-quarter observation as additional instruments in a Heckman selection model (Equation 2.7) estimated using maximum likelihood. λ refers to the coefficient on the inverse Mills ratio, and standard errors are clustered at the quarterly level. Coefficients and $R^2$ values are reported in percentages. OLS $t$-statistics are reported in parentheses, with standard errors clustered by fund and quarter. Stars indicate significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 

<table>
<thead>
<tr>
<th></th>
<th>Principal Weighted</th>
<th></th>
<th>Equal Weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Centrality$_{t-1}$</td>
<td>0.680***</td>
<td>0.688***</td>
<td>0.832***</td>
<td>0.781***</td>
</tr>
<tr>
<td></td>
<td>(3.08)</td>
<td>(3.14)</td>
<td>(2.98)</td>
<td>(3.02)</td>
</tr>
<tr>
<td>Degree$_{t-1}$</td>
<td>-0.414</td>
<td>-0.406</td>
<td>-0.281</td>
<td>-0.275</td>
</tr>
<tr>
<td></td>
<td>(-1.22)</td>
<td>(-1.19)</td>
<td>(-0.64)</td>
<td>(-0.66)</td>
</tr>
<tr>
<td>Eigenvector$_{t-1}$</td>
<td>-0.746**</td>
<td>-0.756**</td>
<td>-0.817**</td>
<td>-0.792**</td>
</tr>
<tr>
<td></td>
<td>(-2.21)</td>
<td>(-2.24)</td>
<td>(-2.22)</td>
<td>(-2.43)</td>
</tr>
<tr>
<td>Expert</td>
<td>0.522***</td>
<td>0.741***</td>
<td>0.550***</td>
<td>0.491***</td>
</tr>
<tr>
<td></td>
<td>(2.86)</td>
<td>(2.73)</td>
<td>(2.59)</td>
<td>(2.72)</td>
</tr>
<tr>
<td>$\lambda(%)$</td>
<td>0.017</td>
<td></td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>s.e.($\lambda$)</td>
<td>0.14</td>
<td></td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Qtr FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$R^2(%)$</td>
<td>1.15</td>
<td>1.19</td>
<td>1.62</td>
<td>1.33</td>
</tr>
<tr>
<td>Obs.</td>
<td>27,219</td>
<td>27,219</td>
<td>17,662</td>
<td>35,008</td>
</tr>
</tbody>
</table>
2.6.3 (No) Timing Sudden Director and Manager Deaths

An important concern is whether central fund managers have genuine information or whether centrality predicts trading performance due to pseudo market timing. For example, Butler, Grullon, and Weston (2005) finds that managers pseudo market time equity markets with their share issuances and appear to be able to forecast unanticipated events such as the Japanese attack of Pearl Harbor.

As a falsification test, I look at round-trip trading around sudden deaths of directors and managers. Because sudden deaths are unpredictable—centrality—as a proxy for access to information, should play no role in the performance of trades made around these events.

Consistent with this intuition, Table 2.7 Column 1 (4) shows that central funds do not have statistically significant higher trading performance relative to the average fund.

In Column 2 (5) I repeat the regression in Column 1 (4) within a sample of central funds matched to peripheral funds based on Degree, Eigenvector, and the same set of controls. I require that the difference in propensity scores does not exceed five basis points in absolute value.\(^{20}\) Again central funds do not have statistically significant higher trading performance relative to the average fund.

The model in Column 3 (6) controls for the propensity to trade around mergers within the matched sample. As before, central funds do not have statistically significant higher trading performance relative to the average fund.

\(^{20}\)Using a caliper of 0.1 basis points reduces the sample size to roughly 200 fund-quarter observations which produces unreliable parameter and standard error estimates.
Table 2.7: Round-trip Trading around Sudden Director & Manager Deaths. Director or manager changes are identified in the S&P Capital IQ Key Events Database as event types 101 (CEO), 102 (CFO), and 16 (Board and Other Executive). I identify sudden deaths by searching the headlines for key phrases including death of, demise of, and passing of. I filter out deaths of individuals over the age of 65, deaths related to cancer, suicide, disease, illness, or complications as reported in the full text articles. Columns 2 and 5 contain central and peripheral funds matched on Degree and Eigenvector (standardized), and controls for trading activity. I require that the difference in propensity scores does not exceed five basis points in absolute value. In Columns 3 and 6 I include centrality indicators based on the first fund-quarter observation as additional instruments in a Heckman selection model (Equation 2.7) estimated using maximum likelihood. \( \lambda \) refers to the coefficient on the inverse Mills ratio, and standard errors are clustered at the quarterly level. Coefficients and \( R^2 \) values are reported in percentages. OLS t-statistics are reported in parentheses, with standard errors clustered by fund and quarter. Stars indicate significance levels: * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).

<table>
<thead>
<tr>
<th></th>
<th>Principal Weighted</th>
<th>Equal Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Centrality_{t-1}</td>
<td>1.059 (0.81)</td>
<td>0.236 (0.23)</td>
</tr>
<tr>
<td>Degree_{t-1}</td>
<td>-1.910 (-1.48)</td>
<td>-0.309 (-0.22)</td>
</tr>
<tr>
<td>Eigenvector_{t-1}</td>
<td>1.001 (0.75)</td>
<td>0.191 (0.11)</td>
</tr>
</tbody>
</table>

\( \lambda (\%) \) | -0.14 | 0.087 |
\( s.e. (\lambda) \) | 0.46  | 0.26  |
Qtr FE             | Yes   | Yes   | Yes  | Yes  | Yes  | Yes  |
Controls           | Yes   | Yes   | Yes  | Yes  | Yes  | Yes  |
\( R^2 (\%) \)    | 0.84  | 0.64  | 0.76 | 0.44 |
Obs.               | 2,277 | 1,880 | 48,362 | 2,277 | 1,880 | 48,362 |
When there is no information to share, there is no information advantage to being central.

2.7 Conclusion

Overall my findings suggest that institutional investors regularly share valuable information with one another. Information Diffusion Centrality predicts trading performance, and is also very persistent, which suggests that the consistent outperformance observed in the literature, e.g. Kacperczyk, Sialm, and Zheng (2008) and Puckett and Yan (2011), is due to a conversational equilibrium in the spirit of Stein (2008). My findings further indicate that word-of-mouth information sharing may be a stylized fact of investor behavior which is not limited to individual investors (Ivković and Weisbenner, 2007) or local institutions (Hong, Kubik, and Stein, 2005; Pool, Stoffman, and Yonker, 2015).

My results support the view that the way in which economic agents converse—i.e., the structure of social and information networks—have important implications for financial markets. Shiller (1995) argues that many behavioral biases may be explained by patterns of conversations, and Hirshleifer (2015) argues that the next step in behavioral finance is to study social finance—in other words, to incorporate theories of sociology into our understanding of financial markets. Herding, home bias, and underreaction, three of the most well-documented behavioral biases, are all arguably related to the localization of information embedded within the community structure of networks. The sociology literature has long recognized that economic
actions such as learning may be “embedded” in social networks (Granovetter, 1985), but network analysis is relatively new to the finance literature. A more detailed subsequent analysis of network structure, beyond centrality, may provide a fuller understanding of the distribution and dispersion of information in financial markets.

Novel information often flows through “weak ties” (Bakshy, Rosenn, Marlow, and Adamic, 2012)—hence empirical work should take advantage of the continuous nature of social interactions.

Finally, I find that using measures of influence as proxies for information diffusion can lead to different conclusions. Empirical researchers in finance should be careful in choosing measures of centrality and understanding the implicit network flows.
Chapter 3

What does the PIN model identify as private information?

The Probability of Informed Trade (PIN) model, developed in a series of seminal papers including, Easley and O’Hara (1987), Easley, Kiefer, O’Hara, and Paperman (1996), and Easley, Kiefer, and O’Hara (1997), has been used extensively in accounting, corporate finance and asset pricing literature as a measure of information asymmetry.\(^1\) The PIN model is based on the notion, originally developed by Glosten and Milgrom (1985), that periods of informed trade can be identified by abnormally large order flow imbalances.\(^2\) Recently, however, several papers have documented PIN anomalies where PINs tend to be at their lowest when information asymmetry should be at its highest (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Joche (2007), and Collin-Dufresne and Fos (2014a)).

\(^{1}\)This is joint work with Jefferson Duarte (Rice University) and Lance Young (University of Washington).

\(^{2}\)A Google scholar search reveals that this series of PIN papers has been cited more than 3,500 times as of this writing. Examples of papers that use PIN in the finance and accounting literature include Duarte, Han, Harford, and Young (2008), Bakke and Whited (2010), Da, Gao, and Jagannathan (2011), and Akins, Ng, and Verdi (2012).

\(^{3}\)Following the literature we define order flow imbalance as the difference between the number of buyer initiated trades less the number of seller initiated trades. In what follows, we refer to buyer initiated trades as ‘buys’ and seller initiated trades as ‘sells.’
We conduct an empirical examination of the PIN model to identify what might cause difficulties in its ability to identify informed trade. This exercise is important because the various possibilities imply very different agendas for this growing area of research. Specifically, we address two possibilities. First, if PIN fails to identify private information because the underlying model does not fit the data well, then the PIN model could, in principle, be corrected by extending the model to better fit the order flow data. Second, it could be that net order flow itself is such a poor indicator of private information that no model based on order flow alone is capable of identifying informed trade, no matter how well it fits the data. Indeed, the theoretical work of Back, Crotty, and Li (2014) and the results in Kim and Stoll (2014) suggest that by itself, order flow imbalance is insufficient to isolate information events. If this is indeed the case, there is little to be gained by extending the PIN model to better fit the order flow data. Instead, a different approach involving variables other than order flow is necessary to generate useful inferences about the arrival of informed trade.

To address these two research questions, we create a variable called the conditional probability of an information event \(CPIE\). To compute the \(CPIE\) implied by the PIN model \(CPIE_{PIN}\), we estimate the PIN model’s parameters using an entire year of data, and then use the observed market data (i.e. buys and sells) to estimate the posterior or model-implied probability of an information event for each day in our sample. We then turn to our first question and examine whether observed variation in \(CPIE_{PIN}\) is consistent with the theory underlying the PIN model. Under the PIN model, private information is identified solely from the absolute order imbalance. In
practice, however, the PIN model may be a poor description of the data and model
misspecification can affect the way it actually identifies private information. To test
this hypothesis, we regress $CPIE_{PIN}$ for each firm-year on absolute order imbalance,
turnover, and their squared terms.

We find that the PIN model primarily identifies information events based on
turnover rather than absolute order flow imbalance. In regressions of $CPIE_{PIN}$ on
absolute order imbalance, turnover, and their squared terms, turnover and turnover
squared account for, on average, around 65% of the overall $R^2$. The identification of
information events through turnover becomes more pronounced late in the sample
with the increase in both the level and variance of turnover.\footnote{Duarte and Young (2009) propose an extension of the PIN model that accounts for the positive
correlation between buys and sells and thus improves the fit of the model. We show in Appendix
A that Duarte and Young (2009) model also performs poorly late in the sample.} For example, for the
median stock after 2002, the PIN model is essentially equivalent to a naive model
that sets the probability of a private information event equal to one on any day with
turnover larger than the yearly mean turnover and zero otherwise. This problem is
due to two limitations of the PIN model. First, under the PIN model, increases in
expected turnover can only come about through the arrival of private information.
Second, the PIN models’ restrictive distributional assumptions make it difficult for
the model to match both the mean and the variance of turnover. As a result of these
limitations, when confronted with actual data the model mechanically interprets
periods of high turnover as periods of private information arrival.

To show that this conflation of turnover with private information is related to
the previous critiques of PIN, we examine an event study similar in spirit to those
in various documented PIN anomalies. For instance, Benos and Johec (2007) find that \( PIN \) is higher after earnings announcements than before.\(^5\) In a similar vein, we examine how well the PIN model identifies information events around earnings announcements. In contrast to Benos and Johec (2007) however, we use \( CPIE_{PIN} \) to conduct this event study instead of \( PIN \). There is a large literature (see Bamber, Barron, and Stevens (2011) for a review) that shows that turnover is substantially higher around earnings announcements and typically remains high for a considerable period after the announcement. Since our concern here is the PIN model’s ability to separate turnover shocks from information events, earnings announcements provide a good opportunity to examine the model’s performance and allows us to connect our results with those in previous studies. As in our regressions above, our event study shows that \( CPIE_{PIN} \) is higher after the announcement simply due to the higher levels of turnover in the post-announcement period.

This mechanical conflation of increases in turnover with the arrival of private information in the PIN model is a problem because it implies that the most popular measure of private information in the literature, PIN, is a poor proxy for private information. There is no theoretical reason why turnover should be mechanically associated with the arrival of private information, once we control for order imbalance. On one hand, trading by informed traders may increase turnover. On the other hand, liquidity traders may postpone trading when the arrival of private information is likely leading to a negative relation between turnover and private information (e.g.

\(^5\)In addition, Aktas, de Bodt, Declerck, and Van Oppens (2007) find that \( PIN \) is higher after merger announcements than before.
Moreover, a model that naively associates turnover with private information arrival does not account for the fact that turnover varies for reasons unrelated to private information. For instance, turnover can increase with disagreement (e.g. Kandel and Pearson (1995), and Banerjee and Kremer (2010)). Turnover is also subject to calendar effects because traders coordinate trade on certain days to reduce trading costs (Admati and Pfleiderer (1988)). Furthermore, turnover can vary due to portfolio rebalancing (Lo and Wang (2000)) and taxation reasons (Lakonishok and Smidt (1986)).

Hence, the PIN model (and the PIN measure) groups all sources of variation in turnover (e.g. disagreement, calendar effects, portfolio rebalancing, taxation, etc.) under the umbrella of private information arrival.

Having demonstrated that the PIN model essentially treats all shocks to turnover as private information because it fits the data so poorly, we turn to our second research question, namely whether a model based on variables other than order flow alone can identify private information significantly better than a model based on order flow alone. To do so, we compare an extension of the PIN model (the EPIN model) with the model developed by Odders-White and Ready (2008) (the OWR model). The EPIN model is based on the same principles as the PIN model. However, the key difference is that the EPIN model does not associate all variation in turnover with private information arrival. The OWR model is based on Kyle

---

6The literature also suggests that turnover after earnings announcements can remain high for many reasons unrelated to the arrival of private information. For instance, traditionally, the literature attributes high turnover after announcements to disagreement (e.g. Bamber, Barron, and Stevens (2011)). Karppoff (1986) suggests that high turnover after earnings announcements may also be due to divergent prior expectations, while Frazzini and Lamont (2007) attribute to small investors' lack of attention.
(1985). In contrast to the PIN and EPIN models, the OWR model uses intraday and overnight returns, along with order imbalance, to identify private information events.

We use the EPIN and OWR CPIEs ($CPIE_{EPIN}$ and $CPIE_{OWR}$) to compare the models in three different ways. First, under the working hypothesis that private information should arrive prior to earnings announcements, rather than after the announcement, we expect that if a model correctly identifies informed trade, its CPIE will increase prior to the announcement. We also anticipate that informed trading, and hence CPIEs, will decline rapidly after the announcement, when investors have the same (now public) information. Second, we follow Cohen, Malloy, and Pomorski (2012) and identify instances of opportunistic insider trades. If either of the models can successfully detect opportunistic insider trading, then its CPIE should increase around these trades. Third, it has long been recognized in the literature (e.g. Hasbrouck (1988, 1991a,b)) that non-information related price changes (e.g. dealer inventory control, price pressure, price discreteness, etc.) should be subsequently reversed, while information related trades should not. Therefore, if a model correctly identifies the arrival of private information, we expect that increases in its CPIE should be associated with smaller future price reversals.

---

7While the PIN and EPIN models allow for a calculation of the probability of informed trade, the OWR model does not. However, all three models have a parameter that controls the unconditional probability of an information event on a given day ($\alpha$) and allow for the calculation of CPIE.

8There is considerable evidence suggesting the possibility of high asymmetric information prior to important announcements. See for example Brooks (1996), Meulbroek (1992) Christophe, Ferri, and Angel (2004), Amin and Lee (1997), Frazzini and Lamont (2007), and Hendershott, Livdan, and Schurhoff (2014)
In answer to our second question, we find that the OWR model performs better than the EPIN model in all three tests. Specifically, we find that the $CPIE_{OWR}$ increases before earnings announcements and decreases rapidly after announcements, while $CPIE_{EPIN}$ decreases before announcements. We also find that order flow imbalance explains only a small fraction of the variation in $CPIE_{OWR}$ in our event studies. Significantly, most of the variation in $CPIE_{OWR}$ is driven by variation in intraday and overnight returns and the interaction between them. $CPIE_{OWR}$ successfully predicts opportunistic insider trading and is strongly negatively associated with price reversals. On the other hand, $CPIE_{EPIN}$ increases only slightly during periods with opportunistic insider trading and is only weakly associated with price reversals.

We contribute to the literature because we show that private information measures based only on order flow (e.g. PIN) perform much worse than those that include variables other than order flow, for instance the OWR’s $\alpha$. The classic microstructure theories (e.g. Glosten and Milgrom (1985), and Kyle (1985)) suggest that order flow imbalances as well as variables such as returns, spreads, and price impacts are related to the arrival of private information. Hence it is not surprising that a model that identifies the arrival of private information solely from order flow imbalance has poorer performance than a model based on returns and order flows. However, it is perhaps surprising that the OWR model performs so much better than the EPIN model in all of our tests. This suggests that order flow, however well modeled, is insufficient to be the sole source of inferences about private information arrival. Therefore, despite the literature’s strong interest on proxies of private infor-
mation based on order flow alone (e.g. Easley, Kiefer, O’Hara, and Paperman (1996), Easley, Kiefer, and O’Hara (1997), and Duarte and Young (2009)) future research aimed at building measures of informed trade should also focus on variables such as returns, spreads, and price impacts (as the classic theory suggests.)

Our paper is also related to a growing literature that analyzes the extent to which $PIN$ actually captures information asymmetry.$^{9}$ In a contemporaneous paper, Gan, Wei, and Johnstone (2014) also show that the PIN model does not fit the order flow data well. We take this result one step further and show that because of this poor fit the PIN model mechanically identifies the arrival of private information from turnover. In addition, we extend the PIN model to correct the mechanical conflation of turnover and private information arrival. This allows us to address whether order flow alone can capture private information arrival or whether we must incorporate the price response mechanism as in the OWR model. Many of the papers analyzing the PIN measure attempt to do so by estimating PINs around events and testing whether PIN is higher before rather than after an announcement. These studies in general document that PIN is higher after announcements than before (i.e. PIN anomalies). For instance, Collin-Dufresne and Fos (2014a) find that PIN and other adverse selection measures are lower when Schedule 13D filers

---

$^{9}$A series of papers addresses the pricing of information asymmetry in the cross section of stock returns (e.g. Easley, Hvidkjaer, and O’Hara (2002), Duarte and Young (2009), Mohanram and Rajgopal (2009), Lambert and Leuz (2012), and Lai, Ng, and Zhang (2014)). In contrast to this paper, Duarte and Young (2009) question whether PIN is priced because it identifies private information or because it is related to illiquidity. We do not address the pricing of PIN, instead we aim to understand how the PIN model identifies private information.
Easley, Engle, O’Hara, and Wu (2008) critique this line of research, noting that PIN is a stock characteristic rather than a measure of the extent to which private information is present in a given calendar time period. To address this critique, Easley, Engle, O’Hara, and Wu (2008) develop an extension of the original model in which PIN is time-varying, and in a paper contemporaneous to ours, Brennan, Huh, and Subrahmanyam (2015) use conditional probabilities similar to CPIE\textsubscript{PIN}. We contribute to this literature in two ways. First, our results indicate that these previously identified PIN anomalies are at least partially related to the strong connection between CPIE\textsubscript{PIN} and turnover that we document. Second, we show that event studies that use daily measures of private information (e.g. Easley, Engle, O’Hara, and Wu (2008)) can be misleading if variation in these measures around event announcements is due to variables not necessarily related to information asymmetry. For instance, Brennan, Huh, and Subrahmanyam (2015) interpret the fact that their CPIE\textsubscript{PIN} measures are higher after earnings announcements than before as evidence of informed trading. We show that CPIE\textsubscript{PIN} is naively related to turnover. This suggests that the findings in Brennan, Huh, and Subrahmanyam (2015) can simply be attributed to the fact that turnover is typically much higher.

\textsuperscript{10}Collin-Dufresne and Fos (2014b) partially attribute this finding to informed traders disguising their trades in periods of high liquidity or timing their trades such that market movements conceal the nature of their information. Our findings cannot speak to this possibility, instead we show that the PIN model mechanically attributes all sources of variation in turnover to the arrival of private information.

\textsuperscript{11}Easley, Lopez de Prado, and O’Hara (2012) develop the volume-synchronized probability of informed trading or VPIN. We do not consider VPIN in this paper because, as Easley, Lopez de Prado, and O’Hara (2012) point out, VPIN is a measure of order flow toxicity at high frequencies rather than a stock characteristic that measures adverse selection at lower frequencies as PIN is widely used in the finance and accounting literature. Moreover, Andersen and Bondarenko (2014) provide detailed critique of the VPIN measure.
after earnings announcements.

The remainder of the paper is as follows. Section 3.1 outlines the data we use for our empirical results. Section 3.2 shows that the PIN model mechanically associates variation in turnover with the arrival of private information. Section 3.3 extends the PIN model to deal with this shortcoming and compares a model based on order flow imbalance alone (EPIN) with a model that identifies private information from both returns and order flow (OWR). Section 3.4 concludes.

3.1 Data

To estimate the PIN, EPIN, and OWR models, we collect trades and quotes data for all NYSE stocks between 1993 and 2012 from the NYSE TAQ database. We require that the stocks in our sample have only one issue (i.e. one PERMNO), are common stocks (share code 10 or 11), are listed on the NYSE (exchange code 1), and have at least 200 days worth of non-missing observations for the year. Our sample contains 1,060 stocks per year on average. Despite our sample selection criteria, about 36% (25%) of the stocks in our sample are in the top (bottom) three Fama-French size deciles. For each stock in the sample, we classify each day’s trades as either buys or sells, following the Lee and Ready (1991) algorithm. In our analysis, we define turnover as the sum of daily buys and sells. Internet Appendix B describes the computation of the number of buys and sells.

We estimate both the PIN and EPIN models using only the daily number of buys and sells ($B_{i,t}$ and $S_{i,t}$). The OWR model, however, also requires intraday and
overnight returns as well as order imbalances. Following Odders-White and Ready (2008) we compute the intraday return at day $t$ as the volume-weighted average price (VWAP) at $t$ minus the opening quote midpoint at $t$ plus dividends at time $t$, all divided by the opening quote midpoint at time $t$.\footnote{The opening quote midpoint is not available in TAQ in many instances. When the opening quote midpoint is not available, we use the matched quote of the first trade in the day as a proxy for the opening quote.} We compute the overnight return at $t$ as the opening quote midpoint at $t + 1$ minus the VWAP at $t$, all divided by the opening quote midpoint at $t$. The total return, or sum of the intraday and overnight returns is the open-to-open return from $t$ to $t + 1$. We compute order imbalance ($y_e$) as the daily share volume of buys minus the share volume of sells, divided by the total share volume. We follow Odders-White and Ready and remove systematic effects from returns to obtain measures of unexpected overnight and intraday returns ($r_{o,i,t}$ and $r_{d,i,t}$). Internet Appendix B describes how we compute $r_{o,i,t}$ and $r_{d,i,t}$.

Like Odders-White and Ready (2008), we remove days around unusual distributions or large dividends, as well as CUSIP or ticker changes. We also drop days for which we are missing overnight returns ($r_{o,i,t}$), intraday returns ($r_{d,i,t}$), order imbalance ($y_e$), buys ($B$), or sells ($S$). Our empirical procedures follow those of Odders-White and Ready with two exceptions. First, OWR estimate $y_e$ as the idiosyncratic component of net order flow divided by shares outstanding. We do not follow the same procedure as OWR in defining $y_e$ because we find that estimating $y_e$ as we do results in less noisy estimates. Specifically, we find that $y_e$ defined as shares bought minus shares sold divided by shares outstanding, as in Odders-White and Ready (2008), suffers from scale effects late in the sample, when order flow is
several orders of magnitude larger than shares outstanding. Second, Odders-White and Ready remove a whole trading year of data surrounding distribution events, but we only remove one trading week \([-2,+2]\) around these events.

For the event study portion of our analysis, we examine earnings announcements. Our sample of earnings announcements includes all CRSP/COMPSTAT firms listed in NYSE, NASDAQ, and AMEX between 1995–2009 for which we have exact timestamps collected from press releases in Factiva which fall within a \([-1,0]\) window relative to COMPSTAT earnings announcement dates following Dong et. al (2015). In the Odders-White and Ready model informed traders submit their orders during the day and their private information is revealed overnight (see Section 3.3.2 for more details). The earnings timestamps allow us to match the timing of the OWR model such that on \(t=0\) we know that the order imbalance \((y_e)\) and intraday returns \((r_d)\) are generated before the earnings numbers are publicly disseminated and are reflected in the overnight returns \((r_o)\).\(^{13}\) Our final sample of earnings announcements includes 21,979 events.

We also examine a sample of opportunistic insider trades, as defined in Cohen, Malloy, and Pomorski (2012), from the Thomson Reuters’ database of insider trades. In order to classify a trader as opportunistic or routine, we require three years of consecutive insider trades. We classify a trader as routine if she places a trade in the same calendar month for at least three years. All non-routine traders’ trades are classified as opportunistic. Cohen, Malloy, and Pomorski (2012) show that opportu-

\(^{13}\)For example, if earnings were announced between midnight and 9:30 AM EST before markets opened on a Tuesday in the COMPUSTAT database, then the event date would actually be Monday in the OWR model.
tunistic insider trades predict abnormal returns, information events, and regulator actions, which is consistent with the presence of private information. Our event sample includes 32,676 opportunistic insider trades.

Table 3.1 contains summary statistics of all the variables used to estimate the models. Panel A gives summary statistics of our entire sample, Panel B displays the summary statistics for the days of earnings announcements, and Panel C displays the summary statistics for opportunistic insider trading days.

### 3.2 Why does PIN fail?

This section addresses whether PIN fails because the underlying model does not fit the data well. Section 3.2.1 briefly describes the PIN model and $CPIE_{PIN}$. Section 3.2.2 shows the results of regressions of $CPIE_{PIN}$ on absolute order imbalance and turnover. Section 3.2.3 shows how $CPIE_{PIN}$ varies around earnings announcements. The results in Sections 3.2.2 and 3.2.3 show that the PIN model identifies the arrival of private information from increases in turnover.

#### 3.2.1 Description of the PIN model

The Easley, Kiefer, O’Hara, and Paperman (1996) PIN model posits the existence of a liquidity provider who receives buy and sell orders from both informed traders and uninformed traders. At the beginning of each day, the informed traders receive a private signal with probability $\alpha$. If the private signal is positive, buy orders from informed and uninformed traders arrive following a Poisson distribution with
Table 3.1: **Summary Statistics.** This table summarizes the full sample and event day (t=0) returns, order imbalance, and number of buys and sells. We compute intraday and overnight returns as well as daily buys and sells for stocks between 1993 and 2012 using data from the NYSE TAQ database. Following Odders-White and Ready (2008), we compute the intraday return, \( r_d \), at time \( t \) as the volume-weighted average price at \( t \) (VWAP) minus the opening quote midpoint at \( t \) plus dividends at time \( t \), all divided by the opening quote midpoint at time \( t \). We compute the overnight return, \( r_o \), at \( t \) as the opening quote midpoint at \( t+1 \) minus the VWAP at \( t \), all divided by the opening quote midpoint at \( t \). The total return, \( r \), is the sum of the intraday and overnight returns and is equal to the open-to-open return from \( t \) to \( t+1 \). We compute \( y_e \) as the daily total volume of buys minus total volume of sells, divided by the total volume. For the PIN and EPIN models, we use the daily total number of buys and sells. Our sample of earnings announcements includes all CRSP/COMPUSTAT firms listed in NYSE, NASDAQ, and AMEX between 1995–2009 for which we have exact timestamps collected from press releases in Factiva which fall within a \([-1,0]\) window relative to COMPUSTAT earnings announcement dates following Dong et. al (2015). Opportunistic insider trades are defined in Cohen et. al (2012).

(a) Full Sample

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_e )</td>
<td>5,286,191</td>
<td>2.766%</td>
<td>31.259%</td>
<td>-10.433%</td>
<td>3.282%</td>
<td>18.996%</td>
</tr>
<tr>
<td>( r_d )</td>
<td>5,286,191</td>
<td>-0.004%</td>
<td>1.500%</td>
<td>-0.707%</td>
<td>-0.024%</td>
<td>0.680%</td>
</tr>
<tr>
<td>( r_o )</td>
<td>5,286,191</td>
<td>0.003%</td>
<td>1.297%</td>
<td>-0.566%</td>
<td>-0.024%</td>
<td>0.525%</td>
</tr>
<tr>
<td># Buys</td>
<td>5,286,191</td>
<td>1.876</td>
<td>6.917</td>
<td>37</td>
<td>220</td>
<td>1,128</td>
</tr>
<tr>
<td># Sells</td>
<td>5,286,191</td>
<td>1.843</td>
<td>6.894</td>
<td>36</td>
<td>194</td>
<td>1,033</td>
</tr>
</tbody>
</table>

(b) Earnings Announcements

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_e )</td>
<td>21,979</td>
<td>5.099%</td>
<td>22.122%</td>
<td>-4.787%</td>
<td>4.373%</td>
<td>16.400%</td>
</tr>
<tr>
<td>( r_d )</td>
<td>21,979</td>
<td>0.002%</td>
<td>2.424%</td>
<td>-1.252%</td>
<td>-0.004%</td>
<td>1.271%</td>
</tr>
<tr>
<td>( r_o )</td>
<td>21,979</td>
<td>0.075%</td>
<td>2.313%</td>
<td>-1.042%</td>
<td>0.013%</td>
<td>1.153%</td>
</tr>
<tr>
<td># Buys</td>
<td>21,979</td>
<td>4.572</td>
<td>13.491</td>
<td>223</td>
<td>956</td>
<td>3,421</td>
</tr>
<tr>
<td># Sells</td>
<td>21,979</td>
<td>4.465</td>
<td>13.546</td>
<td>191</td>
<td>831</td>
<td>3,165</td>
</tr>
</tbody>
</table>

(c) Opportunistic Insider Trades

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_e )</td>
<td>32,676</td>
<td>4.980%</td>
<td>20.425%</td>
<td>-5.106%</td>
<td>3.874%</td>
<td>15.353%</td>
</tr>
<tr>
<td>( r_d )</td>
<td>32,676</td>
<td>0.151%</td>
<td>1.566%</td>
<td>-0.632%</td>
<td>0.086%</td>
<td>0.865%</td>
</tr>
<tr>
<td>( r_o )</td>
<td>32,676</td>
<td>0.056%</td>
<td>1.247%</td>
<td>-0.467%</td>
<td>0.020%</td>
<td>0.528%</td>
</tr>
<tr>
<td># Buys</td>
<td>32,676</td>
<td>3.852</td>
<td>10.645</td>
<td>354</td>
<td>1,129</td>
<td>3,478</td>
</tr>
<tr>
<td># Sells</td>
<td>32,676</td>
<td>3.787</td>
<td>10.554</td>
<td>300</td>
<td>996</td>
<td>3,303</td>
</tr>
</tbody>
</table>
intensity $\mu + \epsilon_B$, while sell orders come only from the uninformed traders and arrive with intensity $\epsilon_S$. If the private signal is negative, sell orders from informed and uninformed traders arrive following a Poisson distribution with intensity $\mu + \epsilon_S$, while buy orders come only from the uninformed traders and arrive with intensity $\epsilon_B$. If the informed traders receive no private signal, they do not trade; thus, all buy and sell orders come from the uninformed traders and arrive with intensity $\epsilon_B$ and $\epsilon_S$, respectively. Fig. 3.1 shows a tree diagram of this model.
Figure 3.1: **PIN Tree.** For a given trading day, private information arrives with probability $\alpha$. When there is no private information, buys and sells are Poisson with intensity $\epsilon_B$ and $\epsilon_S$. Private information is good news with probability $\delta$. The expected number of buys (sells) increases by $\mu$ in case of good (bad) news.

- **No Private Information**
  - Buys $\sim Poi(\epsilon_b)$
  - Sells $\sim Poi(\epsilon_s)$

- **Private Information**
  - Buys $\sim Poi(\epsilon_b + \mu)$
  - Sells $\sim Poi(\epsilon_s + \mu)$

- **Good News**
  - $\delta$

- **Bad News**
  - $1 - \delta$
The difference in arrival rates captures the intuition that on days with positive private information, the arrival rate of buy orders increases over and above the normal rate of noise trading because informed traders enter the market to place buy orders. Similarly, the arrival rate of sell orders rises when the informed traders seek to sell based on their negative private signals. Therefore, the PIN model identifies the arrival of private information through increases in the absolute value of the order imbalance.

The model also ties variations in turnover to the arrival of private information. Specifically, let the indicator $I_{i,t}$ take the value of one if an information event occurs for stock $i$ on day $t$, and zero otherwise. Note that under the model the number of buys plus sells (turnover) is distributed as a Poisson random variable with intensity:

$$\lambda(I_{i,t}) = \begin{cases} 
\epsilon_B + \epsilon_S & \text{when } I_{i,t} = 0 \\
\epsilon_B + \epsilon_S + \mu & \text{when } I_{i,t} = 1
\end{cases} \quad (3.1)$$

Thus, under the PIN model, private information is necessarily the cause of any variation in expected daily turnover.

To formalize the concept of $CPIE_{PIN}$, let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock $i$ on day $t$ and $\Theta_{PIN,i} = (\alpha_i, \mu_i, \epsilon_B, \epsilon_S, \delta_i)$ represent the vector of the PIN model parameters for stock $i$. Let $D_{PIN,i,t} = [\Theta_{PIN,i}, B_{i,t}, S_{i,t}]$. The likelihood function of the Easley, Kiefer, O’Hara, and Paperman (1996) model is $\prod_{t=1}^{T} L(D_{PIN,i,t})$, where $L(D_{PIN,i,t})$ is equal to the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information ($L_{NI}(D_{PIN,i,t})$) added to the likelihood of
Using the PIN model, for each stock-day, we compute the probability of an information event conditional both on the model parameters and on the observed total number of buys and sells. For the PIN model, we compute 

\[ CPIE_{PIN,i,t} = P[I_{i,t} = 1 | D_{PIN,i,t}] \]

This probability is given by

\[ (L_{I^-}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t})) / L(D_{PIN,i,t}) \].

\[ CPIE_{PIN,i,t} \] represents the econometrician’s posterior probability of an information event given the data observed on that day, and assuming that he or she knows the underlying model parameters.

Note that if we condition down with respect to the data, \( CPIE_{PIN,i,t} \) reduces to the model’s unconditional probability of information events \( (\alpha_i) \). The unconditional probability represents the econometrician’s beliefs about the likelihood of an information event before seeing any actual orders or trades. In the absence of buy and sell data, an econometrician would assign a probability \( \alpha_i \) to an information event for stock \( i \) on day \( t \), where \( \alpha_i = E[CPIE_{PIN,i,t}] \) and the expectation is taken with respect to the joint distribution of \( B_{i,t} \) and \( S_{i,t} \). The PIN of a stock, defined as \( \frac{\alpha_i}{\alpha \mu + \epsilon_B + \epsilon_S} \), is the unconditional probability that any given trade is initiated by an informed trader. \( CPIE \) and \( PIN \) are linked via the unconditional probability of an information event, \( \alpha \), which is also the unconditional expectation of \( CPIE \).

We estimate the PIN model numerically via maximum likelihood. For every
firm-year in our sample, we estimate the PIN model by maximizing the likelihood function above. The estimation procedure is similar to that used in Duarte and Young (2009). The parameter estimates are used for computing the CPIEs used in Sections 3.2.2 and 3.2.3. Internet Appendix C provide details about the maximum likelihood procedure and the calculation of CPIEs.

Table 3.2 contains summary statistics for the parameter estimates of the PIN model. Table 3.2 also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{PIN}$. The results in Table 3.2 show that the mean $CPIE_{PIN}$ behaves exactly like $\alpha$. Hence, changes in $CPIE_{PIN}$ and changes in the estimated alphas are analogous. Fig. 3.2 Panel A shows how the distribution of $\alpha$ changes over time. Interestingly, the PIN model $\alpha$ increases over time, with the median PIN $\alpha$ rising from about 30% in 1993 to 50% in 2012.\textsuperscript{14} Panels B of Fig. 3.2 plot the time series of $PIN$. Note that $PIN$ decreases over time in spite of the fact that $\alpha$ increases. This happens because, according to the PIN model, the intensity of noise trading is increasing over time while the intensity of informed trading remains flat as shown in Panels C of Fig. 3.2. It is important to note, however, that the time series pattern of the model parameters in Fig. 3.2 has no implication for how the PIN model identifies private information.

\textsuperscript{14}The increase in our estimated PIN model $\alpha$ parameters is somewhat larger than that in Brennan, Huh, and Subrahmanyam (2015). This small difference arises because Brennan, Huh, and Subrahmanyam (2015) have a larger number of stocks per year due to the fact that we apply sample filters similar to those in Odders-White and Ready (2008). In fact, without these filters, the increase in our estimated PIN model $\alpha$ parameters from 1993 to 2012 is comparable to that in Brennan, Huh, and Subrahmanyam (2015).
Table 3.2: **PIN Parameter Estimates.** This table summarizes parameter estimates of the PIN model for 21,206 PERMNO-Year samples from 1993 to 2012. \( \alpha \) represents the average unconditional probability of an information event at the daily level. \( \delta \) represents the probability of good news, and \( 1 - \delta \) represents the probability of bad news. \( \epsilon_B \) and \( \epsilon_S \) represent the expected number of daily buys and sells given no private information or symmetric order flow shocks. \( \mu \) represents the expected additional order flows given an information event. \( CPIE \) and \( \text{Std}(CPIE) \) are the PERMNO-Year mean and standard deviation of \( CPIE \).

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>21,206</td>
<td>0.372</td>
<td>0.122</td>
<td>0.291</td>
<td>0.375</td>
<td>0.445</td>
</tr>
<tr>
<td>( \delta )</td>
<td>21,206</td>
<td>0.607</td>
<td>0.209</td>
<td>0.484</td>
<td>0.625</td>
<td>0.762</td>
</tr>
<tr>
<td>( \epsilon_B )</td>
<td>21,206</td>
<td>1.625</td>
<td>5.388</td>
<td>33</td>
<td>193</td>
<td>1,039</td>
</tr>
<tr>
<td>( \epsilon_S )</td>
<td>21,206</td>
<td>1.596</td>
<td>5.369</td>
<td>35</td>
<td>186</td>
<td>956</td>
</tr>
<tr>
<td>( \mu )</td>
<td>21,206</td>
<td>312</td>
<td>593</td>
<td>43</td>
<td>160</td>
<td>314</td>
</tr>
<tr>
<td>( CPIE )</td>
<td>21,206</td>
<td>0.382</td>
<td>0.135</td>
<td>0.293</td>
<td>0.379</td>
<td>0.449</td>
</tr>
<tr>
<td>( \text{Std}(CPIE) )</td>
<td>21,206</td>
<td>0.451</td>
<td>0.052</td>
<td>0.427</td>
<td>0.470</td>
<td>0.490</td>
</tr>
</tbody>
</table>
Figure 3.2: **PIN Parameters.** This figure shows the distribution of yearly α, PIN, and μ, ε_B, ε_S parameter estimates for the PIN model. The solid black line represents the median value, and the dotted lines represent the 5, 25, 75, and 95 percentiles.

(a) PIN α  

(b) PIN

(c) PIN Parameters
We also estimate the parameter vectors $\Theta_{PIN,i}$ in the period $t \in [-312, -60]$ before an earnings announcement. These parameter estimates are used to compute the CPIEs in Section 3.2.3. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 3.2 and in Figure 3.2.

3.2.2 How does the PIN model identify private information?

In theory, the PIN model identifies information events from changes in the absolute order flow imbalance. In practice, however, the PIN model may produce a poor description of the data, which can affect the way it actually identifies private information.

To analyze how the PIN model identifies private information in practice, we compare results from data created by simulating the PIN model to results from real data. To create the simulated data, we first estimate the parameters of the PIN model for each firm-year in our sample. Then, for each firm-year, we generate 1,000 artificial firm-years’ worth of data (i.e. $B_{i,t}$ and $S_{i,t}$) using the estimated parameters. We then compute the $CPIE_{PIN,i,t}$ for each trading day in a simulated trading year and regress these CPIEs on the variables that should, in theory, identify information events in the PIN model (absolute order flow imbalance). The results of the regressions using simulated data reveal how the PIN model should perform if the data conform to the model. The simulated data regressions also allow us to build empirical distributions of the $R^2$s of the regressions of CPIEs on order imbalance. We use the empirical distribution of the $R^2$s to test the null hypothesis that the real data conform to the PIN model.
Panel A of Table 3.3 presents the results of yearly multivariate regressions of $CPIE_{PIN}$ on absolute order flow imbalance ($|B - S|$) and $|B - S|^2$. We add squared terms to these regressions to account for nonlinearities in the relation between $CPIE_{PIN}$ and $|B - S|$. We average the simulated results for each PERMNO-Year and report in Panel A of Table 3.3 the median coefficient estimates and $t-$statistics. The coefficients are standardized so they represent the increase in $CPIE_{PIN}$ due to a one standard deviation increase in the corresponding independent variable. We also report the average of the median, the 5$^{th}$, and the 95$^{th}$ percentiles of the empirical distribution of $R^2$s of these regressions generated by the 1,000 simulations. In general, the coefficients are highly statistically significant, and the $R^2$s are high, confirming the theoretical implication that absolute order imbalance can be used to infer the arrival of private information under the PIN model.
Table 3.3: PIN Model Regressions. This table reports real and simulated regressions of the $CPIE_{PIN}$ on absolute order imbalance ($|B - S|$), and order imbalance squared ($|B - S|^2$). In Panel A, we simulate 1,000 instances of the PIN model for each PERMNO-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the $R^2$ ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the $R^2$ attributed to $turn$ and $turn^2$ in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median $R^2$ values, and tests of the hypothesis that the observed variation in $CPIE_{PIN}$ is consistent with the PIN model. The $p$-value of $R^2$ ($R^2_{inc.}$) is the probability of observing an $R^2$ at least as small (large) as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

|   | $|B - S|$ | $|B - S|^2$ | $|B - S|$ | $|B - S|^2$ | 5%  | 50%  | 95%  | 5%  | 50%  | 95%  |
|---|----------|------------|----------|------------|-----|------|------|-----|------|------|
| 1993 | 0.437 | -0.079 | (10.31) | (-1.80) | 71.13% | 76.09% | 80.38% | 7.17% | 10.57% | 15.25% |
| 1994 | 0.422 | -0.072 | (9.63) | (-1.67) | 67.49% | 73.26% | 78.11% | 9.39% | 13.47% | 18.55% |
| 1995 | 0.410 | -0.058 | (9.68) | (-1.36) | 70.32% | 75.39% | 79.85% | 7.64% | 11.39% | 16.02% |
| 1996 | 0.432 | -0.085 | (9.89) | (-1.90) | 69.02% | 74.28% | 78.87% | 8.32% | 12.17% | 16.97% |
| 1997 | 0.450 | -0.089 | (10.30) | (-1.98) | 71.99% | 76.93% | 81.12% | 7.36% | 10.76% | 14.79% |
| 1998 | 0.482 | -0.104 | (10.79) | (-2.36) | 74.32% | 78.71% | 82.46% | 6.65% | 9.53% | 13.30% |
| 1999 | 0.484 | -0.112 | (11.03) | (-2.47) | 75.62% | 79.96% | 83.46% | 6.49% | 9.36% | 12.92% |
| 2000 | 0.529 | -0.137 | (11.88) | (-3.00) | 79.78% | 83.36% | 86.15% | 4.98% | 7.47% | 10.45% |
| 2001 | 0.638 | -0.217 | (13.97) | (-4.61) | 83.34% | 86.13% | 88.57% | 4.17% | 6.00% | 8.35% |
| 2002 | 0.695 | -0.260 | (14.11) | (-5.30) | 82.61% | 85.53% | 88.06% | 4.83% | 6.92% | 9.54% |
| 2003 | 0.665 | -0.244 | (12.38) | (-4.52) | 78.88% | 82.36% | 85.36% | 7.90% | 10.56% | 13.79% |
| 2004 | 0.650 | -0.223 | (11.49) | (-4.16) | 77.84% | 81.38% | 84.59% | 8.92% | 11.67% | 15.03% |
| 2005 | 0.658 | -0.220 | (12.59) | (-4.46) | 80.47% | 83.59% | 86.45% | 7.60% | 10.09% | 12.95% |
| 2006 | 0.650 | -0.221 | (11.96) | (-4.35) | 80.31% | 83.36% | 86.18% | 7.76% | 10.29% | 13.50% |
| 2007 | 0.632 | -0.222 | (9.40) | (-4.07) | 79.72% | 83.35% | 86.15% | 8.53% | 10.93% | 14.05% |
| 2008 | 0.666 | -0.235 | (12.29) | (-4.83) | 82.44% | 85.25% | 88.00% | 6.83% | 9.15% | 11.78% |
| 2009 | 0.709 | -0.269 | (14.37) | (-5.70) | 84.29% | 86.87% | 89.20% | 6.22% | 8.28% | 10.57% |
| 2010 | 0.704 | -0.261 | (14.60) | (-5.68) | 84.99% | 87.41% | 89.64% | 5.66% | 7.55% | 9.89% |
| 2011 | 0.671 | -0.234 | (14.13) | (-5.21) | 85.91% | 88.25% | 90.21% | 5.34% | 7.28% | 9.39% |
| 2012 | 0.693 | -0.251 | (14.92) | (-5.62) | 85.68% | 87.98% | 90.34% | 5.22% | 7.22% | 9.50% |
Table 3.3: PIN Model Regressions. Continued.

(b) Real Data

|       | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ | $|B - S|/|B - S|^2$ |
|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1993  | 0.300           | -0.073          | (5.98)          | (-1.43)         | 35.76%          | 0.39%           | 98.81%          | 36.20%          | 2.57%           |
| 1994  | 0.264           | -0.047          | (5.28)          | (-0.92)         | 32.82%          | 0.39%           | 98.02%          | 40.02%          | 3.36%           |
| 1995  | 0.280           | -0.061          | (5.77)          | (-1.29)         | 34.20%          | 0.73%           | 96.98%          | 36.97%          | 5.05%           |
| 1996  | 0.277           | -0.065          | (5.69)          | (-1.28)         | 30.92%          | 0.51%           | 98.46%          | 38.97%          | 3.85%           |
| 1997  | 0.283           | -0.073          | (5.67)          | (-1.36)         | 30.80%          | 0.35%           | 99.05%          | 38.86%          | 3.54%           |
| 1998  | 0.274           | -0.059          | (5.26)          | (-1.09)         | 30.12%          | 0.24%           | 99.31%          | 39.58%          | 3.54%           |
| 1999  | 0.280           | -0.059          | (5.21)          | (-1.08)         | 29.05%          | 0.18%           | 99.38%          | 39.46%          | 3.29%           |
| 2000  | 0.300           | -0.079          | (5.48)          | (-1.39)         | 29.99%          | 0.17%           | 99.73%          | 39.08%          | 2.59%           |
| 2001  | 0.339           | -0.111          | (5.67)          | (-1.87)         | 29.44%          | 0.17%           | 99.71%          | 39.39%          | 3.53%           |
| 2002  | 0.279           | -0.058          | (4.09)          | (-0.85)         | 23.05%          | 0.10%           | 99.82%          | 44.28%          | 5.59%           |
| 2003  | 0.247           | -0.032          | (3.57)          | (-0.47)         | 21.97%          | 0.17%           | 99.73%          | 41.86%          | 9.55%           |
| 2004  | 0.211           | -0.005          | (3.14)          | (-0.08)         | 19.55%          | 0.00%           | 100.00%         | 45.22%          | 8.78%           |
| 2005  | 0.254           | -0.053          | (3.81)          | (-0.81)         | 19.42%          | 0.38%           | 99.46%          | 46.29%          | 9.21%           |
| 2006  | 0.251           | -0.066          | (3.80)          | (-0.96)         | 16.95%          | 1.26%           | 97.86%          | 48.44%          | 10.83%          |
| 2007  | 0.271           | -0.104          | (4.01)          | (-1.57)         | 14.30%          | 2.47%           | 95.62%          | 50.32%          | 14.04%          |
| 2008  | 0.268           | -0.111          | (4.00)          | (-1.66)         | 13.78%          | 1.93%           | 96.34%          | 50.97%          | 11.49%          |
| 2009  | 0.280           | -0.117          | (4.15)          | (-1.74)         | 14.59%          | 1.88%           | 96.79%          | 49.91%          | 10.08%          |
| 2010  | 0.291           | -0.124          | (4.39)          | (-1.82)         | 15.96%          | 1.96%           | 96.23%          | 47.64%          | 10.62%          |
| 2011  | 0.295           | -0.131          | (4.56)          | (-2.03)         | 15.94%          | 1.01%           | 97.34%          | 46.60%          | 11.14%          |
| 2012  | 0.319           | -0.145          | (4.96)          | (-2.23)         | 17.56%          | 3.35%           | 94.75%          | 45.61%          | 13.31%          |

The columns of Table 3.3 labeled as ‘$R_{inc}^2$’ include statistics on the increase in the $R^2$ that is due to the inclusion of turnover (turn) and turnover squared ($turn^2$) in the regressions. $R_{inc}^2$ is equal to the difference between the $R^2$ of the extended regression model with turnover terms and the $R^2$ of the regression with only order imbalance terms. We report the average of the median, the 5th, and the 95th percentiles of the $R_{inc}^2$s of these regressions across the 1,000 simulations. The incremental increase in $R^2$s are relatively low, with an average value of around 10%, which implies that, under the data generating process of the model, turnover has only modest power in explaining $CPIE_{PIN}$ once controlling for absolute order imbalance and its square.
The picture that emerges from these regressions is that if the PIN model were a perfectly accurate representation of trading activity, $CPIE_{PIN}$ would be determined solely by the order flow imbalance on each day. Since the order flow imbalance and turnover on any given day are highly positively correlated, turnover adds little to explaining $CPIE_{PIN}$ in theory.

Panel B of Table 3.3 reports regression results for the real rather than simulated data. With the real data, the picture is very different. The $R^2$s of the regressions of $CPIE_{PIN}$ on $|B - S|$ and $|B - S|^2$ are much smaller than those in the simulations. We test the hypothesis that the real data $R^2$s and $R^2_{inc.}$s are consistent with those generated by the PIN model. Panel B reports the average across all stocks’ $p$-value (the probability of observing an $R^2$ in the simulations at least as small as what we observe in the data), and the frequency that we reject the null at the 5% level implied by the distribution of simulated $R^2$s. The PIN model is rejected in about 98% of the stock-years in our sample, and there is on average less than a 1% chance of the PIN model generating $R^2$s as low as what we see in the data. On the other hand, the incremental $R^2$s of turnover are much higher than those in the Panel A. The incremental $R^2$ increases over time with a value of about 36% in 1993, but nearly 46% in 2012. This implies that turnover and turnover squared explain a much larger degree of variation in $CPIE_{PIN}$ than order imbalance. In fact, the average ratio of the median $R^2$s, $R^2/(R^2 + R^2_{inc.})$, is about 65%. The difference arises because, in the real data, absolute order flow and turnover are only weakly correlated. For instance, large absolute order flow imbalances are possible when turnover is below average, and vice versa. Under the PIN model, however, the two are highly correlated. The results
in Table 3.3 indicate that the PIN model frequently identifies private information in periods of increased turnover, as opposed to periods of large order imbalances. This is especially true in the later portion of the sample, as order imbalance becomes less important in explaining $CPIE_{PIN}$ over time.

To understand our results, we plot in Fig. 3.3 simulated and real order flow data along with $CPIE_{PIN}$ as function of turnover for Exxon-Mobil during 1993 and 2012. Panels A and B plot simulated and real order flow for Exxon-Mobil in 1993 and 2012 respectively, with buys on the horizontal axis and sells on the vertical axis. Real data are marked as +, and simulated data as transparent dots. The real data are shaded according to the model-specific $CPIE$, with darker points (+) representing low and lighter points (+) high $CPIEs$. Panel C and D plot the $CPIE_{PIN}$ as function of turnover. The vertical lines in these panels represent the mean turnover in the year.

Panel A of Fig. 3.3 illustrates the central intuition behind the PIN model. The simulated data comprise three types of days, which create three distinct clusters. Two of the clusters are made up of days characterized by relatively large order flow imbalance, with a large number of sells (buys) and relatively few buys (sells). The third group of days has relatively low numbers of buys and sells; these days have no private information arrival. Generalizing from this figure, days with large order flow imbalances (i.e. absolute value of the number of buys minus the number of sells) are likely to correspond to informed traders entering the market. The real data, on the other hand, show no distinct clusters in Panel A, and in Panel B of Fig. 3.3 the PIN model’s three clusters barely fit even a small minority of the data. This implies that the model cannot account for existence of the majority of the daily observations
Figure 3.3: XOM EO. This figure compares the real and simulated data for XOM in 1993 using the PIN model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{PIN}$, with darker markers (+) representing high and lighter markers (+) low $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark “cloud” of points, and low probability states appear as a light “cloud” of points. The PIN model has three states: no news, good news, and bad news. Panels C and D plot the CPIE values for the real data as a function of turnover along with a dotted vertical line indicating the mean turnover for the year.

(a) XOM 1993  

(b) XOM 2012  

(c)  

(d)
of order flow for Exxon-Mobil in 2012. In essence, the model classifies almost all daily observations as extreme outliers. The intuition for this is that the PIN model assumes that order flow is distributed as a mixture of three bivariate Poisson random variables (i.e. the three clusters in Panels A and B). The mean and the variance of a Poisson random variable are equal and, as a consequence, the Poisson mixtures behind the PIN model cannot accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample.

Panel C and D show that, the PIN model essentially classifies days with above average turnover as private information days (i.e. $CPIE_{PIN}$ equal to one) and days with below average turnover as days without private information (i.e. $CPIE_{PIN}$ equal to zero) with no intermediate values. While this identification of information events is not as stark in 1993 as in 2012, a large number of information event days, even in 1993, are identified according to this rule. The reason for this is that under the PIN model expected turnover varies only because of the arrival of private information (see Equation 3.1). Hence the poor fit to the turnover data along with the connection between turnover and arrival of private information in the PIN model causes the model to mechanically identify shocks to turnover as the arrival of private information.

Fig. 3.3 also emphasizes the mechanical nature of the relation between $CPIE_{PIN}$ and turnover. In 2012, the PIN model essentially identifies almost all days with higher than average turnover as being days with private information events. Note that this identification does not necessarily relate to the possibility, suggested by Collin-Dufresne and Fos (2014b), that informed traders sometimes choose to trade
in days with high liquidity or turnover. Naturally, it is possible that informed traders do in fact trade in some days with high turnover. However, the PIN model identifies all high turnover days as information events.

In addition to being mechanical, the identification of information events under the PIN model is naive because the PIN model groups all sources of variation in turnover (e.g. disagreement, calendar effects, portfolio rebalancing, taxation, etc.) under the umbrella of private information arrival. Although Fig. 3.3 is a highly stylized example of the PIN model’s naive identification of private information events, the problem is widespread. To quantify how often the PIN model classifies information events as simple function of turnover we define

\[
CPIE_{\text{Naive},i,t} = \begin{cases} 
0, & \text{if } turn_{i,t} < \overline{turn_i} \\
1, & \text{if } turn_{i,t} \geq \overline{turn_i}
\end{cases} 
\] (3.2)

That is, \(CPIE_{\text{Naive},i,t}\) is a dummy variable equal to one when turnover for stock \(i\) on day \(t\) \((turn_{i,t})\) is larger or equal to the average turnover of stock \(i\) for the entire calendar \((\overline{turn_i})\) year and zero otherwise. To our knowledge there is no paper in the literature that proposes identifying private information in similar manner.\(^{15}\) It is clear, however, from Panel D of Fig. 3.3 that the PIN model essentially identifies the arrival of private information for Exxon-Mobil in 2012 according to this rule. We use \(CPIE_{\text{Naive}}\) to gauge the extent to which the PIN model conflates the arrival

\(^{15}\)Stickel and Verrecchia (1994) propose identifying information arrival in general with a similar measure, but not private information in particular.
of private information with turnover. Specifically, Panel A of Fig. 3.4 shows the distribution of the fraction of days for which $CPIE_{PIN}$ is essentially identical to $CPIE_{Naive}$ ($|CPIE_{PIN} - CPIE_{Naive}| < 10^{-10}$). $CPIE_{PIN}$ and $CPIE_{Naive}$ are identical for about 85% of the annual observations for the median stock since 2002.

Another way to gauge the extent to which the PIN model breaks down later in our sample period is to count the number of days that the PIN model classifies as outliers. Panel B of Fig. 3.4 shows the fraction of days for the median stock-year which the PIN model classifies as “outliers” (likelihoods smaller than $10^{-10}$). According to the PIN model, for the median stock about 60% (90%) of the annual observations are classified as outliers in 2005 (2010).
Figure 3.4: **Breakdown of the PIN Model.** Panel A shows the distribution of the percent of trading days in a year in which the PIN model identifies private information essentially in the same way as the naive identification scheme. That is, Panel A plots the percentage of days where the $|CPIE_{PIN} - CPIE_{Naive}| < 10^{-10}$. Panel B shows the distribution of the percent of days where the total likelihood, given the model parameters and observed order flow data is less than $10^{-10}$—days, according to the model, with near-zero probability of occurring. $CPIE_{Naive}$ is based on a naive scheme for identifying private information arrival. Specifically, $CPIE_{Naive}$ is one for a given stock-day if turnover is higher than average yearly turnover, and is zero otherwise. The solid black line represents the median stock, and the dotted lines represent the 5, 25, 75, and 95 percentiles.

(a) Days with $CPIE_{PIN} \approx CPIE_{Naive}$

(b) Days with Near-Zero Probability
Figs. 3.3 and 3.4 also give the intuition for why the median PIN α increases over time in Fig. 3.2. To see this, recall that α is the unconditional expected value of CPIEPIN. Therefore, as we observe more CPIEPIN values approaching one, the estimated PIN α must increase. In fact, the median PIN α becomes close to 50% later in the sample which consistent with the fact that the PIN model assigns a CPIEPIN equal to one (zero) to days with turnover above (below) the average.

Given the strong connection between CPIEs and the unconditional probability of information arrival, (α) these results call into question the use of PIN as proxy for private information. While there are other parameters in the model (i.e. μ, εB and εS), these parameters are jointly identified with α. Hence it seems extremely unlikely that in the joint identification of the model parameters, biases in the other parameters ‘correct’ the biases in α in such a way that PIN is ‘rescued’ as a reasonable proxy for private information. Thus, while our CPIE results do not speak directly to μ, εB and εS, they still call into question PIN as a measure of private information. Moreover, it seems unlikely that PIN can possibly measure private information later in the sample if the model on which it is based naively identifies the arrival of private information from turnover.16

16O’Hara, Yao, and Ye (2014) find that high-frequency trading is associated with an increase in the use of odd lot trades, which do not appear in the TAQ database. Therefore, estimates of the PIN model parameters computed using recent TAQ data may be systematically biased. More broadly, Fig. 3.4 indicates that even if the PIN model are estimated using data that includes odd lot trades, the model will still be badly misspecified late in the sample.
3.2.3 Explaining PIN anomalies with turnover

The previous section shows that the PIN model often identifies private information from turnover. The question remains, however, whether this is merely an inconsequential specification issue or whether this changes the interpretation of results in the existing literature (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), Brennan, Huh, and Subrahmanyam (2015), and Easley, Engle, O’Hara, and Wu (2008)). To address this, we examine how well the PIN model identifies information events around earnings announcements. Turnover is typically much higher around earnings announcements (e.g. Bamber, Barron, and Stevens (2011)) hence earnings announcements provide a good laboratory to examine this question.

Unlike a standard event study, we focus on movements in \( CPIE \) rather than price movements. For each model, we examine the period \( t \in [-20, 20] \) around the event. To do so, we estimate the parameter vector \( \Theta_{PIN,i} \) in the period \( t \in [-312, -60] \) before the event and then compute the daily \( CPIEs \) for the period \( t \in [-20, 20] \) surrounding the announcement. Prior studies estimate the parameters of the model in various windows around an event in order to compute the \( PIN \). Our procedure is different in that we estimate the parameters of the model one year prior to the event and then employ the estimated parameters as if we were market makers observing the market data (i.e. buys and sells) and attempting to infer whether an information event occurred. Table 3.1 Panel B presents summary statistics for order imbalance, intraday returns, overnight returns, number of buys, and the number of sells for earnings announcement days \( (t = 0) \).
Panel A of Fig. 3.5 shows the average $CPIE_{PIN}$ in event time for our sample of earnings announcements. The graph shows that, under the PIN model, the probability of an information event increases prior to the event, starting below 55% 20 days before the announcement and peaking above 80% on the day after the announcement. The rise in the probability of an information event prior to the announcement is consistent with a world where informed traders generate signals about earnings and trade on this information before earnings are announced to the public. However, $CPIE_{PIN}$ is also higher after the actual earnings become public information.

Panels B and C of Fig. 3.5 shed light on the features of the data that produce the observed pattern in the average $CPIE_{PIN}$ in Panel A. Panel B shows the average predictions from OLS regressions of $CPIE_{PIN}$ on order imbalance and absolute order imbalance squared across all of the stocks in the event study sample. The solid line indicates that order imbalance explains only a small fraction of the movement in $CPIE_{PIN}$ during the event window. Panel C shows the average predictions from regressions of $CPIE_{PIN}$ on turnover and turnover squared. The solid line indicates that the variation in $CPIE_{PIN}$ around earnings announcements is explained almost entirely by turnover. The intuition follows directly from the results in Section 3.2.2, which illustrates that $CPIE_{PIN}$ is mechanically driven by turnover increases. The higher post-event turnover levels are enough to keep $CPIE_{PIN}$ above its pre-event mean for a substantial period.

To formalize the intuition behind Panels B and C of Fig. 3.5, we run regressions similar to those in Table 3.3 using our event sample. Specifically, we run regressions of $CPIE_{PIN}$ on absolute value of order imbalance and its squared term during the
Figure 3.5: Earnings Announcements - PIN. Panel A shows the average $CPIE_{PIN}$ for the PIN model in event time surrounding earnings announcements. Panels B and C compare the average $CPIE_{PIN}$ with the $CPIE_{PIN}$ predicted with either the absolute order imbalance or turnover, respectively. To obtain the predictions, we run regressions of daily $CPIE_{PIN}$ on $|B - S|$ or $\text{turn}$, and their respective squared terms.

(a) $CPIE_{PIN}$

(b) Prediction using $|B - S|$ and $|B - S|^2$

(c) Prediction using $\text{turn}$ and $\text{turn}^2$
event window [-20, +20]. The results of these regressions (see Table 3.4) indicate that absolute order imbalance explains little of the variation in $CPIE_{PIN}$ in the event window while turnover explains most of the variation in $CPIE_{PIN}$. In fact, Table 3.4 shows that for the median stock, adding turnover and turnover squared to these regressions nearly quadruples the $R^2$s.
Table 3.4: **PIN Regressions Around Earnings Announcements.** This table reports regression results for $CPIE_{PIN}$ around Earnings Announcements. For each announcing firm in our sample we run regressions of $CPIE_{PIN}$ on absolute order imbalance ($|B - S|$) and absolute order imbalance squared ($|B - S|^2$) from $[-20, +20]$ and report median estimates across all the events. We compute the incremental $R^2_{inc.}$ as the increase in $R^2$ attributed to $turn$ and $turn^2$ in an extended regression model. We report standardized coefficients.

| $|B - S|$ | $|B - S|^2$ | $|B - S|$ | $|B - S|^2$ | $R^2$ | $R^2_{inc.}$ |
|--------|-----------|--------|-----------|------|------------|
| 0.143  | -0.032    | (1.07) | (-0.35)   | 15.42% | 44.44%    |
The event study results suggest that the variation in $PIN$ around events documented in the literature could be due to variation in $\alpha$ that is mechanically driven by turnover, rather than order imbalance. For instance, Benos and Jochec (2007) show that $PIN$ increases after earnings announcements, while Aktas, de Bodt, Declerck, and Van Oppens (2007) show that $PIN$ increases after M&A target announcements due to increases in both $\mu$ and $\alpha$. Therefore, our evidence suggests that these $PIN$ results are at least partially explained by the fact that the PIN model attributes increases in turnover to private information.

Turnover around earnings announcement can vary for many reasons unrelated to the arrival of private information. For instance, traditionally, the literature has attributed high turnover after announcements to disagreement (e.g. Bamber, Barron, and Stevens (2011)). Karpoff (1986) suggests that high turnover after earnings announcements may also be due to divergent prior expectations, while Frazzini and Lamont (2007) attributes to small investors’ lack of attention. None of these studies suggest that the higher turnover around announcements is necessarily the result of increased informed trade, per se. Even the PIN model suggests that once we control for order imbalance, turnover should have little power to identify informed trade.

Another important implication of these results for the literature is that event studies based on daily measures of private information, like $CPIE_{PIN}$ (e.g. Easley, Engle, O’Hara, and Wu (2008) and Brennan, Huh, and Subrahmanyam (2015)) can also be misleading. To see this point consider the results in Panel A of Fig. 3.5. It may appear at first glance that the results in Panel A of Fig. 3.5 suggest that the PIN model identifies private information in a sensible way since $CPIE_{PIN}$ increases
dramatically from 55% before the announcement to over 75% on the day of the announcement then falls after the announcement, albeit over a period of weeks. However, the decomposition of the CPIEs in Panels B and C of Fig. 3.5 points to a different interpretation, namely that the dramatic increase in CPIE around the event is actually result of variation in turnover, and variations in turnover may not be due to arrival of private information as we point out above.

3.3 Does order flow alone reveal private information?

The previous section shows that a mechanical conflation of private information with turnover plagues the PIN model. However, it could be that net order flow itself is such a poor indicator of private information that no model based on order flow alone is capable of identifying informed trade (e.g. Back, Crotty, and Li (2014) and Kim and Stoll (2014)) . This section gauges the extent to which a model involving variables other than order flow generates better inferences about the arrival of informed trade than a model based on order flow alone. To do so, we first propose an extension of the PIN model (the EPIN model) that removes the mechanical conflation of turnover and arrival of private information that plagues the PIN model. We then compare the OWR model, which infers the arrival of private information from returns and order flow, with the EPIN model, which is solely based on order flow. Section 3.3.1 presents the EPIN model. Section 3.3.2 describes the OWR model and Section 3.3.3 presents the results of a horse race between the OWR and the EPIN models.
3.3.1 Extending the PIN model

Our results in Sections 3.2.2 and 3.2.3 show that the PIN model naively identifies information events based on turnover. This happens because of two limitations of the PIN model. First, under the PIN model, increases in expected turnover can only come about through the arrival of private information (see Equation 3.1). Second, the PIN model assumes that order flow is distributed as a mixture of three bivariate Poisson random variables (i.e. the three clusters in Panels A and B of Fig. 3.3). This assumption is too restrictive to accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample. In this section, we propose an extension of the PIN model to fix the issues with the PIN model.

Before doing so, it is useful to formalize why the model fails in the way that we discuss above. Panel A of Fig. 3.6 displays a reparameterization of the PIN model in terms of the intensity of the number of trades (buys plus sells). Specifically, we parameterize the PIN model in terms of the intensity of total number trades λ. Recall that Equation 3.1 defines λ such that on days without private information λ(0) = ε_B + ε_S and, on days with private information, λ(1) = λ(0) × (1 + η), where η = μ/λ(0). The bottom node of Panel A in Fig. 3.6 shows that, on days without private information, the ratio of the intensity of buyer initiated trades to the intensity of the total number of trades is θ = ε_B/λ(0), while the intensity of seller initiated trades to the intensity of the number of trades (1 − θ) = ε_S/λ(0). The central node of Panel A in Fig. 3.6 shows that on negative private information days, the ratio of the intensity of buys to the intensity of the total number of trades
drops to $\theta/(1 + \eta)$. Similarly, the top node Panel A of Fig. 3.6 shows that on days with positive private information, the ratio of the intensity of sells to the intensity of the total number of trades drops to $(1 - \theta)/(1 + \eta)$. Essentially, Panel A of Fig. 3.6 shows a parameterization of the PIN model using the parameters $\lambda(0)$, $\eta$, and $\theta$ instead of $\epsilon_B$, $\epsilon_S$, and $\mu$. 


Figure 3.6: EPIN Tree. Panel A presents a reparameterization of the PIN model in terms of the ratio, \(\theta = \epsilon_B/(\epsilon_B + \epsilon_S)\), the intensity of the number of trades in non-information days, \(\lambda(0) = \epsilon_B + \epsilon_S\), and the percent increase in the intensity of the number of trades due to the arrival of private information, \(\eta = \lambda(1)/\lambda(0) - 1\) where \(\lambda(1) = \lambda(0) \times (1 + \mu)\). Panel B presents the EPIN model. The EPIN model extends the PIN model by allowing the intensity of the number of trades on a given day \(t(\lambda_t)\) to be drawn from a Gamma distribution with location and scale parameters \(r\) and \(p/(1-p)\), respectively. The information structure remains the same as the one in the PIN model. For a given trading day, private information arrives with probability \(\alpha\). When there is no private information, the number of buys (sells) is distributed as a Poisson with intensity \(\theta \times \lambda_t ((1 - \theta) \times \lambda_t)\). Private information is good (bad) news with probability \(\delta (1 - \delta)\). When there is good news, the number of sells (buys) is Poisson with intensity \(\frac{(1 - \theta)}{1 + \eta} \lambda_t \left(1 - \frac{(1 - \theta)}{1 + \eta}\right)\lambda_t\). When there is bad news, the number of buys (sells) is Poisson with intensity \(\frac{\theta}{1 + \eta} \lambda_t \left(1 - \frac{\theta}{1 + \eta}\right)\lambda_t\).
The limitations of the PIN model are clearly formalized with the parameterization in Panel A of Fig. 3.6. To wit, the PIN model does not allow for enough variability in $\lambda$ to accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample. Moreover, increases in $\lambda$ can only come about through the arrival of private information. That is, $\lambda$ is function of information arrival ($I_t$).

Hence one way to fix the limitations of the PIN model while keeping its private information arrival structure is to draw $\lambda_t$ independently of the arrival of private information. In the extended model (EPIN model), the market maker still sees only the number of buys and sells on a given day. However, the market maker knows that the unobservable overall intensity of trade, $\lambda_t$, varies from day to day. The market maker also knows that private information arrives with probability $\alpha$. If private information is positive (negative), then the ratio of expected number of buys to the expected total number of trades rises (falls). In this way, we keep the underlying information structure of the PIN model, but allow the market maker to take account of all of the other potential reasons why turnover could vary apart from private information arrival. Appendix D shows that all other aspects of the PIN model including the interpretation of the PIN measure itself and its implied relation to spreads are the same.

Specifically, the EPIN model in Panel B of Fig. 3.6 draws $\lambda_t$ from a $\text{Gamma}(r, p/(1 - p))$ distribution with shape parameter $r$ and scale parameter $p/(1 - p)$. The fact that $\lambda_t$ is drawn from $\text{Gamma}$ distribution makes the model particularly tractable since the mixture of the $\text{Poisson}$ and $\text{Gamma}$ distributions is the well-known $\text{Negative Binomial}$ distribution. In the EPIN model, the number of trades ($B + S$) is dis-
tributed as *Negative Binomial* (see Appendix D for proof). The fact that the number of trades is distributed as a *Negative Binomial* simplifies the numerical estimation of the model by maximum likelihood quite a bit because we estimate the parameters $r$ and $p$ from the distribution of number trades, while in a second step we estimate the other model parameters with maximum likelihood.

Let $\Theta_{EPIN,i} = (\alpha_i, \delta_i, \eta_i, \theta_i, r_i, p_i)$ be the vector of parameters of the EPIN model for stock $i$. Let $B_{i,t}$ and $S_{i,t}$ be the number of buys and sells, respectively, for stock $i$ on day $t$. Let $D_{EPIN,i,t} = [B_{i,t}, S_{i,t}, \Theta_{EPIN,i}]$. The likelihood function of the extended model is $Q_T = \prod_{t=1}^T L(D_{EPIN,i,t})$, where $L(D_{EPIN,i,t})$ is the sum of $L_{NI}(D_{EPIN,i,t})$ (the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information), $L_{I+}(D_{EPIN,i,t})$ (the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with positive information), and $L_{I-}(D_{EPIN,i,t})$ (the likelihood on a day with negative information). Conditional on $\lambda_t$ and analogous to the original PIN model, each term in the likelihood function corresponds to a branch in the tree in Fig. 3.6, Panel B. See Internet Appendix D for the formulas for the likelihood functions.

As with the PIN model, for each stock-day, we compute the probability of an information event conditional on both the model parameters and on the number of buys and sells observed that day. Specifically, recall that we define the indicator $I_{i,t}$ as a variable that takes the value of one if an information event occurs for stock $i$ on day $t$ and zero otherwise. We compute $CPIE_{EPIN,i,t} = P[I_{i,t} = 1|D_{DY,i,t}]$, which is equal to $(L_{I-}(D_{EPIN,i,t}) + L_{I+}(D_{EPIN,i,t}))/L(D_{EPIN,i,t})$. Internet Appendix D provides details about the maximum likelihood procedure and the calculation of $CPIEs$. 
Table 3.5: **EPIN Parameter Estimates.** This table summarizes parameter estimates of the EPIN model for 21,206 PERMNO-Year samples from 1993 to 2012. \( \alpha \) represents the average unconditional probability of an information event at the daily level. \( \delta \) represents the probability of good news, and \( 1 - \delta \) represents the probability of bad news. The total number of trades in any given day \((t)\) is drawn from a Poisson distribution with intensity \( \lambda_t \), where \( \lambda_t \) is draw from a Gamma distribution with shape parameter \( r \) and scale parameter \( p/(1 - p) \). The number of buys (sells) on a day with no private information is draw from a Poisson distribution with intensity \( \theta \times \lambda_t \) \(((1 - \theta) \times \lambda_t)\). On days with negative (positive) news, the number of buys (sells) is drawn from a Poisson with intensity \( \theta/(1 + \eta) \times \lambda_t \) \(((1 - \theta)/(1 + \eta) \times \lambda_t)\). \( \bar{CPIE} \) and \( \text{Std}(CPIE) \) are the PERMNO-Year mean and standard deviation of \( CPIE \).

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>21,206</td>
<td>0.493</td>
<td>0.088</td>
<td>0.448</td>
<td>0.498</td>
<td>0.543</td>
</tr>
<tr>
<td>( \delta )</td>
<td>21,206</td>
<td>0.495</td>
<td>0.184</td>
<td>0.372</td>
<td>0.492</td>
<td>0.616</td>
</tr>
<tr>
<td>( r )</td>
<td>21,206</td>
<td>7.210</td>
<td>4.724</td>
<td>4.056</td>
<td>5.976</td>
<td>8.960</td>
</tr>
<tr>
<td>( p )</td>
<td>21,206</td>
<td>0.948</td>
<td>0.080</td>
<td>0.932</td>
<td>0.984</td>
<td>0.997</td>
</tr>
<tr>
<td>( \theta )</td>
<td>21,206</td>
<td>0.515</td>
<td>0.049</td>
<td>0.493</td>
<td>0.514</td>
<td>0.546</td>
</tr>
<tr>
<td>( \eta )</td>
<td>21,206</td>
<td>0.316</td>
<td>0.242</td>
<td>0.152</td>
<td>0.240</td>
<td>0.413</td>
</tr>
<tr>
<td>( \bar{CPIE} )</td>
<td>21,206</td>
<td>0.494</td>
<td>0.087</td>
<td>0.449</td>
<td>0.499</td>
<td>0.543</td>
</tr>
<tr>
<td>( \text{Std}(CPIE) )</td>
<td>21,206</td>
<td>0.414</td>
<td>0.082</td>
<td>0.367</td>
<td>0.445</td>
<td>0.478</td>
</tr>
</tbody>
</table>

Table 3.5 contains summary statistics for the parameter estimates of the EPIN model. Table 3.5 also contains summary statistics of the cross-sectional sample means and standard deviations of \( CPIE_{EPIN} \). We see that the mean \( CPIE_{EPIN} \) behaves exactly like \( \alpha \). We also estimate the parameter vectors \( \Theta_{EPIN,i} \) in the period \( t \in [-312, -60] \) before earnings announcements and opportunistic insider trades. These parameter estimates are used to compute the \( CPIE \)s in Section 3.3.3. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 3.5.

To illustrate how the EPIN model works, we present a stylized example of the
EPIN in Fig. 3.7. Analogous to the PIN model plot in Fig. 3.3, we plot simulated and real order flow data for Exxon-Mobil during 1993 and 2012, with buys on the horizontal axis and sells on the vertical axis. Panels A and B of Fig. 3.7 illustrate the central intuition behind the EPIN model. The simulated data comprise three types of days, which create three distinct clusters. Two of the clusters are made up of days characterized by large order flow imbalances relative to turnover (large $\frac{|B-S|}{B+S}$), with a large number of sells (buys) and relatively few buys (sells). The third group of days has low order imbalance relative to turnover–these days have no private information arrival and are clustered around the dotted line in the center of the scatter plot.

An econometrician using the EPIN model, moving along the dotted line, would observe that high turnover days–days the PIN model classifies as information events–are no longer classified as such, because higher turnover is driven by a large draw of the parameter $\lambda_t$ under the EPIN model. Instead, the EPIN model identifies private information when moving away from the dotted line–when buys are greater than sells and vice-versa.

Panels C and D plot $CPIE_{EPIN}$ as function of turnover. As opposed to the analogous plot of the PIN model in Fig. 3.3, Panels C and D do not suggest any relation between turnover and $CPIE_{EPIN}$. As a formal test of the EPIN model we run regressions of $CPIE_{EPIN}$ on absolute order imbalance scaled by turnover ($\frac{|B-S|}{B+S}$) and a squared term ($\left(\frac{|B-S|}{B+S}\right)^2$). We use $\frac{|B-S|}{B+S}$ to analyze the EPIN model because, as Fig. 3.7 suggests, the EPIN model implies that days with information events are the ones in which the absolute order imbalance relative to turnover is large. We report median coefficient estimates and $t$–statistics across all firms within a particular year. The
Figure 3.7: **XOM EPIN.** This figure compares the real and simulated data for XOM in 1993 using the EPIN model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{EPIN}$, with darker markers (+) representing high and lighter markers (+) low $CPIEs$. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark “cloud” of points, and low probability states appear as a light “cloud” of points. The EPIN model has three states: no news, good news, and bad news. Panels C and D plot the CPIE values for the real data as a function of turnover along with a dotted vertical line indicating the mean turnover for the year.

(a) XOM 1993  
(b) XOM 2012  
(c)  
(d)
coefficients are standardized as in Table 3.3. We report the average of the median, the 5th, and the 95th percentiles of the $R^2$s and $R^2_{inc,s}$.

Panel A of Table 3.6 presents the results of regressions based on simulated data. As before, we report the median coefficient estimates and t-statistics. The coefficients are standardized so they represent the increase in $CPIE_{EPIN}$ due to a one standard deviation increase in the corresponding independent variable. We also report the average of the median, the 5th, and the 95th percentiles of the empirical distribution of $R^2$s of these regressions generated by the 1,000 simulations. In general the EPIN model identifies private information from absolute order imbalance scaled by turnover, and not from turnover. The median $R^2$ values are high, ranging from 61%-92%, while the incremental $R^2$ from turnover is small-typically below 4%.
Table 3.6: EPIN Model Regressions. This table reports real and simulated regressions of the $CPIE_{EPIN}$ on the proportion of imbalanced trades ($\frac{|B-S|}{B+S}$) and its square. In Panel A, we simulate 1,000 instances of the EPIN model for each PERMNO-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the $R^2 (R^2_{inc.})$ values. We compute the incremental $R^2_{inc.}$ as the $R^2$ attributed to turn and $turn^2$ in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median $R^2$ values, and tests of the hypothesis that the observed variation in $CPIE_{EPIN}$ is consistent with the EPIN model. The $p$-value of $R^2 (R^2_{inc.})$ is the probability of observing an $R^2$ at least as small (large) as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$R^2$</th>
<th>$R^2_{inc.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\frac{</td>
<td>B-S</td>
<td>}{B+S}$</td>
<td>$\left(\frac{</td>
</tr>
<tr>
<td>1993</td>
<td>0.382</td>
<td>-0.134</td>
<td>(8.22)</td>
<td>(-3.04)</td>
</tr>
<tr>
<td>1994</td>
<td>0.355</td>
<td>-0.119</td>
<td>(8.19)</td>
<td>(-2.83)</td>
</tr>
<tr>
<td>1995</td>
<td>0.350</td>
<td>-0.113</td>
<td>(7.86)</td>
<td>(-2.59)</td>
</tr>
<tr>
<td>1996</td>
<td>0.364</td>
<td>-0.122</td>
<td>(8.31)</td>
<td>(-2.90)</td>
</tr>
<tr>
<td>1997</td>
<td>0.369</td>
<td>-0.126</td>
<td>(8.03)</td>
<td>(-2.84)</td>
</tr>
<tr>
<td>1998</td>
<td>0.388</td>
<td>-0.131</td>
<td>(8.93)</td>
<td>(-3.03)</td>
</tr>
<tr>
<td>1999</td>
<td>0.465</td>
<td>-0.190</td>
<td>(10.90)</td>
<td>(-4.26)</td>
</tr>
<tr>
<td>2000</td>
<td>0.447</td>
<td>-0.171</td>
<td>(9.34)</td>
<td>(-3.60)</td>
</tr>
<tr>
<td>2001</td>
<td>0.425</td>
<td>-0.123</td>
<td>(6.77)</td>
<td>(-2.08)</td>
</tr>
<tr>
<td>2002</td>
<td>0.243</td>
<td>0.007</td>
<td>(2.86)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>2003</td>
<td>0.033</td>
<td>0.202</td>
<td>(0.30)</td>
<td>(1.95)</td>
</tr>
<tr>
<td>2004</td>
<td>-0.477</td>
<td>0.679</td>
<td>(-4.25)</td>
<td>(6.10)</td>
</tr>
<tr>
<td>2005</td>
<td>0.343</td>
<td>-0.062</td>
<td>(3.38)</td>
<td>(-0.67)</td>
</tr>
<tr>
<td>2006</td>
<td>0.294</td>
<td>-0.018</td>
<td>(3.16)</td>
<td>(-0.21)</td>
</tr>
<tr>
<td>2007</td>
<td>0.778</td>
<td>-0.338</td>
<td>(17.81)</td>
<td>(-7.59)</td>
</tr>
<tr>
<td>2008</td>
<td>0.784</td>
<td>-0.335</td>
<td>(18.60)</td>
<td>(-7.90)</td>
</tr>
<tr>
<td>2009</td>
<td>0.777</td>
<td>-0.321</td>
<td>(19.72)</td>
<td>(-8.04)</td>
</tr>
<tr>
<td>2010</td>
<td>0.773</td>
<td>-0.318</td>
<td>(19.47)</td>
<td>(-7.97)</td>
</tr>
<tr>
<td>2011</td>
<td>0.783</td>
<td>-0.335</td>
<td>(19.80)</td>
<td>(-8.16)</td>
</tr>
<tr>
<td>2012</td>
<td>0.781</td>
<td>-0.332</td>
<td>(19.89)</td>
<td>(-8.23)</td>
</tr>
</tbody>
</table>
Table 3.6: **EPIN Model Regressions.** Continued.

(b) Real Data

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$R^2$</th>
<th>$R^2_{inc.}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>0.336</td>
<td>(-2.93)</td>
<td>57.90%</td>
<td>24.42%</td>
</tr>
<tr>
<td>1994</td>
<td>0.321</td>
<td>(-2.92)</td>
<td>56.55%</td>
<td>24.76%</td>
</tr>
<tr>
<td>1995</td>
<td>0.317</td>
<td>(-2.62)</td>
<td>58.03%</td>
<td>26.23%</td>
</tr>
<tr>
<td>1996</td>
<td>0.339</td>
<td>(-3.06)</td>
<td>59.28%</td>
<td>22.31%</td>
</tr>
<tr>
<td>1997</td>
<td>0.339</td>
<td>(-2.98)</td>
<td>57.53%</td>
<td>24.71%</td>
</tr>
<tr>
<td>1998</td>
<td>0.362</td>
<td>(-3.34)</td>
<td>61.34%</td>
<td>25.83%</td>
</tr>
<tr>
<td>1999</td>
<td>0.433</td>
<td>(-4.88)</td>
<td>62.95%</td>
<td>20.51%</td>
</tr>
<tr>
<td>2000</td>
<td>0.419</td>
<td>(-3.95)</td>
<td>58.88%</td>
<td>22.18%</td>
</tr>
<tr>
<td>2001</td>
<td>0.402</td>
<td>(-2.62)</td>
<td>50.55%</td>
<td>8.71%</td>
</tr>
<tr>
<td>2002</td>
<td>0.255</td>
<td>(-0.27)</td>
<td>42.07%</td>
<td>5.25%</td>
</tr>
<tr>
<td>2003</td>
<td>0.126</td>
<td>(1.36)</td>
<td>40.55%</td>
<td>4.18%</td>
</tr>
<tr>
<td>2004</td>
<td>-0.067</td>
<td>(3.54)</td>
<td>38.32%</td>
<td>2.76%</td>
</tr>
<tr>
<td>2005</td>
<td>0.249</td>
<td>(-0.20)</td>
<td>41.68%</td>
<td>2.38%</td>
</tr>
<tr>
<td>2006</td>
<td>0.264</td>
<td>(-0.34)</td>
<td>43.41%</td>
<td>1.89%</td>
</tr>
<tr>
<td>2007</td>
<td>0.762</td>
<td>(-9.57)</td>
<td>66.36%</td>
<td>0.90%</td>
</tr>
<tr>
<td>2008</td>
<td>0.800</td>
<td>(-11.20)</td>
<td>70.98%</td>
<td>0.64%</td>
</tr>
<tr>
<td>2009</td>
<td>0.813</td>
<td>(-11.49)</td>
<td>71.79%</td>
<td>0.48%</td>
</tr>
<tr>
<td>2010</td>
<td>0.814</td>
<td>(-11.44)</td>
<td>72.77%</td>
<td>0.67%</td>
</tr>
<tr>
<td>2011</td>
<td>0.809</td>
<td>(-11.21)</td>
<td>71.67%</td>
<td>0.54%</td>
</tr>
<tr>
<td>2012</td>
<td>0.804</td>
<td>(-11.14)</td>
<td>72.72%</td>
<td>0.91%</td>
</tr>
</tbody>
</table>
Panel B of Table 3.6 reports regression results for the real rather than simulated data. In contrast to the PIN model, in the real data the EPIN model identifies private information from order imbalance and not turnover. The median $R^2$ values are high, ranging from 38%–72%, while the incremental $R^2$ from turnover is small—typically below 1%. Because the $R^2$ values using the real data are on average lower than those in the simulated data, our statistical tests frequently reject the null that the $CPIE_{EPIN}$ is driven by absolute order imbalance scaled by turnover. However, the EPIN model passes the more important test, namely in the majority of stock-year observations in the real data the incremental $R^2$ due to turnover is at least as large as the incremental $R^2$ in the simulated data. Therefore, the EPIN model, while not a perfect description of the order flow data, fixes the problem of the PIN model which mechanically identifies private information from higher turnover.

### 3.3.2 The OWR model

Odders-White and Ready (2008) extend Kyle (1985) by allowing for days with information events and days without information events. Private information arrives before the opening of the trading day with probability $\alpha$. On days when private information arrives, the model assumes that the information is publically revealed after the close of trade. The OWR model identifies the arrival of private information through order flow imbalance, $y_e$, the immediate intraday price response to order imbalance, $r_d$, and through subsequent overnight price changes, $r_o$.\(^{17}\) The vector $(y_e, r_d, r_o)$ is assumed to be multivariate normal with mean zero and a covariance matrix.

\(^{17}\)We suppress the $t$ subscript for ease of exposition.
that differs between information days and noninformation days.\footnote{We follow Odders-White and Ready and remove systematic effects from returns to obtain measures of unexpected overnight and intraday returns \( (r_o \text{ and } r_d) \). See Section 3.1 and Internet Appendix B for a detailed description of how we compute \( y_e, r_o, \text{ and } r_d \).} Fig. 3.8 shows the time line of the model. The intuition behind the OWR model is that the market maker updates prices in response to order flow because the order flow could reflect an information event. However, the subsequent price pattern is different depending on whether there actually was an information event or not. If an information event occurs, the overnight price response reflects a continuation of the market makers’ intraday reaction. If no information event occurs, the overnight price response reverses the market makers’ initial price reaction. Therefore, an econometrician can make inferences about the probability of an information event in the OWR model because the covariance matrix of the three variables \( (y_e, r_d, r_o) \) differs between days when private information arrives and days when only public information is available.\footnote{Unlike the market maker who must update prices before observing the overnight revelation of information, the econometrician in the OWR model can make inferences about the arrival of private information after viewing the overnight price response.}
Figure 3.8: **OWR Tree.** In the OWR model, prior to markets opening, private information arrives with probability $\alpha$. Once markets open investors submit their trades generating order imbalance ($y_e$), and the intraday return ($r_d$). After markets close, private information becomes public and is reflected in the overnight return ($r_o$). The variables ($y_e$, $r_d$, $r_o$) are normally distributed with mean zero and covariance $\Sigma$, where $\Sigma$ is function of the information arrival indicator ($I$). For instance, when there is no private information, there is a reversal in the returns ($cov(r_d, r_o) < 0$) and when there is private information there is a continuation in the returns ($cov(r_d, r_o) > 0$).
To see how the covariance matrix of \((y_e, r_d, r_o)\) differs between information and noninformation days, consider first the covariance of the intraday and overnight returns, \(cov(r_o, r_d)\). This covariance is positive for information events, reflecting the fact that the information event is not completely captured in prices during the day and the revelation of the private information means that the overnight return continues the partial intraday price reaction. In contrast, \(cov(r_o, r_d)\) is negative in the absence of an information event since the market maker’s reaction to the noise trade during the day is reversed when she learns that there was no private signal.

The other moments in the covariance matrix of \((y_e, r_d, r_o)\) are also affected by the arrival of private information. If no information event occurs, then \(Var(y_e)\) is composed of only the variances of the uninformed order flow and the noise in the data. However, if an event occurs, \(Var(y_e)\) increases because the order flow reflects at least some informed trading. Similarly, \(Var(r_d)\) is higher for an information event, because it reflects the market maker’s partial reaction to the day’s increased order flow. Since the private signal is revealed after trading closes, \(Var(r_o)\) also increases in the wake of an information event, as it reflects the remainder of the market maker’s partial reaction to the informed trade component in order flow. Likewise, information events make \(cov(y_e, r_d)\) and \(cov(y_e, r_o)\) rise. The higher covariance between order flow and intraday returns occurs because, in an information event, both order flow and the intraday return (partially) reflect the impact of informed trading. Along these same lines, because the market maker cannot separate the informed from the uninformed order flow, she is unable to fully adjust the price during the day to reflect the informed trader’s private signal. However, since the private signal is publically
revealed and fully reflected in prices after the close, $cov(y_e, r_o)$ is higher during an information event because the overnight returns incorporate any additional reaction to the private signal that was not captured in prices during the day.

In contrast to the PIN model, the OWR model does not contain a direct analog to the probability of informed trading. To understand this result, note that the probability of informed trade in the PIN and EPIN models is given by the ratio of the expected number of informed trades to the expected total number of trades on a given day. Since the OWR model employs only the difference between buys and sells, it does not make assumptions about the distribution of number of trades. Thus, the OWR is mute regarding the ratio of informed to expected number of trades. This may appear to be a limitation of the OWR model, but this is actually an advantage because it allows the OWR model to disentangle variations in turnover from the arrival of informed trading.

Even though the OWR model does not have a measure analogous to the PIN measure, the OWR model admits multiple measures of private information. Odders-White and Ready (2008) motivate their model as a tool to separate the expected liquidity provider losses due to trading with informed traders into the frequency of private information arrival and the expected magnitude of the private information. Hence, the OWR allows for the construction of private information measures that are based on both dimensions. The PIN and EPIN models, on the other hand, focus only on the frequency of information arrival and are silent with respect to the expected magnitude of the private information. Hence, our comparison of the EPIN and OWR models with $CPIE_{EPIN}$ with $CPIE_{OWR}$ focuses on the dimension
of private information that both models have in common, namely the frequency of information arrival. The fact that we are using $\alpha$ and CPIEs to compare the models does not imply that we are taking the position that frequency measures are the only private information metrics that are worthy of consideration.

Analogous to the EPIN and PIN models, let $\Theta_{OWR,i}$ represent the vector of OWR model parameters for stock $i$, $r_{d,i,t}$ and $r_{o,i,t}$ represent the intraday and overnight returns for stock $i$ on day $t$, and $y_{e,i,t}$ represents the order flow imbalance for stock $i$ on day $t$. Let $D_{OWR,i,t} = [\Theta_{OWR,i}, y_{e,i,t}, r_{d,i,t}, r_{o,i,t}]$. The likelihood function of the OWR model is $Q_{T,t} = \prod_{t=1}^{T} L(D_{OWR,i,t})$ where $L(D_{OWR,i,t})$ is the sum of $L_{NI}$ (the likelihood of observing $y_{e,i,t}$, $r_{d,i,t}$, and $r_{o,i,t}$ on a day without private information), and $L_I$ (the likelihood of observing $y_{e,i,t}$, $r_{d,i,t}$, and $r_{o,i,t}$ on a day with an information event). See Internet Appendix E for the formulas for the likelihood functions. The probability of an information event, conditional on $D_{OWR,i,t}$, is therefore $CPIE_{OWR,i,t} = P[I_{i,t} = 1|D_{OWR,i,t}]$. This probability is given by $L_I(D_{OWR,i,t})/L(D_{OWR,i,t})$. As in the PIN and EPIN models, if we condition down with respect to the data, $CPIE_{OWR,i,t}$ reduces to the model’s unconditional probability of information events ($\alpha_i$).

As with the PIN and EPIN models, we estimate the OWR model numerically via maximum likelihood. Table 3.7 contains summary statistics for the parameter estimates of the OWR model. Table 3.7 also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{OWR}$. As in the PIN and EPIN models, we see that the mean $CPIE_{OWR}$ behaves exactly like $\alpha$ in the OWR model. The estimated OWR $\alpha$ parameters are in general higher than those
in Odders-White and Ready (2008). This is due to the fact that our definition of $y_e$ is different from that in Odders-White and Ready (2008) (see the discussion in Section 3.1 above).\textsuperscript{20} We also estimate the parameter vector $\Theta_{OWR,i}$ in the period $t \in [-312, -60]$ before earnings announcements and opportunistic insider trades. These parameter estimates are used to compute the CPIEs in Sections 3.3.3 and 3.3.3. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 3.7 and are in Internet Appendix D. This appendix also displays results indicating that the OWR model does not suffer from the same conflation of turnover and arrival of private information that plagues the PIN model.

\textsuperscript{20}In fact, we get $\alpha$ estimates close to those reported in Odders-White and Ready (2008) if we define $y_e$ in the same way that they do.
Table 3.7: **OWR Parameter Estimates.** This table summarizes parameter estimates of the OWR model for 21,206 PERMNO-Year samples from 1993 to 2012. \( \alpha \) represents the average unconditional probability of an information event at the daily level. \( \sigma_u \) represents the standard deviation of the order imbalance due to uninformed traders, which is observed with normally distributed noise with variance \( \sigma^2_z \). \( \sigma_i \) represents the standard deviation of the informed trader’s private signal. \( \sigma_{pd} \) and \( \sigma_{po} \) represent the standard deviation of intraday and overnight returns, respectively. CPIE and Std(CPIE) are the PERMNO-Year mean and standard deviation of CPIE.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>21,206</td>
<td>0.437</td>
<td>0.257</td>
<td>0.214</td>
<td>0.436</td>
<td>0.639</td>
</tr>
<tr>
<td>( \sigma_u )</td>
<td>21,206</td>
<td>0.075</td>
<td>0.068</td>
<td>0.022</td>
<td>0.062</td>
<td>0.109</td>
</tr>
<tr>
<td>( \sigma_z )</td>
<td>21,206</td>
<td>0.239</td>
<td>0.143</td>
<td>0.137</td>
<td>0.221</td>
<td>0.332</td>
</tr>
<tr>
<td>( \sigma_i )</td>
<td>21,206</td>
<td>0.030</td>
<td>0.286</td>
<td>0.013</td>
<td>0.021</td>
<td>0.027</td>
</tr>
<tr>
<td>( \sigma_{pd} )</td>
<td>21,206</td>
<td>0.010</td>
<td>0.005</td>
<td>0.006</td>
<td>0.009</td>
<td>0.012</td>
</tr>
<tr>
<td>( \sigma_{po} )</td>
<td>21,206</td>
<td>0.006</td>
<td>0.004</td>
<td>0.004</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>CPIE</td>
<td>21,206</td>
<td>0.451</td>
<td>0.258</td>
<td>0.227</td>
<td>0.455</td>
<td>0.656</td>
</tr>
<tr>
<td>Std(CPIE)</td>
<td>21,206</td>
<td>0.137</td>
<td>0.047</td>
<td>0.109</td>
<td>0.142</td>
<td>0.171</td>
</tr>
</tbody>
</table>
3.3.3 A horse race between EPIN and OWR models

A fundamental problem in the literature related to testing and proposing measures of private information is the lack of cleanly identifiable periods in which private information is present in the market. To address this issue, we use three different methods to analyze the performance of the OWR and EPIN models. In Section 3.3.3 we analyze how $CPIE_{OWR}$ and $CPIE_{EPIN}$ vary around earnings announcements. The working hypothesis in this test is that private information arrival is more likely before than after the announcement. In Section 3.3.3 we analyze how $CPIE_{OWR}$ and $CPIE_{EPIN}$ vary around insider trading events. In Section 3.3.3 we analyze how $CPIE_{OWR}$ and $CPIE_{EPIN}$ are related to return autocorrelations.

Each of these three methods has its own unique limitations. For instance, it is possible that, for some reason, private information is more prevalent after important announcements than before. Similar critiques could be levied against the other two methods. However, if all of these methods point to the same conclusions, it seems unlikely that our overall interpretation would be biased due to the limitations of any specific method.

Information event probabilities under the EPIN and OWR models

Panel A of Fig. 3.9 illustrates the average $CPIE_{EPIN}$ in event time for our sample of earnings announcements. In contrast to the PIN model, the probability of an information event decreases from around 51% 20 days before the announcement and drops on the announcement date to around 46%. This pattern is not consistent with
informed traders acting on private information before the announcement. Panel B of Fig. 3.9 illustrates the average $CPIE_{OWR}$ in event time for our sample of earnings announcements. Similar to the PIN model, the probability of an information event increases from around 40% 20 days before the announcement and peaks on the announcement date at around 45%. Panel B indicates that the $CPIE_{OWR}$ is far outside of two standard deviations from its mean (estimated between $t \in [-40, -21]$) on the announcement date $t = 0$. This pattern is consistent with the timing of the OWR model where informed traders act on private information during the day before the public announcement occurs overnight ($t \in [0, 1]$). Unlike the PIN model, the $CPIE_{OWR}$ drops back to its pre-event mean within a few days after the announcement. This is consistent with the intuition that there is more scope for informed trading before the announcement than after.
Figure 3.9: Earnings Announcements. Panel A (B) shows the average $CPIE_{EPIN}$ ($CPIE_{OWR}$) for the EPIN (OWR) model in event time surrounding earnings announcements.

(a) $CPIE_{EPIN}$

(b) $CPIE_{OWR}$
What causes the EPIN results to be so different from the PIN results above? Fig. 3.10 sheds light on this question. Panel A of Fig. 3.10, shows the actual $CPI_{EPIN}$ along with predicted values from a regression of $CPI_{EPIN}$ on the proportion of imbalanced trades $\left( \frac{|B-S|}{(B+S)} \right)$ and its square. The results indicate that $CPI_{EPIN}$ drops because the imbalance is small relative to the absolute amount of trade on the announcement day. This is consistent with the results in Easley, Engle, O’Hara, and Wu (2008), who show that in their sample of 834 announcements that the average proportion of imbalanced trades decreases on earnings announcement days. The PIN model interprets the increase in turnover as indicative of the arrival of private information, but the EPIN model, on the other hand, is able to use the information in the order flow imbalance to draw the opposite conclusion. Panel B provides support for this notion by showing that the $CPI_{EPIN}$ does not respond to the increase in total trade. Panel B shows the predicted $CPI_{EPIN}$ based on a regression of $CPI_{EPIN}$ on $(B + S)$ and its square. The results indicate that, consistent with the motivation for the extended model, $CPI_{EPIN}$ responds to order imbalance not total trade.
Figure 3.10: Earnings Announcements - EPIN Decomposition. Panels A and B compare the average $CPI_{EPI}N$ with the $CPI_{EPI}N$ predicted using either $\frac{|B-S|}{B+S}$ or turnover, respectively. To obtain the predictions, we run regressions of daily $CPI_{EPI}N$ on $\frac{|B-S|}{B+S}$ or $turn$, and their respective squared terms.

(a) Prediction using $\frac{|B-S|}{B+S}$ and $\frac{|B-S|^2}{B+S}$

(b) Prediction using $turn$ and $turn^2$
As we saw in Section 3.3.2, the OWR model identifies private information from the covariance matrix of the three variables in the model \((y_{e,i,t}, r_{o,i,t}, r_{d,i,t})\). Therefore, to analyze how the OWR model identifies private information around earnings announcements, we decompose \(CPIE_{OWR}\) on to the squared and interaction terms of \((y_{e,i,t}, r_{o,i,t}, r_{d,i,t})\). Panels A–F of Fig. 3.11 show that the majority of the variation in measured private information \((CPIE_{OWR})\) comes from intraday returns squared (Panel B) and the interaction between the intraday and overnight returns (Panel F). Order imbalance squared (Panel A) provides no explanatory power, although the interaction between the order imbalance and returns (Panels D and E) does have some impact.

Our results suggest that order flow, however well modeled, is insufficient to be the sole source of inferences about private information arrival. Under the working hypothesis that there is more informed trade before rather than after earnings announcements, our findings suggest that the OWR model identifies private information in a sensible way while the EPIN does not. Even though the magnitude of the increase in \(CPIE_{OWR}\) around the event date may be considered small, \(CPIE_{OWR}\) increases before the event day while \(CPIE_{EPIN}\) counter-intuitively decreases. Since both models use order flow to identify private information, the marked difference in the results highlights the importance of using variables other than just order flow. The use of returns, particularly intraday returns, allows the OWR model to reach a different and more economically sensible conclusion. Moreover, the fact that order imbalance alone explains very little of the variation in \(CPIE_{OWR}\) around earnings announcements also emphasizes the relatively low contribution of order flow relative
Figure 3.11: Earnings Announcements - OWR Decomposition. Panels A–F compare the average $CPIE_{OWR}$ with the $CPIE_{OWR}$ predicted using the squared and interaction terms of $y_e$, $r_d$, and $r_o$.

(a) Prediction using $y_e^2$
(b) Prediction using $r_d^2$
(c) Prediction using $r_o^2$
(d) Prediction using $y_e \times r_d$
(e) Prediction using $y_e \times r_o$
(f) Prediction using $r_d \times r_o$
to returns in identifying private information. Our results therefore provide empirical support for the proposition in Back, Crotty, and Li (2014) and in Kim and Stoll (2014) that researchers cannot use order flow alone to successfully identify periods of informed trade.

*CPIE*<sub>EPIN</sub> and *CPIE*<sub>OWR</sub> around insider trading

In this section we investigate whether the OWR and EPIN models are capable of identifying opportunistic insider trades using the insider trade classification scheme developed in Cohen, Malloy, and Pomorski (2012).<sup>21</sup> Cohen, Malloy, and Pomorski (2012) show that a long-short portfolio that exploits the trades of opportunistic traders (opportunistic buys minus opportunistic sells) earns value-weighted abnormal return of 82 basis points per month (9.8 percent annualized, t=2.15). They also show that the trades of opportunistic insiders show significant predictive power for future news about the firm, and that the fraction of traders who are opportunistic in a given month is negatively related to the number of recent news releases by the SEC regarding illegal insider trading cases. Their results are all consistent with opportunistic insider trades, as opposed to routine insider trades, being based on private information. Opportunistic insider trades therefore, provide a convenient laboratory to examine the models’ ability to detect the arrival of actionable private information.

Panel A (B) of Fig. 3.12 presents the average *CPIE*<sub>EPIN</sub> (*CPIE*<sub>OWR</sub>) in event time for our sample of opportunistic insider trades. There is no clear pattern in the

<sup>21</sup>See Section 3.1 for a further discussion of the classification of insider trades as opportunistic.
$CPIE_{EPI}N$ indicating the arrival of private information before the opportunistic insider trade, though there is an increase in $CPIE_{EPI}N$ on the day of the opportunistic insider trading.
Figure 3.12: **Opportunistic Insider Trades.** Panel A (B) shows the average $CPI_{EPIN}$ ($CPI_{OWR}$) for the EPIN (OWR) model in event time surrounding opportunistic insider trades.
In contrast, Panel B shows that the $CPIE_{OWR}$ identifies the arrival of private information in the days leading up to an opportunistic insider trade. Beginning at $t = -4$, the $CPIE_{OWR}$ is more than two standard deviations higher than the mean estimated between $t \in [-40, 21]$. However, $CPIE_{OWR}$ begins to drift strongly upward and very nearly crosses the two standard deviation bound as early at day $t = -10$. Strikingly, at $t = 1$, immediately after the trade, $CPIE_{OWR}$ drops precipitously back to average levels. We interpret this as strong evidence that the OWR model’s use of both order flow and returns is successful in uncovering informed trade.

Taken together, the insider trading event study evidence further supports the claim that order flows alone may be insufficient to identify private information. $CPIE_{EPIN}$, which varies based only on changes in order imbalances, is unable to clearly detect the imminent arrival of insider trades. $CPIE_{OWR}$, on the other hand, is able to predict insider trading based on small variations in intraday and overnight returns.

**Are $CPIE_{EPIN}$ and $CPIE_{OWR}$ related to return continuation?**

The market microstructure literature has long held that price changes related to informed trades should not be subsequently reversed while non-information related price changes (e.g. dealer inventory control, price pressure, price discreteness etc.) are transient (e.g. Hasbrouck (1988, 1991a,b)). In this section, we investigate whether $CPIE_{EPIN}$ and $CPIE_{OWR}$ are associated with subsequent return reversals. In particular, we examine the relation between $CPIEs$ and return autocorrelations. The intuition is that if a model’s $CPIE$ on day $t$ actually reflects a high
probability of informed trade then we expect that the return on day $t$ will be continued over the subsequent day as the information gradually becomes public and gets fully impounded in prices. To capture this idea we model return autocorrelations as linear functions of $CPIE$. Specifically, we consider the following regressions:

$$r_{i,t+1} = a + b_{OWR,1} r_{i,t} + b_{OWR,2} CPIE_{OWR,t} + b_{OWR,3} (r_{i,t} \times CPIE_{OWR,t}) + e_{i,t+1},$$

and

$$r_{i,t+1} = c + b_{EPIN,1} r_{i,t} + b_{EPIN,2} CPIE_{EPIN,t} + b_{EPIN,3} (r_{i,t} \times CPIE_{EPIN,t}) + w_{i,t+1}$$

In the above regressions, $r_{i,t}$ is the open-to-open, risk adjusted return ($r_{i,d,t} + r_{i,o,t}$) on day $t$. Thus, there is no overlap between the intraday and overnight returns that are used to compute $CPIE_{OWR,i,t}$ on day $t$ and the return on day $t + 1$. The coefficients $b_{OWR,2}$ and $b_{EPIN,2}$ reflect the impact of the model’s $CPIE$ on the correlation between the return on day $t$ and the return the next trading day. We estimate the regressions above using a panel regression approach including firm and year fixed effects with standard errors clustered by firm and year. Table 3.8 reports the coefficient estimates and t-statistics for these regressions. The results in Table 3.8 show that the estimates for both $b_{OWR,2}$ and $b_{EPIN,2}$ are positive and significant, indicating that both $CPIE_{EPIN}$ and $CPIE_{OWR}$ are associated with future return continuation. To see this note that both regressions show a tendency of daily returns to reverse because the coefficients on lagged returns in both regressions are negative. However, a one standard deviation shock to $CPIE_{OWR}$ is associated with a 0.02 (0.08 × 0.25) decline in the subsequent reversal, while a one standard deviation shock to $CPIE_{EPIN}$ is associated with only a 0.003 (0.006 × 0.49) drop in the subsequent reversal. Thus, while the point coefficient estimates for both the OWR and EPIN models suggest that $CPIE_{EPIN}$ and $CPIE_{OWR}$ capture information that has a persistent impact
Table 3.8: **Return Reversals.** This table reports regressions of the daily return at time $t+1$ on the return, $CPIE$, and the interaction at time $t$. Returns are measured from open to open. We include stock and year fixed effects and cluster standard errors by stock and year. * indicates statistical significance at the 10% level, ** at the 5%, and *** at the 1% level.

<table>
<thead>
<tr>
<th></th>
<th>$r_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OWR</td>
</tr>
<tr>
<td>$r_t$</td>
<td>-0.126***</td>
</tr>
<tr>
<td></td>
<td>(-6.06)</td>
</tr>
<tr>
<td>$CPIE_t$</td>
<td>0.000458***</td>
</tr>
<tr>
<td></td>
<td>(4.35)</td>
</tr>
<tr>
<td>$CPIE_t \times r_t$</td>
<td>0.0816***</td>
</tr>
<tr>
<td></td>
<td>(4.16)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.0061</td>
</tr>
<tr>
<td>Obs.</td>
<td>5,284,135</td>
</tr>
</tbody>
</table>

on prices, the effect is ten times stronger with the OWR $CPIE$. We view this as further evidence that the using variables such as returns to augment order flow allow researchers to make stronger inferences about private information arrival.

### 3.4 Conclusion

The PIN measure, developed in the seminal work of Easley and O’Hara (1987), Easley, Kiefer, O’Hara, and Paperman (1996), and Easley, Kiefer, and O’Hara (1997), is arguably the most widely used measure of information asymmetry in the accounting, corporate finance and asset pricing literature today. Recent work however suggests that PIN fails to capture private information (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Joche (2007), and Collin-Dufresne and Fos
This paper analyzes why the model might incorrectly identify informed trade.

Our findings indicate that the PIN model fits the data so poorly that it mechanically groups all sources of variation in turnover (e.g. disagreement, calendar effects, portfolio rebalancing, taxation, etc.) under the umbrella of private information arrival. This is at odds with a vast literature that suggests turnover varies for many reasons unrelated to the arrival of private information. This failure of the PIN model is particularly strong after the increase in turnover in the early 2000s. In fact, after 2002 for the median stock in our sample, the PIN model is essentially equivalent to a naive model that assigns a probability of one to the arrival of private information on any day where turnover is above average and zero probability to the arrival of private information on any other day. These findings suggest some important insights for future research that tests, constructs, or uses proxies for informed trade.

Our results suggest that event study based tests of private information proxies (e.g. Easley, Engle, O'Hara, and Wu (2008) and Brennan, Huh, and Subrahmanyam (2015)) can be misleading if one fails to account for the fact that patterns in their private information measures may simply reflect event related patterns in turnover that have nothing to do with private information arrival. For instance, Brennan, Huh, and Subrahmanyam (2015) interpret the fact that their $CPIE_{PIN}$ measures are higher after earnings announcements than before as evidence of informed trading. However, we show that $CPIE_{PIN}$ is naively related to turnover. This suggests that the findings in Brennan, Huh, and Subrahmanyam (2015) can simply be attributed to the fact that turnover is typically much higher after earnings announcements.
Our findings also suggest that future research aimed at building measures of informed trade should focus on the use of variables other than simply net order flow alone because order flow, however well modeled, appears insufficient to identify private information. Specifically, we use three different methods to compare the OWR model, which infers the arrival of private information from returns and order flow, with an extension of the PIN model (the EPIN model), which is solely based on order flow but corrects the PIN model’s mechanical association of private information arrival with variation in turnover. The OWR model performs better than the EPIN model in all three tests. First, the EPIN model actually predicts a decrease in private information arrival before earnings announcements while the OWR model captures a pattern of increasing private information arrival prior to the announcement and a marked decrease after the announcement. Second, $CPIE_{OWR}$ predicts periods of opportunistic insider trading and decreases dramatically immediately following the insider trades, while $CPIE_{EPIN}$ displays no such clear pattern around these events. Lastly, the relation between $CPIE_{OWR}$ and future return continuation is ten times larger than that of the $CPIE_{EPIN}$. These results provide empirical support for the theoretical proposition in Back, Crotty, and Li (2014) and in Kim and Stoll (2014) that researchers cannot use net order flow alone to successfully identify periods of informed trade.

Our findings also suggest that future research in corporate finance, accounting or asset pricing that uses information asymmetry measures should consider using proxies for private information based on the OWR model, for instance $CPIE_{OWR}$ or its $\alpha$, instead of using proxies based on the PIN model (e.g. $PIN$).
Appendix A

WTCN and IDC

A.1 Information Sharing Matrix

Suppose that each agent $i = 1, \ldots, N$ is endowed with a private signal $s_i$, and each agent updates their signal by taking weighted averages of the differences between their current signal and the signals of their immediate neighbors such that the evolution of agent $i$’s signal is given by:

$$\frac{ds_i}{dt} = -\sum_j \frac{A_{ij}}{\sqrt{d_i} \sqrt{d_j}} (s_j - s_i),$$

$$= -s_i + \sum_j \frac{A_{ij}}{\sqrt{d_i} \sqrt{d_j}} s_j.$$  \hfill (A.1)

The second equality implies agent $i$ “sends” his signal $s_i$ to his neighbors and receives $\frac{A_{ij}}{\sqrt{d_i} \sqrt{d_j}} s_j$ from each of his neighbors. Sharing is bilateral because $\frac{A_{ij}}{\sqrt{d_i} \sqrt{d_j}}$ is symmetric.

In vector form,

$$\frac{ds}{dt} = -(I - D^{-1/2} AD^{-1/2})s = -\mathcal{L}s,$$  \hfill (A.2)

where $I$ is the identity matrix, $D$ is the diagonal matrix of each agents’ degrees, $A$ is the adjacency matrix, and $\mathcal{L}$ is the symmetric graph Laplacian. The differential
equation has the solution:

$$s_t = e^{-Lt} s_0, \quad \text{(A.3)}$$

given an initial distribution of signals $s_0$. I call $S_t = e^{-Lt}$ the symmetric information “sharing” matrix. The non-symmetric version $S_t$ is also called the “heat kernel” in the spectral graph theory literature. Chung and Yau (1998) introduces the symmetric heat kernel, and Chung (2007) develops a centrality measure based on the non-symmetric version $e^{-Lt}$.

### A.2 Eigenvector centralities

One measure which has not been popular in the finance literature, Personalized PageRank, is useful because it is a close proxy for Information Diffusion Centrality and can be used to derive the “eigenvector centralities” popular in the finance literature (Bonacich, Katz, and Eigenvector). Recall that Information Diffusion Centrality can be written as the exponential sum of symmetric random walks (2.4). Analogously, Personalized PageRank can be written as the geometric sum of random walks:

$$\varepsilon(\alpha, \beta) = \beta \sum_{k=0}^{\infty} \alpha^k W^k p. \quad \text{(A.4)}$$

Or in vector form, $\varepsilon = (I - \alpha W)^{-1} \beta p$. Here, $p$ represents the “preference” vector or probability that a random walker will start from any given node which is analogous to the arrival of information above. $\alpha$ acts as a damping factor in the geometric
series, and $\beta$ is a positive constant.\(^1\) The standard random walk matrix $W$ can also be replaced with its symmetric counterpart to be consistent with the bilateral exchange of information.

PageRank, named after Google co-founder Larry Page, is designed to measure the importance of websites based on the notion that a website is important if other important websites link to it. As we have shown, it can also be expressed in terms of a “random (web) surfer” clicking links on pages such that important sites tend to be on the paths of many random walks (see Brin and Page, 1998; Page, Brin, Motwani, and Winograd, 1999). However it would be a mistake to call PageRank a measure of information diffusion as PageRank contains no description of the underlying information sharing process. Hence the canonical form of PageRank (with $p = 1$) is not the random walk form shown above, but the recursive form:

$$\epsilon = \alpha W \epsilon + \beta \mathbf{1}.$$ \hspace{1cm} (A.5)

Bonacich, Katz, and Eigenvector centrality are also defined in recursive form, and can in fact be derived as special cases of PageRank. To see this, we only need to examine the typical element of the PageRank vector:

$$\epsilon_i = \alpha \sum_j A_{ij} \frac{\epsilon_j}{d_j} + \beta.$$ \hspace{1cm} (A.6)

PageRank can be viewed as a scaled or random walk version of Bonacich centrality.

\(^1\) $\beta$ is sometimes set to $(1 - \alpha)$ with $\alpha < 1$ such that the above random walk can be interpreted as a “teleporting random walk.”
When $\varepsilon_j$ is not scaled by the degree $d_j$ inside the summation (i.e. if $d_j = 1$ above) PageRank becomes Bonacich centrality. Katz centrality is a special case of Bonacich centrality when $\beta = 1$, and the standard Eigenvector centrality is a special case of Bonacich centrality with $\alpha = 1$ and $\beta = 0$.

Although these “eigenvector centralities” may be positively correlated with Information Diffusion Centrality, only PageRank has the random walk interpretation, and even then PageRank lacks the information sharing foundations of information diffusion.

A.3 ANCerno Data

Each observation in the ANCerno dataset consists of the execution of a trade along with the corresponding order-level identifiers, and trade-day liquidity measures.

Stocks are identified by CUSIP at the time of execution, ANCerno clients are identified by clientcode, their money managers by managercode, and the broker-dealer responsible for each execution by brokercode. Following Puckett and Yan (2011) I define a fund as a client-manager pair. Using the ANCerno provided cross-reference metadata, I identify 10,355 unique funds, which are composed of 1,358 clients and 1,007 money managers. A sample of the cleaned data is shown below.

Order identifies an order ticket, Date indicates when the trade was made, Fund is the composite clientcode-brokercode identifier, Broker is the broker identifier, Stock is the PERMNO corresponding to the provided CUSIP, Price and Volume are at the trade level, Side is 1 if the trade was a buy and -1 if the trade was a sell, and Days
refers to the number of days over which the stock was traded as part of the Order.

The funds in my sample trade a total of 8,555 unique common stocks listed across the NYSE, NASDAQ, and AMEX exchanges. Overall, the funds make 141.85 million trades of 1.11 trillion shares valued at 34.26 trillion dollars USD. Puckett and Yan (2011) estimate that the ANcerno institutions account for approximately 8% of the total dollar value of CRSP trading volume between 1999 and 2005. More detailed summary statistics are provided in Table A.1 below.

Puckett and Yan (2011) obtains a sample of 64 ANcerno client institution names which include pension plan sponsors such as CalPERS, the Commonwealth of Virginia, the YMCA retirement fund, money managers such as Massachusetts Financial Services, Putnam Investments, Lazard Asset Management, and several broker-dealers. Although ANcerno does not report the identities of their clients, the dataset does include the names of their clients’ money managers and brokers. These money managers include investment management companies such as AllianceBernstein, Black-Rock, and Wellington as well as investment banks such as Goldman Sachs, JP Mor-
Table A.1: **ANCerno Data (1999Q1–2012Q1).** A fund is defined as a unique *clientcode–managercode* pair. The sample of stocks traded includes all common stock listed on the NYSE, AMEX, and NASDAQ exchanges which I am able to match to CRSP PERMNOs using CUSIPs and tickers. Brokers are identified by *brokercode*. Prices and volume are adjusted for stock splits using adjustment factors from the CRSP database.

<table>
<thead>
<tr>
<th>Year</th>
<th># of Funds</th>
<th># Stocks Traded</th>
<th># Brokers</th>
<th># of Trades</th>
<th>Volume Traded</th>
<th>Dollar Value Traded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>1,992</td>
<td>4,921</td>
<td>651</td>
<td>1.98M</td>
<td>18.17B</td>
<td>904.88B</td>
</tr>
<tr>
<td>2000</td>
<td>1,901</td>
<td>4,814</td>
<td>641</td>
<td>3.15M</td>
<td>32.67B</td>
<td>1.72T</td>
</tr>
<tr>
<td>2001</td>
<td>1,996</td>
<td>4,557</td>
<td>693</td>
<td>3.59M</td>
<td>40.97B</td>
<td>1.32T</td>
</tr>
<tr>
<td>2002</td>
<td>2,105</td>
<td>4,348</td>
<td>747</td>
<td>3.57M</td>
<td>39.81B</td>
<td>1.02T</td>
</tr>
<tr>
<td>2003</td>
<td>2,068</td>
<td>4,315</td>
<td>729</td>
<td>3.77M</td>
<td>34.45B</td>
<td>859.37B</td>
</tr>
<tr>
<td>2004</td>
<td>1,950</td>
<td>4,256</td>
<td>671</td>
<td>4.42M</td>
<td>33.64B</td>
<td>985.22B</td>
</tr>
<tr>
<td>2005</td>
<td>1,612</td>
<td>4,057</td>
<td>687</td>
<td>4.39M</td>
<td>31.42B</td>
<td>1.02T</td>
</tr>
<tr>
<td>2006</td>
<td>1,404</td>
<td>3,984</td>
<td>665</td>
<td>7.32M</td>
<td>47.10B</td>
<td>1.55T</td>
</tr>
<tr>
<td>2007</td>
<td>1,296</td>
<td>3,928</td>
<td>638</td>
<td>8.64M</td>
<td>50.87B</td>
<td>1.84T</td>
</tr>
<tr>
<td>2008</td>
<td>1,285</td>
<td>3,771</td>
<td>668</td>
<td>7.32M</td>
<td>57.82B</td>
<td>1.71T</td>
</tr>
<tr>
<td>2009</td>
<td>1,304</td>
<td>3,712</td>
<td>648</td>
<td>6.49M</td>
<td>58.47B</td>
<td>1.28T</td>
</tr>
<tr>
<td>2010</td>
<td>1,220</td>
<td>3,350</td>
<td>652</td>
<td>5.19M</td>
<td>29.57B</td>
<td>870.19B</td>
</tr>
<tr>
<td>2011</td>
<td>1,233</td>
<td>3,387</td>
<td>659</td>
<td>4.80M</td>
<td>17.47B</td>
<td>561.19B</td>
</tr>
<tr>
<td>2012</td>
<td>238</td>
<td>2,657</td>
<td>186</td>
<td>506.477</td>
<td>1.45B</td>
<td>50.49B</td>
</tr>
</tbody>
</table>

gan, and Deutsche Bank.\(^2\) Broker-dealers include independent broker-dealers such as LPL Financial, Raymond James Financial Services, Ameriprise Financial Services as well as dealer banks such as Bank of America, Citigroup, and Credit Suisse.

### A.4 Propensity Score Matching

\(^2\)The database also includes the Carlyle Group, KKR & Co., and Blackstone Group LP which were formerly external managers investing on behalf of CalPERS, an ANcerno client. Source: “Calpers to Cut External Money Managers by Half.” *WSJ.* Web. 21 July 2015.
Table A.2: **Normalized Differences.** The propensity score matched sample of central and peripheral funds is based on the previous quarter’s Degree, Eigenvector, number of trades, share volume traded, and principal-weighted (equal-weighted) average abnormal interim trading performance. I require that the difference in propensity scores does not exceed 0.1 basis points in absolute value to ensure covariate balance in the matched sample. Imbens and Wooldridge (2009) suggest a rule of thumb threshold for normalized differences of 0.25, above which linear regression methods tend to be affected by covariate balance.

<table>
<thead>
<tr>
<th></th>
<th>Peripheral</th>
<th>Central</th>
<th>Norm. Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>3.73</td>
<td>4.16</td>
<td>0.24</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.0025</td>
<td>0.0027</td>
<td>0.021</td>
</tr>
<tr>
<td>Trades</td>
<td>382.6</td>
<td>872.7</td>
<td>0.079</td>
</tr>
<tr>
<td>Volume</td>
<td>2970581.3</td>
<td>6701747.0</td>
<td>0.052</td>
</tr>
<tr>
<td>PW</td>
<td>0.0033</td>
<td>0.0079</td>
<td>0.048</td>
</tr>
<tr>
<td>EW</td>
<td>0.0038</td>
<td>0.0090</td>
<td>0.059</td>
</tr>
<tr>
<td>N</td>
<td>5267</td>
<td>4982</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

What does the PIN model identify as private information?

B.1 The DY model

Duarte and Young (2009) propose an extension of the PIN model that accounts for the positive correlation between buys and sells. We show in this Appendix that the Duarte and Young (2009) model also performs poorly late in our sample from 1993–2012.

B.1.1 The DY model

Duarte and Young (2009) extend the PIN model to address some of its shortcomings in matching the order flow data. Specifically, the authors note that the PIN model implies that the number of buys and sells are negatively correlated; however, in the data the correlation between the number of buys and sells is overwhelmingly positive. To correct this problem, the DY model partially disentangles turnover variation from private information arrival. As in the PIN model, the DY model posits that at the beginning of each day, informed investors receive a private signal with probability $\alpha$. If the private signal is positive, buy orders from the informed traders arrive according to a Poisson distribution with intensity $\mu_B$. If the private
signal is negative, informed sell orders arrive according to a Poisson distribution with intensity $\mu_S$. If the informed traders receive no private signal, they do not trade.

In contrast to the PIN model, the DY model allows for symmetric order flow shocks. These shocks increase both the number of buyer- and seller-initiated trades but are unrelated to private information events. Symmetric order flow shocks can happen for a variety of reasons, such as disagreement among traders about the interpretation of public news. Alternatively, liquidity shocks may occur that cause investors holding different collections of assets to simultaneously rebalance their portfolios, resulting in increases to both buys and sells. Regardless of the mechanism, symmetric order flow shocks arrive on any given day with probability $\theta$. On days with symmetric order flow shocks, both the number of buyer- and seller-initiated trades increase by amounts drawn from independent Poisson distributions with intensity $\Delta_B$ or $\Delta_S$, respectively. Buy and sell orders from uninformed traders arrive according to a Poisson distribution with intensities $\epsilon_B (\epsilon_B + \Delta_B)$ and $\epsilon_S (\epsilon_S + \Delta_S)$ on days without (with) symmetric order flow shocks. Fig. B.1 shows the structure of the DY model.

Under the DY model, turnover can increase due to either symmetric order flow shocks or the arrival of private information. To see this, note that the expected number of buys plus sells on days with positive (negative) information and without symmetric order flow shocks is $\epsilon_B + \epsilon_S + \mu_B (\epsilon_B + \epsilon_S + \mu_S)$; the expected number of trades on days with symmetric order flow shocks and without private information shocks is $\epsilon_B + \epsilon_S + \Delta_B + \Delta_S$, and the expected number of trades is $\epsilon_B + \epsilon_S$ on days without either.
B.1.2 Estimation of the DY model

As with the PIN model, we estimate the DY model numerically via maximum likelihood. Let $\Theta_{DY,i} = (\alpha_i, \mu_{B_i}, \mu_{S_i}, \epsilon_{B_i}, \epsilon_{S_i}, \delta_i, \Delta_{B_i}, \Delta_{S_i})$ be the vector of parameters of the DY model for stock $i$. Let $B_{i,t}$ and $S_{i,t}$ be the number of buys and sells, respectively, for stock $i$ on day $t$. Let $D_{DY,i,t} = [B_{i,t}, S_{i,t}, \Theta_{DY,i}]$. The likelihood function of the extended model is $\prod_{t=1}^{T} L(D_{DY,i,t})$:

$$L(D_{DY,i,t}) = L_{NI,NS}(D_{DY,i,t}) + L_{NI,S}(D_{DY,i,t}) + L_{I-,NS}(D_{DY,i,t})$$
$$+ L_{I-,S}(D_{DY,i,t}) + L_{I+,NS}(D_{DY,i,t}) + L_{I+,S}(D_{DY,i,t})$$

where $L_{NI,NS}(D_{DY,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information or a symmetric order flow shock; $L_{NI,S}(D_{DY,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day without private information but with a symmetric order flow shock; $L_{I-,NS}$ ($L_{I-,S}$) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information and without (with) a symmetric order flow shock; and $L_{I+,NS}$ ($L_{I+,S}$) is the probability on a day with positive information and without (with) a symmetric order flow shock. Analogous to the original PIN model, each term in the likelihood function corresponds to a branch in the tree in Fig. B.1 and each term is given by:
\begin{align}
L_{NI,NS}(D_{DY,i,t}) &= (1 - \alpha_i)(1 - \theta_i)e^{-\epsilon B_i} \frac{e^{B_i,t}}{B_{i,t}!} e^{-\epsilon S_i} \frac{S_i,t}{S_{i,t}!} \\
L_{NI,S}(D_{DY,i,t}) &= (1 - \alpha_i) \theta_i e^{-\epsilon B_i + \Delta B_i} \frac{(\epsilon B_i + \Delta B_i)B_{i,t}}{B_{i,t}!} e^{-(\epsilon S_i + \Delta S_i)} \frac{S_i + \Delta S_i}{S_{i,t}!} \\
L_{I-,NS}(D_{DY,i,t}) &= \alpha_i(1 - \theta_i)(1 - \delta_i) e^{-\epsilon B_i} \frac{e^{B_i,t}}{B_{i,t}!} e^{-(\mu S_i + \epsilon S_i)} \frac{(\mu S_i + \epsilon S_i)S_i,t}{S_{i,t}!} \\
L_{I-,S}(D_{DY,i,t}) &= \alpha_i(1 - \theta_i) \delta_i e^{-\epsilon B_i + \Delta B_i} \frac{(\epsilon B_i + \Delta B_i)B_{i,t}}{B_{i,t}!} e^{-(\mu S_i + \epsilon S_i + \Delta S_i)} \frac{(\mu S_i + \epsilon S_i + \Delta S_i)S_i,t}{S_{i,t}!} \\
L_{I+,NS}(D_{DY,i,t}) &= \alpha_i(1 - \theta_i) \delta_i e^{-(\mu S_i + \epsilon S_i)} \frac{(\mu B_i + \epsilon B_i)B_{i,t}}{B_{i,t}!} e^{-(\mu S_i)} \frac{S_i,t}{S_{i,t}!} \\
L_{I+,S}(D_{DY,i,t}) &= \alpha_i(1 - \theta_i) \delta_i e^{-(\mu S_i + \epsilon S_i + \Delta B_i)} \frac{(\mu B_i + \epsilon B_i + \Delta B_i)B_{i,t}}{B_{i,t}!} e^{-(\mu S_i + \Delta S_i)} \frac{(\mu S_i + \Delta S_i)S_i,t}{S_{i,t}!}
\end{align}

In order to avoid local optima, we use the maximum of the likelihood maximization with ten different starting points as in Duarte and Young (2009). In addition, for one of the starting points we choose \((\epsilon_B, \epsilon_S)\) values, and \((\epsilon_B + \Delta B, \epsilon_S + \Delta S)\) equal to the sample means of buys and sells computed by the k-means algorithm with \(k=2\). The k-means algorithm looks for clusters in the buys and sells such that each observation belongs to the cluster with the nearest mean. Because we know a priori that buys and sells have a strong positive correlation (see Duarte and Young (2009)), we partition the sample into high and low order flow clusters, which correspond to the symmetric order flow shock/no symmetric order flow shock states in the DY model. The other nine starting points are randomized. This procedure ensures that at least one of the starting points is centered properly, as the numerical likelihood estimation using purely random starts often stops at points outside of the central clusters of
B.1.3 $CPIE_{DY}$

As with the PIN model, for each stock-day, we compute the probability of an information event conditional on both the model parameters and on the number of buys and sells observed that day. Specifically, let the indicator $I_{i,t}$ take the value of one if an information event occurs for stock $i$ on day $t$ and zero otherwise. We compute $CPIE_{DY,i,t} = P[I_{i,t} = 1|D_{DY,i,t}]$ as:

$$CPIE_{DY,i,t} = \frac{L_{I^+,NS}(D_{DY,i,t}) + L_{I^+,S}(D_{DY,i,t}) + L_{I^-,S}(D_{DY,i,t}) + L_{I^-,NS}(D_{DY,i,t})}{L(D_{DY,i,t})}$$

(B.8)

Analogous to the PIN model, the Adj. PIN of a stock is $\frac{\alpha(\delta \mu B + (1-\delta) \mu S)}{\alpha(\delta \mu B + (1-\delta) \mu S) + \epsilon B + \epsilon S + \theta(\Delta B + \Delta S)}$. This is the unconditional probability that any given trade is initiated by an informed trader. $CPIE_{DY}$ and Adj. PIN are linked via the unconditional probability of an information event, $\alpha$, which is also the unconditional expectation of $CPIE_{DY}$.

Table B.1 contains summary statistics for the parameter estimates for the DY model as well as summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{DY}$. We see that the mean $CPIE$ behaves exactly like $\alpha$. Hence, changes in $CPIE_{DY}$ and changes in the estimated alphas are analogous.

B.1.4 How does the DY model identify private information?

To illustrate how the $CPIE_{DY}$ works, we present a stylized example of the DY model in Fig. B.2. In Panel A we plot simulated and real order flow data for Exxon-Mobil...
during 1993, with buys on the horizontal axis and sells on the vertical axis. Real data are marked as +, and simulated data as transparent dots. The real data are shaded according to the\textit{CPIE}, with lighter points (+) representing low and darker points (+) high CPIEs.

The DY model generates six data clusters, greatly improving upon the PIN model’s coverage of the data in 1993. The two clusters on the dotted line are not related to private information, but the other four clusters are. An econometrician using the DY model, moving along the dotted line, would observe that high turnover–considered information days under the PIN model–are no longer classified as such, because higher turnover may be driven by symmetric order flow shocks under the DY model. Instead, the DY model identifies private information when moving away from the dotted line; when buys are greater than sells and vice versa.

Unfortunately late in the sample the DY model breaks down. Panel B of Fig. B.2 shows that the DY model, like the PIN model, fails to fit the majority of the order flow data for Exxon-Mobil in 2012. The problem of fitting the data is not limited to our stylized example. Fig. B.3 shows that after 2005 the DY model estimates that the total likelihood for 80% of the order flow data of the median stock is less than $10^{-10}$.

As a more formal test of the DY model, Table B.2 presents regressions of $CPIE_{DY}$ based on simulated and real data. The right-hand side variables are the absolute order imbalance adjusted for buy/sell correlations ($|adj.OIB|$), turnover and its squared term. We define the adjusted absolute order imbalance as the absolute value of the residual from a regression of buys on sells. We use this measure to analyze the DY
model because, as Fig. B.2 suggests, the DY model implies that days with information events are far from the dashed line in this figure.\textsuperscript{1} Turnover, as before, is defined as the sum of buys and sells. We report median coefficient estimates and $t$-statistics across all firms within a particular year. The coefficients are standardized as above. We report the average of the median, the $5^{th}$, and the $95^{th}$ percentiles of the $R^2$s and $R^2_{incs}$.

As with the $CPIE_{PIN}$, in theory, turnover has little additional power in explaining $CPIE_{DY}$. The incremental $R^2$s in Table B.2 Panel A are low with an average value close to 4%. This is smaller than the average incremental $R^2$s of the PIN model. The intuition for this result is that the DY model disentangles turnover and order flow shocks by including the possibility of symmetric order flow shocks. Buying and selling activity can simultaneously be higher than average, but this is not indicative of private information unless there is a large order flow imbalance.

Panel B of Table B.2 reports regression results for the real, rather than simulated, data. The DY model behaves very differently when using real data as opposed to data generated from the model. The $R^2$s for the real data are much lower than those in the simulated data, declining from 35% in 1993 to 12% in 2012. The $p$-values (frequency of rejection) also decreases (increases) over time. For example, in 1993, our hypothesis test based on $R^2$ rejects the model at 5% significance for 81% of the stocks, while in 2012 this percentage increases to around 95%. The incremental $R^2$ indicate that turnover and turnover squared explain a large degree of variation in

\textsuperscript{1}Our results are qualitatively similar if we use absolute order imbalance instead of adjusted absolute order imbalance.
Indeed, the average ratio of the median $R^2$s, $R^2/(R^2 + R^2_{inc})$, is about $40\%$.

**B.2 Estimating Order Flow, $r_{o,i,t}$ and $r_{d,i,t}$**

Wharton Research Data Services (WRDS) provides trades matched to National Best Bid and Offer (NBBO) quotes at 0, 1, 2, and 5 second delay intervals. We use only “regular way” trades, with original time and/or corrected timestamps to avoid incorrect quotes or non-standard settlement terms, for instance, trades that are settled in cash or settled the next business day.\(^2\) Prior to 2000, we match “regular way” trades to quotes delayed for 5 seconds; between 2000 and 2007, we match trades to quotes delayed for 1 second; and after 2007, we match trades to quotes without any delay.

We classify the matched trades as either buys or sells following the Lee and Ready (1991) algorithm, which classifies all trades occurring above (below) the bid-ask mid-point as buyer (seller) initiated. We use a tick test to classify trades that occur at the mid-point of the bid and ask prices. The tick test classifies trades as buyer (seller) initiated if the price was above/(below) that of the previous trade.

To estimate $r_{o,i,t}$ and $r_{d,i,t}$, we run daily cross-sectional regressions of overnight and intraday returns on a constant, historical $\beta$ (based on the previous 5 years of monthly CRSP returns), log market cap, log book-to-market (following Fama and French (1992), Fama and French (1993), and Davis, Fama, and French (2000)). We impose min/max values for book equity (before taking logs) of 0.017 and 3.13,

\(^2\)Trade COND of ("@","*", or ") and CORR of (0,1)
respectively. If book equity is negative, we set it to 1 before taking logs, so that it is zero after taking logs. We use the residuals from these daily cross-sectional regressions, winsorized at the 1 and 99% levels as our idiosyncratic intraday and overnight returns.

B.3 Details about the PIN model

B.3.1 PIN Likelihood

The likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without an information event, on a day with positive information event, and on a day with a negative information event are:

\begin{align*}
L_{NI}(D_{PIN,i,t}) &= (1 - \alpha_i)e^{-\epsilon_{B_i}}\frac{B_{i,t}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{S_{i,t}}{S_{i,t}!} \\
L_{I+}(D_{PIN,i,t}) &= \alpha_i \bar{\delta} e^{-(\mu_i + \epsilon_{B_i})}\frac{(\mu_i + \epsilon_{B_i})B_{i,t}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{S_{i,t}}{S_{i,t}!} \\
L_{I-}(D_{PIN,i,t}) &= \alpha_i (1 - \bar{\delta})e^{-\epsilon_{B_i}}\frac{B_{i,t}}{B_{i,t}!}e^{-(\mu_i + \epsilon_{i,S})}\frac{(\mu_i + \epsilon_{i,S})S_{i,t}}{S_{i,t}!}
\end{align*}

where $L_{NI,NS}(D_{DY,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information trading or symmetric order flow shock; $L_{NI,S}(D_{DY,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day without private information and with a symmetric order flow shock; $L_{I-,NS}$ ($L_{I-,S}$) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information and without (with) symmetric order flow shock; $L_{I+,NS}$ ($L_{I+,S}$) is the probability on a day with positive information and without (with) a symmetric
order flow shock.

B.3.2 Maximum likelihood procedure

To estimate the PIN likelihood function, we use the maximum of the likelihood maximization with ten different starting points as in Duarte and Young (2009). We note, however, that late in the sample, the likelihood functions of the PIN are very close to zero. After 2006, the PIN model suggests that 90% of the observed daily order flows for the median stock have a near-zero probability (i.e. smaller than $10^{-10}$) of occurring. This makes the estimation susceptible to local optima. To get around this problem, we choose one of our ten starting points to be such that the PIN model clusters are close to the observed mean of the number of buys and sells. Specifically, we choose $\varepsilon_B$ and $\varepsilon_S$ values equal to the sample means of buys and sells, $\alpha$ equal to 1%, and delta equal to the mean absolute value of order imbalance. The other nine starting points are randomized. We do this in order to ensure that at least one of the starting points is centered properly, as the numerical likelihood estimation using purely random starts often stops at points outside of the central cluster of data.

B.3.3 Computing $CPIE_{PIN}$

In Section 2 of the paper, we define the $CPIE$ as the ratio of the “news” likelihood functions to the sum total of the likelihood functions. In practice, there are many cases in the PIN model for which the data a near-zero probability of occurring, meaning $L_{NI}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t}) + L_{I^-}(D_{PIN,i,t})$ is smaller than $10^{-10}$. As a result the $CPIE$ ratio frequently results in a divide by zero error.
In order to compute \( CPIE \) for these days, we “center” the likelihoods around the state with the highest log-likelihood before computing the \( CPIE \). For example, consider the PIN model with:

\[
L_{\text{max}} = \max\{L_{NI}, L_{I+}, L_{I-}\}, \quad (B.12)
\]

\[
\ell_{\text{max}} = \log(L_{\text{max}}) \quad (B.13)
\]

where \( \ell \) represents the log of the corresponding likelihood function. We compute the centered versions of each of the likelihood functions:

\[
\ell'_{NI} = \ell_{NI} - \ell_{\text{max}}, \quad \text{(B.14)}
\]

\[
\ell'_{I+} = \ell_{I+} - \ell_{\text{max}}, \quad \text{(B.15)}
\]

\[
\ell'_{I-} = \ell_{I-} - \ell_{\text{max}}. \quad \text{(B.16)}
\]

We compute the \( CPIE' \) as:

\[
CPIE'_{PIN} = \frac{L'_{I+} + L'_{I-}}{L'_{NI} + L'_{I+} + L'_{I-}} \quad (B.17)
\]

such that the most likely state has \( L' = 1 \). For a high turnover day, it may be the case that \( L'_{I+} = 1, L'_{I-} = 0 \) and \( L'_{NI} = 0 \); hence, the \( CPIE' \) will be 1. We follow a similar procedure to compute \( CPIE'_{DY} \).
B.3.4 $CPIE_{PIN}$ of M&A targets around announcements

Aktas, de Bodt, Declerck, and Van Oppens (2007) find that $PIN$ is higher after merger announcements than before, partially as a result of increases in PIN model’s $\alpha$. In this section we show that their results are related to our main finding that the PIN model identifies private information from turnover.

We examine the period $t \in [-30, 30]$ around the event. To do so, we estimate the parameter vector $\Theta_{PIN,i}$ in the period $t \in [-312, -60]$ before the event and then compute the daily $CPIE$s for the period $t \in [-30, 30]$ surrounding the announcement.

Panel A of Fig. B.4 shows the average $CPIE_{PIN}$ in event time for our sample of M&A targets. The graph shows that, under the PIN model, the probability of an information event increases prior to the event, starting at around 55% 20 days before the announcement and peaking around 80% on the after day of the announcement. The rise in the probability of an information event prior to the announcement is consistent with a world where informed traders generate signals about potential mergers and acquisitions and trade on this information before the events are announced to the public. However, $CPIE_{PIN}$ is also higher after the actual announcements become public information. In fact, $CPIE_{PIN}$ remains above the average $CPIE_{PIN}$ observed in the gap period, $[-60, -31]$, for 20 trading days after the announcement.

Panels B and C of Fig. B.4 shed light on the features of the data that produce the observed pattern in the average $CPIE_{PIN}$ in Panel A. Panel B shows the average
predictions from OLS regressions of $CPIE_{PIN}$ on order imbalance and absolute order imbalance squared across all of the stocks in the event study sample. The solid line indicates that order imbalance explains only a small fraction of the movement in $CPIE_{PIN}$ during the event window. Panel C shows the average predictions from regressions of $CPIE_{PIN}$ on turnover and turnover squared. The solid line indicates that the variation in $CPIE_{PIN}$ around M&A announcements is explained almost entirely by turnover. The intuition follows directly from the main results, which illustrates that $CPIE_{PIN}$ is mechanically driven by turnover increases. The higher post-event turnover levels are enough to keep $CPIE_{PIN}$ above its pre-event mean for a substantial period.

B.4 Details about the EPIN model

The EPIN model extends the PIN model to allow for continuous variation in turnover unrelated to private information arrival.

B.4.1 Negative binomial distribution in EPIN model

The probability that $B + S$ is equal to $x$ in a given day is:

$$f(x; r, p) = \int_0^{\infty} \frac{\lambda^x}{x!} \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{(1-p^r)^r \Gamma(r)} d\lambda = \frac{(1-p)^r p^{-r}}{\Gamma(r)} p^{r+x} \Gamma(r + x)$$ (B.18)
B.5 Details about the OWR model

B.5.1 OWR Likelihood

Let \( \Theta_{OWR,i} = (\alpha_i, \sigma_{u_i}, \sigma_{z_i}, \sigma_{p.d_i}, \sigma_{p.o_i}) \) be the vector of parameters of this model. The parameter \( \alpha_i \) is the probability that there is an information event on a given day. \( \sigma_{z_i}^2 \) is the variance of the noise of the observed net order flow \( (y_e) \); \( \sigma_{u_i}^2 \) is the variance of the net order flow from noise traders; \( \sigma_{d_i}^2 \) is the variance of the private signal received by the informed trader; \( \sigma_{p.d_i}^2 \) is the variance of the intraday return; \( \sigma_{p.o_i}^2 \) is the variance of the overnight return.

The likelihood of observing \( D_{OWR,i,t} \) on a day without and with an information event is:

\[
L_{NI} = (1 - \alpha) f_{NI}(D_{OWR,i,t}) \quad \text{(B.19)}
\]
\[
L_I = \alpha f_I(D_{OWR,i,t}) \quad \text{(B.20)}
\]

where \( f_{NI}(D_{OWR,i,t}) \) is the joint probability density of \((y_{e,i,t}, r_{o,i,t}, r_{d,i,t})\) on days without information, \( f_I(D_{OWR,i,t}) \) is the density of \((y_{e,t}, r_{o,t}, r_{d,t})\) on days with information events. Both \( f_{NI}(D_{OWR,i,t}) \) and \( f_I(D_{OWR,i,t}) \) are multivariate normal with zero means and covariance matrices \( \Omega_{NI} \) and \( \Omega_I \). The covariance matrix \( \Omega_{NI} \) has ele-
\[ \text{Var}(y_e) = \sigma_u^2 + \sigma_z^2, \quad \text{(B.21)} \]
\[ \text{Var}(r_d) = \sigma_{pd}^2 + \alpha \sigma_t^2/4, \quad \text{(B.22)} \]
\[ \text{Var}(r_o) = \sigma_{po}^2 + \alpha \sigma_t^2/4, \quad \text{(B.23)} \]
\[ \text{Cov}(r_d, r_o) = -\alpha \sigma_t^2/4, \quad \text{(B.24)} \]
\[ \text{Cov}(r_d, y_e) = \alpha^{1/2} \sigma_i \sigma_u/2, \quad \text{(B.25)} \]
\[ \text{Cov}(r_o, y_e) = -\alpha^{1/2} \sigma_i \sigma_u/2 \quad \text{(B.26)} \]

And \( \Omega_I \):\[
\text{Var}(y_e) = (1 + 1/\alpha) \sigma_u^2 + \sigma_z^2, \quad \text{(B.27)}
\]
\[ \text{Var}(r_d) = \sigma_{pd}^2 + (1 + \alpha) \sigma_t^2/4, \quad \text{(B.28)} \]
\[ \text{Var}(r_o) = \sigma_{po}^2 + (1 + \alpha) \sigma_t^2/4, \quad \text{(B.29)} \]
\[ \text{Cov}(r_d, r_o) = (1 - \alpha) \sigma_t^2/4, \quad \text{(B.30)} \]
\[ \text{Cov}(r_d, y_e) = \alpha^{-1/2} \sigma_i \sigma_u/2 + \alpha^{1/2} \sigma_i \sigma_u/2, \quad \text{(B.31)} \]
\[ \text{Cov}(r_o, y_e) = \alpha^{-1/2} \sigma_i \sigma_u/2 - \alpha^{1/2} \sigma_i \sigma_u/2 \quad \text{(B.32)} \]

**B.5.2 How does the OWR model identify private information?**

In theory, the OWR model identifies private information from the covariance matrix of the three variables in the model \((y_{e,i,t}, r_{o,i,t}, r_{d,i,t})\). To analyze the model, we run
the regression of $CPIE_{OWR}$ on the squared and interaction terms of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$:

$$CPIE_{OWR,i,t} = \beta_0 + \beta_1 y_{e,i,t}^2 + \beta_2 r_{d,i,t}^2 + \beta_3 r_{o,i,t}^2 + \beta_4 y_{e,i,t} r_{d,i,t} + \beta_5 y_{e,i,t} r_{o,i,t} + \beta_6 r_{d,i,t} r_{o,i,t} + u_{i,t}. \tag{B.33}$$

Panel A of Table B.3 presents median coefficient estimates, $t$-statistics, and three percentiles of $R^2$s across all firms within a particular year using simulated data. The results highlight the intuition behind the model. The probability of an information event on any given day is increasing in the square of intraday returns, the interaction between imbalance and intraday (or overnight) returns, and the interaction between intraday and overnight returns. The coefficient estimates on the square of the order imbalance and on the square of overnight returns are too small to be precisely measured. The high $R^2$s indicate that, practically speaking, the square of intraday returns, the interaction between intraday and overnight returns and the interaction between intraday returns and order flow imbalance are sufficient to explain a large part of the variation in $CPIE_{OWR}$.

Panel B of Table B.3 shows the median coefficient estimates, $t$-statistics, and the results of the hypothesis tests based on $R^2$s across all firms within a particular year using real data. Unlike the PIN and DY models, the coefficient estimates are consistent across the simulated and real data. For instance in simulated data regressions in Panel A, 2008 is the only year in which $y_{e}^2$ is the most important term. In the real data regressions in Panel B, 2008 is also the only year in which $y_{e}^2$ is the most important term, indicating that the model matches the features of the data quite well, even for clear outliers like 2008. Furthermore, as with the simulated
data regressions, the high median $R^2$s indicate that a large part of the variation in $CPIE_{OWR}$ is explained by the squared and interaction terms of $(y_{e,t}, r_{o,t}, r_{d,t})$ as implied by the model. The average across years of the $R^2$s in Panel B is about 83% and these $R^2$s increase over time, reaching 90% in 2012. Moreover, we reject the null hypothesis that the $R^2$s observed in the real data are consistent with the OWR model at 5% level for about 40% of the sample in 1993 and for about 8% of the sample in 2012.

The high $R^2$s in Panel B imply that, in principle, any variable unrelated to private information under the OWR model has only a small incremental value in explaining the $CPIE_{OWR}$. To see this note that the typical $R^2$ in Panel B is around 85%. This suggests that any additional regressor, even if it explained 100% of the residual variation in the regressions in Panel B, could only marginally improve the $R^2$ from 85% to 100%. Note that in the case of the PIN and DY models, our results show that turnover, which in principle is a poor measure of private information, largely drives the PIN and DY models’ identification of private information. In contrast, under the OWR model the variables related to private information in the model (squares and interactions of $y_{e}$, $r_{o}$, and $r_{d}$) can explain a fairly large amount of the variation in $CPIE_{OWR}$. As a result, any variable that is not related to private information in the OWR model can only explain a relatively small fraction of the variation in $CPIE_{OWR}$.

B.6 Tables and Figures
Table B.1: **DY Estimates.** This table summarizes parameter estimates of the DY model for 21,206 PERMNO-Year samples from 1993–2012. $\alpha$ represents the average unconditional probability of an information event at the daily level. $\epsilon_B$ and $\epsilon_S$ represent the expected number of daily buys and sells given no private information or symmetric order flow shocks. $\mu$, $\mu_B$, and $\mu_S$ represent the expected additional order flows given an information event, which is good news with probability $\delta$ and bad news with probability $1 - \delta$. A symmetric order flow shock occurs with probability $\theta$, in which case the expected number of buys and sells increase by $\Delta_B$ and $\Delta_S$, respectively. $\text{CPIE}$ and $\text{Std(CPIE)}$ are the PERMNO-Year mean and standard deviation of $\text{CPIE}$.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>21,206</td>
<td>0.456</td>
<td>0.092</td>
<td>0.409</td>
<td>0.464</td>
<td>0.509</td>
</tr>
<tr>
<td>$\delta$</td>
<td>21,206</td>
<td>0.550</td>
<td>0.192</td>
<td>0.441</td>
<td>0.541</td>
<td>0.680</td>
</tr>
<tr>
<td>$\theta$</td>
<td>21,206</td>
<td>0.249</td>
<td>0.137</td>
<td>0.149</td>
<td>0.253</td>
<td>0.344</td>
</tr>
<tr>
<td>$\epsilon_B$</td>
<td>21,206</td>
<td>1.418</td>
<td>4.571</td>
<td>26</td>
<td>158</td>
<td>866</td>
</tr>
<tr>
<td>$\epsilon_S$</td>
<td>21,206</td>
<td>1.397</td>
<td>4.570</td>
<td>28</td>
<td>148</td>
<td>807</td>
</tr>
<tr>
<td>$\Delta_B$</td>
<td>21,206</td>
<td>2.148</td>
<td>10.058</td>
<td>41</td>
<td>190</td>
<td>989</td>
</tr>
<tr>
<td>$\Delta_S$</td>
<td>21,206</td>
<td>2.097</td>
<td>9.934</td>
<td>34</td>
<td>160</td>
<td>908</td>
</tr>
<tr>
<td>$\mu_B$</td>
<td>21,206</td>
<td>290</td>
<td>575</td>
<td>29</td>
<td>119</td>
<td>310</td>
</tr>
<tr>
<td>$\mu_S$</td>
<td>21,206</td>
<td>284</td>
<td>574</td>
<td>27</td>
<td>107</td>
<td>302</td>
</tr>
<tr>
<td>$\text{CPIE}$</td>
<td>21,206</td>
<td>0.455</td>
<td>0.092</td>
<td>0.409</td>
<td>0.461</td>
<td>0.506</td>
</tr>
<tr>
<td>$\text{Std(CPIE)}$</td>
<td>21,206</td>
<td>0.454</td>
<td>0.056</td>
<td>0.431</td>
<td>0.479</td>
<td>0.493</td>
</tr>
</tbody>
</table>
Table B.2: DY Model Regressions. This table reports real and simulated regressions of the \( CPIE_{DY} \) on absolute adjusted order imbalance (\(|\text{adj. OIB}|\)), and absolute adjusted order imbalance squared (\(|\text{adj. OIB}|^2\)). In Panel A, we simulate 1,000 instances of the PIN model for each \( \text{PERMNO-YEAR} \) in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the \( R^2 \) (\( R^2_{\text{inc.}} \)) values. We compute the incremental \( R^2_{\text{inc.}} \) as the \( R^2 \) attributed to \( \text{turn} \) and \( \text{turn}^2 \) in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median \( R^2 \) values, and tests of the hypothesis that the observed variation in \( CPIE_{DY} \) is consistent with the DY model. The \( p \)-value of \( R^2 \) (\( R^2_{\text{inc.}} \)) is the probability of observing an \( R^2 \) at least as small (large) as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

<table>
<thead>
<tr>
<th>Year</th>
<th>( \beta ) (adj. OIB)</th>
<th>( \beta ) (adj. OIB)^2</th>
<th>( t ) (adj. OIB)</th>
<th>( t ) (adj. OIB)^2</th>
<th>5%</th>
<th>50%</th>
<th>95%</th>
<th>5%</th>
<th>50%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>0.518</td>
<td>-0.230</td>
<td>(10.88)</td>
<td>(4.74)</td>
<td>52.28%</td>
<td>59.44%</td>
<td>66.01%</td>
<td>5.55%</td>
<td>9.86%</td>
<td>15.29%</td>
</tr>
<tr>
<td>1994</td>
<td>0.484</td>
<td>-0.214</td>
<td>(10.47)</td>
<td>(4.42)</td>
<td>50.66%</td>
<td>58.06%</td>
<td>64.97%</td>
<td>5.56%</td>
<td>9.46%</td>
<td>14.95%</td>
</tr>
<tr>
<td>1995</td>
<td>0.475</td>
<td>-0.214</td>
<td>(9.96)</td>
<td>(4.32)</td>
<td>46.81%</td>
<td>54.46%</td>
<td>61.69%</td>
<td>7.01%</td>
<td>11.71%</td>
<td>17.54%</td>
</tr>
<tr>
<td>1996</td>
<td>0.516</td>
<td>-0.229</td>
<td>(10.54)</td>
<td>(4.60)</td>
<td>51.36%</td>
<td>58.62%</td>
<td>65.21%</td>
<td>5.18%</td>
<td>9.09%</td>
<td>14.31%</td>
</tr>
<tr>
<td>1997</td>
<td>0.513</td>
<td>-0.221</td>
<td>(10.33)</td>
<td>(4.40)</td>
<td>50.55%</td>
<td>57.80%</td>
<td>64.50%</td>
<td>4.78%</td>
<td>8.57%</td>
<td>14.03%</td>
</tr>
<tr>
<td>1998</td>
<td>0.537</td>
<td>-0.236</td>
<td>(10.60)</td>
<td>(4.49)</td>
<td>52.85%</td>
<td>60.14%</td>
<td>66.63%</td>
<td>4.00%</td>
<td>7.45%</td>
<td>12.31%</td>
</tr>
<tr>
<td>1999</td>
<td>0.607</td>
<td>-0.281</td>
<td>(11.92)</td>
<td>(5.45)</td>
<td>56.53%</td>
<td>63.49%</td>
<td>69.68%</td>
<td>3.07%</td>
<td>6.11%</td>
<td>10.47%</td>
</tr>
<tr>
<td>2000</td>
<td>0.597</td>
<td>-0.272</td>
<td>(11.43)</td>
<td>(5.09)</td>
<td>55.69%</td>
<td>62.59%</td>
<td>69.09%</td>
<td>2.82%</td>
<td>5.65%</td>
<td>9.73%</td>
</tr>
<tr>
<td>2001</td>
<td>0.729</td>
<td>-0.350</td>
<td>(13.81)</td>
<td>(6.75)</td>
<td>65.81%</td>
<td>71.48%</td>
<td>76.83%</td>
<td>0.62%</td>
<td>1.87%</td>
<td>4.09%</td>
</tr>
<tr>
<td>2002</td>
<td>0.769</td>
<td>-0.371</td>
<td>(15.03)</td>
<td>(7.28)</td>
<td>71.90%</td>
<td>76.37%</td>
<td>80.55%</td>
<td>0.24%</td>
<td>1.04%</td>
<td>2.41%</td>
</tr>
<tr>
<td>2003</td>
<td>0.805</td>
<td>-0.394</td>
<td>(16.06)</td>
<td>(7.99)</td>
<td>74.77%</td>
<td>78.95%</td>
<td>82.78%</td>
<td>0.34%</td>
<td>1.19%</td>
<td>2.71%</td>
</tr>
<tr>
<td>2004</td>
<td>0.798</td>
<td>-0.385</td>
<td>(15.94)</td>
<td>(7.61)</td>
<td>77.39%</td>
<td>81.40%</td>
<td>84.70%</td>
<td>0.23%</td>
<td>0.95%</td>
<td>2.22%</td>
</tr>
<tr>
<td>2005</td>
<td>0.787</td>
<td>-0.365</td>
<td>(16.23)</td>
<td>(7.40)</td>
<td>79.40%</td>
<td>83.08%</td>
<td>86.23%</td>
<td>0.25%</td>
<td>0.97%</td>
<td>2.26%</td>
</tr>
<tr>
<td>2006</td>
<td>0.761</td>
<td>-0.332</td>
<td>(15.52)</td>
<td>(6.74)</td>
<td>79.38%</td>
<td>83.00%</td>
<td>86.15%</td>
<td>0.45%</td>
<td>1.41%</td>
<td>2.88%</td>
</tr>
<tr>
<td>2007</td>
<td>0.736</td>
<td>-0.311</td>
<td>(12.97)</td>
<td>(5.97)</td>
<td>69.81%</td>
<td>74.50%</td>
<td>79.19%</td>
<td>1.23%</td>
<td>2.93%</td>
<td>5.99%</td>
</tr>
<tr>
<td>2008</td>
<td>0.755</td>
<td>-0.317</td>
<td>(15.14)</td>
<td>(6.52)</td>
<td>77.82%</td>
<td>81.67%</td>
<td>85.36%</td>
<td>0.34%</td>
<td>1.21%</td>
<td>2.82%</td>
</tr>
<tr>
<td>2009</td>
<td>0.768</td>
<td>-0.331</td>
<td>(16.09)</td>
<td>(7.01)</td>
<td>79.54%</td>
<td>83.16%</td>
<td>86.36%</td>
<td>0.63%</td>
<td>1.70%</td>
<td>3.51%</td>
</tr>
<tr>
<td>2010</td>
<td>0.769</td>
<td>-0.329</td>
<td>(15.95)</td>
<td>(7.01)</td>
<td>78.65%</td>
<td>82.63%</td>
<td>86.22%</td>
<td>0.56%</td>
<td>1.64%</td>
<td>3.66%</td>
</tr>
<tr>
<td>2011</td>
<td>0.754</td>
<td>-0.313</td>
<td>(15.47)</td>
<td>(6.73)</td>
<td>77.75%</td>
<td>81.79%</td>
<td>85.71%</td>
<td>0.63%</td>
<td>1.87%</td>
<td>4.16%</td>
</tr>
<tr>
<td>2012</td>
<td>0.763</td>
<td>-0.328</td>
<td>(15.65)</td>
<td>(7.01)</td>
<td>77.64%</td>
<td>81.93%</td>
<td>85.61%</td>
<td>0.89%</td>
<td>2.25%</td>
<td>4.69%</td>
</tr>
</tbody>
</table>
Table B.2: DY Model Regressions. Continued.

(b) Real Data

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$R^2$</th>
<th>$%$ Rej.</th>
<th>$p$-value</th>
<th>Rej.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>0.369</td>
<td>-0.170</td>
<td>(7.61)</td>
<td>(3.48)</td>
<td>34.07%</td>
<td>6.15%</td>
</tr>
<tr>
<td>1994</td>
<td>0.348</td>
<td>-0.150</td>
<td>(7.51)</td>
<td>(3.16)</td>
<td>33.55%</td>
<td>7.07%</td>
</tr>
<tr>
<td>1995</td>
<td>0.342</td>
<td>-0.149</td>
<td>(6.99)</td>
<td>(3.00)</td>
<td>30.15%</td>
<td>6.52%</td>
</tr>
<tr>
<td>1996</td>
<td>0.358</td>
<td>-0.164</td>
<td>(7.33)</td>
<td>(3.42)</td>
<td>31.11%</td>
<td>6.85%</td>
</tr>
<tr>
<td>1997</td>
<td>0.334</td>
<td>-0.140</td>
<td>(6.49)</td>
<td>(2.78)</td>
<td>28.00%</td>
<td>5.52%</td>
</tr>
<tr>
<td>1998</td>
<td>0.329</td>
<td>-0.136</td>
<td>(6.21)</td>
<td>(2.62)</td>
<td>26.26%</td>
<td>4.50%</td>
</tr>
<tr>
<td>1999</td>
<td>0.365</td>
<td>-0.166</td>
<td>(6.91)</td>
<td>(3.16)</td>
<td>27.89%</td>
<td>3.53%</td>
</tr>
<tr>
<td>2000</td>
<td>0.333</td>
<td>-0.145</td>
<td>(5.75)</td>
<td>(2.55)</td>
<td>23.49%</td>
<td>3.06%</td>
</tr>
<tr>
<td>2001</td>
<td>0.374</td>
<td>-0.176</td>
<td>(6.38)</td>
<td>(3.06)</td>
<td>25.25%</td>
<td>1.43%</td>
</tr>
<tr>
<td>2002</td>
<td>0.328</td>
<td>-0.130</td>
<td>(4.82)</td>
<td>(1.90)</td>
<td>21.31%</td>
<td>1.03%</td>
</tr>
<tr>
<td>2003</td>
<td>0.334</td>
<td>-0.135</td>
<td>(4.84)</td>
<td>(1.98)</td>
<td>21.55%</td>
<td>0.63%</td>
</tr>
<tr>
<td>2004</td>
<td>0.295</td>
<td>-0.104</td>
<td>(4.15)</td>
<td>(1.46)</td>
<td>18.31%</td>
<td>0.39%</td>
</tr>
<tr>
<td>2005</td>
<td>0.279</td>
<td>-0.103</td>
<td>(4.03)</td>
<td>(1.51)</td>
<td>16.23%</td>
<td>0.40%</td>
</tr>
<tr>
<td>2006</td>
<td>0.243</td>
<td>-0.083</td>
<td>(3.40)</td>
<td>(1.17)</td>
<td>12.46%</td>
<td>1.16%</td>
</tr>
<tr>
<td>2007</td>
<td>0.219</td>
<td>-0.086</td>
<td>(3.14)</td>
<td>(1.25)</td>
<td>9.66%</td>
<td>1.94%</td>
</tr>
<tr>
<td>2008</td>
<td>0.217</td>
<td>-0.086</td>
<td>(3.05)</td>
<td>(1.23)</td>
<td>8.83%</td>
<td>2.16%</td>
</tr>
<tr>
<td>2009</td>
<td>0.230</td>
<td>-0.093</td>
<td>(3.24)</td>
<td>(1.30)</td>
<td>10.04%</td>
<td>2.04%</td>
</tr>
<tr>
<td>2010</td>
<td>0.241</td>
<td>-0.103</td>
<td>(3.41)</td>
<td>(1.49)</td>
<td>10.59%</td>
<td>2.45%</td>
</tr>
<tr>
<td>2011</td>
<td>0.245</td>
<td>-0.102</td>
<td>(3.45)</td>
<td>(1.50)</td>
<td>10.35%</td>
<td>2.04%</td>
</tr>
<tr>
<td>2012</td>
<td>0.275</td>
<td>-0.127</td>
<td>(4.04)</td>
<td>(-1.86)</td>
<td>12.22%</td>
<td>2.54%</td>
</tr>
</tbody>
</table>
Table B.3: OWR Model Regressions. This table reports real and simulated regressions of the $CPIE_{OWR}$ on the squared and interaction terms of $y_e$, $r_d$, and $r_o$. In Panel A, we simulate 1,000 instances of the OWR model for each PERMNO-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the $R^2$ values. In Panel B, we report standardized estimates for the median stock using real data, along with the median $R^2$ values, and tests of the null that the model fits the data. The $p$-value of $R^2$ is the probability of observing an $R^2$ at least as small as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the null at the 5% level.

(a) Simulated Data

<table>
<thead>
<tr>
<th>Year</th>
<th>$\hat{\beta}$</th>
<th>$t$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>$t$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$y_e$^2</td>
<td>$y_e \times r_d$</td>
<td>$y_e \times r_o$</td>
</tr>
<tr>
<td>1993</td>
<td>0.002</td>
<td>0.068</td>
<td>-0.003</td>
</tr>
<tr>
<td>1994</td>
<td>0.002</td>
<td>0.065</td>
<td>-0.003</td>
</tr>
<tr>
<td>1995</td>
<td>0.003</td>
<td>0.065</td>
<td>-0.003</td>
</tr>
<tr>
<td>1996</td>
<td>0.003</td>
<td>0.066</td>
<td>-0.003</td>
</tr>
<tr>
<td>1997</td>
<td>0.003</td>
<td>0.063</td>
<td>-0.003</td>
</tr>
<tr>
<td>1998</td>
<td>0.002</td>
<td>0.070</td>
<td>-0.004</td>
</tr>
<tr>
<td>1999</td>
<td>0.003</td>
<td>0.060</td>
<td>-0.003</td>
</tr>
<tr>
<td>2000</td>
<td>0.003</td>
<td>0.051</td>
<td>-0.002</td>
</tr>
<tr>
<td>2001</td>
<td>0.002</td>
<td>0.066</td>
<td>-0.004</td>
</tr>
<tr>
<td>2002</td>
<td>0.001</td>
<td>0.066</td>
<td>-0.003</td>
</tr>
<tr>
<td>2003</td>
<td>0.002</td>
<td>0.071</td>
<td>-0.005</td>
</tr>
<tr>
<td>2004</td>
<td>0.001</td>
<td>0.068</td>
<td>-0.005</td>
</tr>
<tr>
<td>2005</td>
<td>0.002</td>
<td>0.061</td>
<td>-0.005</td>
</tr>
<tr>
<td>2006</td>
<td>0.001</td>
<td>0.063</td>
<td>-0.004</td>
</tr>
<tr>
<td>2007</td>
<td>0.001</td>
<td>0.051</td>
<td>-0.003</td>
</tr>
<tr>
<td>2008</td>
<td>0.076</td>
<td>0.000</td>
<td>-0.001</td>
</tr>
<tr>
<td>2009</td>
<td>0.002</td>
<td>0.039</td>
<td>-0.002</td>
</tr>
<tr>
<td>2010</td>
<td>0.002</td>
<td>0.038</td>
<td>-0.002</td>
</tr>
<tr>
<td>2011</td>
<td>0.001</td>
<td>0.042</td>
<td>-0.002</td>
</tr>
<tr>
<td>2012</td>
<td>0.001</td>
<td>0.046</td>
<td>-0.003</td>
</tr>
</tbody>
</table>
Table B.3: **OWR Model Regressions.** Continued.

(b) Real Data

<table>
<thead>
<tr>
<th>Year</th>
<th>$\beta_{y_e}$</th>
<th>$\beta_{y_e \times r_d}$</th>
<th>$\beta_{y_e \times r_o}$</th>
<th>$\beta_{r_d}$</th>
<th>$\beta_{r_d \times r_o}$</th>
<th>$\beta_{r_o}^2$</th>
<th>$t_{y_e}$</th>
<th>$t_{y_e \times r_d}$</th>
<th>$t_{y_e \times r_o}$</th>
<th>$t_{r_d}$</th>
<th>$t_{r_d \times r_o}$</th>
<th>$t_{r_o}^2$</th>
<th>$R^2$</th>
<th>p-value</th>
<th>% Rej.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>-0.000</td>
<td>0.053</td>
<td>-0.000</td>
<td>0.032</td>
<td>0.029</td>
<td>0.055</td>
<td>(-0.03)</td>
<td>(7.24)</td>
<td>(-0.13)</td>
<td>(4.41)</td>
<td>(4.56)</td>
<td>(8.11)</td>
<td>69.97%</td>
<td>15.59%</td>
<td>40.02%</td>
</tr>
<tr>
<td>1994</td>
<td>0.000</td>
<td>0.053</td>
<td>-0.001</td>
<td>0.032</td>
<td>0.027</td>
<td>0.060</td>
<td>(0.06)</td>
<td>(8.11)</td>
<td>(-0.17)</td>
<td>(4.69)</td>
<td>(4.68)</td>
<td>(9.44)</td>
<td>72.00%</td>
<td>17.13%</td>
<td>38.49%</td>
</tr>
<tr>
<td>1995</td>
<td>0.001</td>
<td>0.052</td>
<td>-0.001</td>
<td>0.033</td>
<td>0.029</td>
<td>0.059</td>
<td>(0.15)</td>
<td>(7.92)</td>
<td>(-0.17)</td>
<td>(4.74)</td>
<td>(4.89)</td>
<td>(9.35)</td>
<td>72.73%</td>
<td>17.59%</td>
<td>41.21%</td>
</tr>
<tr>
<td>1996</td>
<td>0.001</td>
<td>0.055</td>
<td>-0.003</td>
<td>0.032</td>
<td>0.028</td>
<td>0.062</td>
<td>(0.28)</td>
<td>(8.61)</td>
<td>(-0.52)</td>
<td>(4.77)</td>
<td>(4.81)</td>
<td>(9.83)</td>
<td>73.65%</td>
<td>16.42%</td>
<td>41.69%</td>
</tr>
<tr>
<td>1997</td>
<td>0.002</td>
<td>0.054</td>
<td>-0.002</td>
<td>0.029</td>
<td>0.027</td>
<td>0.061</td>
<td>(0.36)</td>
<td>(8.90)</td>
<td>(-0.53)</td>
<td>(4.85)</td>
<td>(4.84)</td>
<td>(10.17)</td>
<td>74.72%</td>
<td>16.34%</td>
<td>41.04%</td>
</tr>
<tr>
<td>1998</td>
<td>0.002</td>
<td>0.069</td>
<td>-0.004</td>
<td>0.025</td>
<td>0.023</td>
<td>0.074</td>
<td>(0.37)</td>
<td>(11.25)</td>
<td>(-0.89)</td>
<td>(4.43)</td>
<td>(4.15)</td>
<td>(12.61)</td>
<td>77.46%</td>
<td>16.14%</td>
<td>35.82%</td>
</tr>
<tr>
<td>1999</td>
<td>0.002</td>
<td>0.057</td>
<td>-0.003</td>
<td>0.025</td>
<td>0.025</td>
<td>0.065</td>
<td>(0.56)</td>
<td>(9.59)</td>
<td>(-0.64)</td>
<td>(4.33)</td>
<td>(4.58)</td>
<td>(11.66)</td>
<td>76.48%</td>
<td>19.65%</td>
<td>31.28%</td>
</tr>
<tr>
<td>2000</td>
<td>0.003</td>
<td>0.050</td>
<td>-0.003</td>
<td>0.021</td>
<td>0.022</td>
<td>0.066</td>
<td>(0.82)</td>
<td>(10.58)</td>
<td>(-0.98)</td>
<td>(4.50)</td>
<td>(5.15)</td>
<td>(14.37)</td>
<td>79.83%</td>
<td>28.77%</td>
<td>20.29%</td>
</tr>
<tr>
<td>2001</td>
<td>0.001</td>
<td>0.068</td>
<td>-0.003</td>
<td>0.018</td>
<td>0.016</td>
<td>0.078</td>
<td>(0.47)</td>
<td>(14.62)</td>
<td>(-0.94)</td>
<td>(4.10)</td>
<td>(3.81)</td>
<td>(16.91)</td>
<td>83.25%</td>
<td>34.94%</td>
<td>20.48%</td>
</tr>
<tr>
<td>2002</td>
<td>0.002</td>
<td>0.072</td>
<td>-0.002</td>
<td>0.016</td>
<td>0.014</td>
<td>0.081</td>
<td>(0.47)</td>
<td>(16.83)</td>
<td>(-0.72)</td>
<td>(3.88)</td>
<td>(3.71)</td>
<td>(19.17)</td>
<td>84.71%</td>
<td>36.14%</td>
<td>16.94%</td>
</tr>
<tr>
<td>2003</td>
<td>0.002</td>
<td>0.080</td>
<td>-0.003</td>
<td>0.017</td>
<td>0.015</td>
<td>0.080</td>
<td>(0.60)</td>
<td>(20.66)</td>
<td>(-0.94)</td>
<td>(4.38)</td>
<td>(3.93)</td>
<td>(20.51)</td>
<td>87.22%</td>
<td>42.18%</td>
<td>13.72%</td>
</tr>
<tr>
<td>2004</td>
<td>0.001</td>
<td>0.077</td>
<td>-0.005</td>
<td>0.016</td>
<td>0.012</td>
<td>0.074</td>
<td>(0.54)</td>
<td>(24.74)</td>
<td>(-1.74)</td>
<td>(4.48)</td>
<td>(3.58)</td>
<td>(21.11)</td>
<td>88.70%</td>
<td>41.66%</td>
<td>14.95%</td>
</tr>
<tr>
<td>2005</td>
<td>0.002</td>
<td>0.072</td>
<td>-0.005</td>
<td>0.013</td>
<td>0.010</td>
<td>0.065</td>
<td>(0.83)</td>
<td>(25.08)</td>
<td>(-2.12)</td>
<td>(4.36)</td>
<td>(3.32)</td>
<td>(20.58)</td>
<td>89.54%</td>
<td>44.67%</td>
<td>12.29%</td>
</tr>
<tr>
<td>2006</td>
<td>0.002</td>
<td>0.072</td>
<td>-0.005</td>
<td>0.013</td>
<td>0.010</td>
<td>0.066</td>
<td>(0.74)</td>
<td>(25.53)</td>
<td>(-1.61)</td>
<td>(4.12)</td>
<td>(3.36)</td>
<td>(20.42)</td>
<td>89.47%</td>
<td>43.32%</td>
<td>12.28%</td>
</tr>
<tr>
<td>2007</td>
<td>0.002</td>
<td>0.058</td>
<td>-0.003</td>
<td>0.004</td>
<td>0.005</td>
<td>0.058</td>
<td>(0.98)</td>
<td>(18.17)</td>
<td>(-0.97)</td>
<td>(1.40)</td>
<td>(1.79)</td>
<td>(17.59)</td>
<td>89.34%</td>
<td>45.41%</td>
<td>10.21%</td>
</tr>
<tr>
<td>2008</td>
<td>0.077</td>
<td>0.004</td>
<td>-0.002</td>
<td>0.003</td>
<td>0.006</td>
<td>0.007</td>
<td>(22.41)</td>
<td>(1.10)</td>
<td>(-0.55)</td>
<td>(1.07)</td>
<td>(2.00)</td>
<td>(1.54)</td>
<td>88.02%</td>
<td>42.49%</td>
<td>7.83%</td>
</tr>
<tr>
<td>2009</td>
<td>0.003</td>
<td>0.038</td>
<td>-0.002</td>
<td>0.004</td>
<td>0.006</td>
<td>0.053</td>
<td>(1.55)</td>
<td>(15.99)</td>
<td>(-0.87)</td>
<td>(1.85)</td>
<td>(2.42)</td>
<td>(22.33)</td>
<td>89.34%</td>
<td>46.12%</td>
<td>7.45%</td>
</tr>
<tr>
<td>2010</td>
<td>0.002</td>
<td>0.035</td>
<td>-0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.038</td>
<td>(1.39)</td>
<td>(16.80)</td>
<td>(-0.69)</td>
<td>(1.02)</td>
<td>(1.53)</td>
<td>(15.83)</td>
<td>89.54%</td>
<td>48.77%</td>
<td>7.53%</td>
</tr>
<tr>
<td>2011</td>
<td>0.002</td>
<td>0.043</td>
<td>-0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.050</td>
<td>(1.27)</td>
<td>(17.71)</td>
<td>(-0.84)</td>
<td>(1.04)</td>
<td>(1.50)</td>
<td>(18.56)</td>
<td>89.84%</td>
<td>49.68%</td>
<td>7.88%</td>
</tr>
<tr>
<td>2012</td>
<td>0.002</td>
<td>0.045</td>
<td>-0.003</td>
<td>0.002</td>
<td>0.003</td>
<td>0.039</td>
<td>(1.14)</td>
<td>(20.34)</td>
<td>(-1.05)</td>
<td>(1.20)</td>
<td>(1.54)</td>
<td>(17.30)</td>
<td>90.29%</td>
<td>51.35%</td>
<td>7.71%</td>
</tr>
</tbody>
</table>
Figure B.1: **DY Tree.** For a given trading day, private information arrives with probability $\alpha$. When there is no private information, buys and sells are Poisson with intensity $\epsilon_B$ and $\epsilon_S$. Private information is good news with probability $\delta$. The expected number of buys (sells) increases by $\mu$ in case of good (bad) news. Non-information related order flow shocks arrive with probability $\theta$. In the event of an order flow shock, buys and sells increase by $\delta_b$ and $\delta_s$ respectively.
Figure B.2: **XOM DY.** This figure compares the real and simulated data for XOM in 1993 using the DY model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{DY}$, with darker markers (+) representing high and lighter markers (+) low $CPIEs$. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark “cloud” of points, and low probability states appear as a light “cloud” of points. The DY model extends the three states of the PIN model corresponding to no news, good news, and bad news with three additional states with higher order flows due to non-information symmetric order flow shocks.

(a) XOM 1993          (b) XOM 2012

Figure B.3: **Breakdown of the DY Model.** This figure shows the distribution of the percent of days where the total likelihood, given the model parameters and observed order flow data is less than $10^{-10}$—days, according to the model, with near-zero probability of occurring. The solid black line represents the median stock, and the dotted lines represent the 5, 25, 75, and 95 percentiles.
Figure B.4: M&A Targets - PIN. Panel A shows the average $CPIE_{PIN}$ for the PIN model in event time surrounding M&A announcements in target stocks. Panels B and C compare the average $CPIE_{PIN}$ with the $CPIE_{PIN}$ predicted with either the absolute order imbalance or turnover, respectively. To obtain the predictions, we run regressions of daily $CPIE_{PIN}$ on $|B - S|$ or $\text{turn}$, and their respective squared terms.

(a) $CPIE_{PIN}$

(b) Prediction using $|B - S|$ and $|B - S|^2$

(c) Prediction using $\text{turn}$ and $\text{turn}^2$
Bibliography


Akins, Brian K., Jeffrey Ng, and Rodrigo S. Verdi, 2012, Investor competition over information and the pricing of information asymmetry, *The Accounting Review* 87, 35–58.


Back, Kerry, Kevin Crotty, and Tao Li, 2014, Can information asymmetry be identified from order flows alone?, *Working paper*.


Bamber, Linda S., Orire E. Barron, and Douglas E. Stevens, 2011, Trading volume around earnings announcements and other financial reports: Theory, research


Bonacich, Phillip, 1972a, Factoring and weighting approaches to status scores and clique identification, *Journal of Mathematical Sociology* 2, 113–120.


Chandrasekhar, A, 2015, Econometrics of network formation, .


Gan, Quan, Wang C. Wei, and David J. Johnstone, 2014, Does the probability of informed trading model fit empirical data?, *FIRN Research Paper*.


Heckman, James J, 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, in *Annals of Economic and Social Measurement* vol. 5 pp. 475–492.


Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies and com-


