

Large-scale insider trading analysis: patterns and discoveries

Acar Tamersoy · Elias Khalil · Bo Xie ·
 Stephen L. Lenkey · Bryan R. Routledge ·
 Duen Horng Chau · Shamkant B. Navathe

Received: 8 January 2014 / Revised: 30 April 2014 / Accepted: 12 May 2014
 © Springer-Verlag Wien 2014

Abstract How do company insiders trade? Do their trading behaviors differ based on their roles (e.g., chief executive officer vs. chief financial officer)? Do those behaviors change over time (e.g., impacted by the 2008 market crash)? Can we identify insiders who have similar trading behaviors? And what does that tell us? This work presents the first academic, large-scale exploratory study of insider filings and related data, based on the complete Form 4 filings from the U.S. Securities and Exchange Commission. We analyze 12 million transactions by 370 thousand insiders spanning 1986–2012, the largest reported in academia. We explore the temporal and network-based aspects of the trading behaviors of insiders, and make surprising and counterintuitive discoveries. We study how the trading behaviors of insiders differ based on their roles in their companies, the types of their transactions, their

companies' sectors, and their relationships with other insiders. Our work raises exciting research questions and opens up many opportunities for future studies. Most importantly, we believe our work could form the basis of novel tools for financial regulators and policymakers to detect illegal insider trading, help them understand the dynamics of the trades, and enable them to adapt their detection strategies toward these dynamics.

1 Introduction

Illegal insider trading—defined by statutes, regulations, and common law—means exploiting one's role in an organization to gain information to profitably trade in financial markets. Public policy debates related to insider trading usually weigh the harm to financial markets through reduced liquidity (“adverse selection”) and undesirable effects on managerial incentives (“moral hazard”) against the economic benefit from any information that is indirectly revealed via the trading process (see Bernhardt et al. 1995). As many recent high profile cases highlight, illegal insider trading is actively prosecuted.

This is an extended version of the IEEE/ACM ASONAM 2013 paper “Inside Insider Trading: Patterns & Discoveries from a Large Scale Exploratory Analysis” Tamersoy et al. (2013). The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statement by any of its employees. The views expressed herein are those of the author and do not necessarily reflect the views of the Commission or of the author's colleagues on the staff of the Commission.

A. Tamersoy (✉) · E. Khalil · B. Xie · D. H. Chau ·
 S. B. Navathe
 College of Computing, Georgia Institute of Technology, Atlanta,
 GA, USA
 e-mail: tamersoy@gatech.edu

E. Khalil
 e-mail: lyes@gatech.edu

B. Xie
 e-mail: bo.xie@gatech.edu

D. H. Chau
 e-mail: polo@gatech.edu

S. B. Navathe
 e-mail: sham@cc.gatech.edu

S. L. Lenkey
 Division of Economic and Risk Analysis, U.S. Securities and
 Exchange Commission, Washington, DC, USA
 e-mail: lenkeys@sec.gov

B. R. Routledge
 Tepper School of Business, Carnegie Mellon University,
 Pittsburgh, PA, USA
 e-mail: routledge@cmu.edu

Most trades by insiders, however, are not illegal. Insiders are defined as corporate officers, directors, or beneficial owners of more than 10 % of a company's stock. Illegal insider trading involves using *material nonpublic* information about the company as a basis for trade. Most often, insiders trade simply to adjust their portfolio to alter the risk profile (diversify) or liquidity (cash-out). To monitor trades by insiders, the U.S. Securities and Exchange Commission (SEC) requires these trades to be disclosed via a form called *Form 4*. Detecting illegal trades in the large pool of reported trades is challenging.

1.1 Opportunities for data mining

Government regulators are increasingly interested in applying data mining techniques to detect fraud and illegal insider trading (Hamilton 2011). These techniques can provide a way to quickly sift through large volumes of transactions to spot illegal trades.

Our work aims to help regulators and policymakers better understand how insiders trade based on factors such as corporate roles, company sectors, and how insiders' connections with each other affect their trades. This knowledge could eventually help detect potential illegal activities at a large scale. We utilize techniques from time series data mining and social network mining. First, tools that explore the time series of insiders' trades are important because, as we show, insiders' trading behaviors are affected by corporate and government regulations, and major economic events in the past decades. By understanding the temporal patterns of insiders' trading behaviors, we could flag the ones that exhibit anomalous activities for further examination. Second, network-based analysis is crucial for detecting illegal insider trading since insiders often share information through their social networks. With network-based techniques, we could uncover the hidden communication channels through which the inside information flows and better understand how insiders operate collectively.

To the best of our knowledge, very few published research is available that uses computational techniques to help financial regulators and policymakers streamline or automate the analysis process of insiders' trades. Our work explores a large dataset of the SEC Form 4 filings, which describe changes in the ownership interests of insiders in their firms. As such, we present the first effort to systematically analyze insider trades in a large-scale setting.

1.2 Benefits for regulators

Our analysis may benefit financial regulators and policymakers in a number of ways. Our analysis could provide a useful and novel tool for detecting illegal insider trading.

Our methodology uncovers individuals' trading patterns and compares their transactions in a nonparametric way. As such, our results could form a basis to initiate an examination of a particular set of insiders' transactions that seem suspicious. We envision use by financial regulators and policymakers as the most likely avenue for deploying our research. Our analysis has the potential to spur future research by economists and legal scholars as well.

1.3 Contributions

We conduct an extensive large-scale analysis of insiders' trades using the Form 4 filings. Our analysis consists of three major components. The first is based on time series data mining; in this component, we discover temporal patterns by partitioning the trades on several properties, such as corporate roles, company sectors, and transaction types. The second is the correlational analysis of prices of insiders' transactions and market closing prices of their companies' stocks, where we develop a statistical approach to determine the insiders who are skilled at timing their transactions. The third is based on social network mining; in this component, we construct networks of insiders based on the similarity of insiders' timings of their transactions. Our main contributions include the following:

- We perform the first academic, large-scale exploratory study of the insider SEC Form 4 filings;
- We discover distinctive temporal patterns in insiders' trades that may be explained by government regulations, corporate policies, employment positions, company sectors, and macroeconomic factors;
- We determine that a significant portion of the insiders makes short-swing profits despite the existence of a rule designed to prevent short-swing trading;
- We discover a set of insiders who time their trades well: They buy when the price is low or sell when the price is high in comparison with the market closing price;
- We find strong evidence that insiders form small clusters in which trade-related information might propagate both **vertically** (between higher- and lower-level insiders) and **horizontally** (among lower-level insiders).

Our work takes a computational and statistical modeling approach toward the challenging problem of uncovering correlations among insiders. As we show, our approach discovers a number of interesting and rare findings that may otherwise be buried among the large amount of insider data. We note, however, that our conclusions are based only on publicly available data. In addition, the relationships we uncover are statistical in nature and do not necessarily imply that any particular insider has traded

illegally. We hence replace the names of insiders and companies with generic symbols (e.g., company A) throughout the paper.

Next, we describe our data, survey related work, present our methods and results, and discuss their implications. Finally, we close with a summary.

2 Dataset

United States federal law requires corporate insiders to report their open-market transactions and other ownership changes to the SEC within 2 business days via Form 4. This form consists of two parts, namely Part 1 and Part 2. Part 1 is used for transactions related to stocks and non-derivatives, whereas Part 2 is used to report transactions about derivatives, such as options, warrants, and convertible securities. In this work, we focus on analyzing Part 1 of each Form 4 filed with the SEC.

The forms we analyze range from January 1986 to August 2012, including more than 12 million transactions in more than 15 thousand companies, mostly located in the USA. Table 1 provides a set of summary statistics for the dataset. Each record in the dataset consists of information about a transaction by an insider. The fields in a record include the name and company of the insider, transaction date and type, number of shares traded, transaction price, role of the insider in the company, and information about the company, including its sector and address. There are over 50 different role codes an insider may report in a Form 4, ranging from chairman of the board to retired. Since a role code's job nature is loosely defined, occasionally insiders may report different but related role codes in subsequent trades. This is a minor issue when we consider high-level aggregate data, such as all transactions by presidents since 1986. However, when we focus on a particular insider, it becomes difficult to associate that trader with a role in the company. Previous work has proposed heuristics to map specific role codes to more general ones. Our low-level insider-specific analyses (i.e., analyses other than those in Sects. 4.1 and 4.2) use the mapping from Edelson and Whisenant (2009), which converts a role code from the raw data into one of the four *general* codes: chief executive officer (CEO), chief financial officer (CFO),

director (D), or other officer (OO). In some analyses, we also consider beneficial owners, which we represent with the role code *B*. This mapping is effective in that it assigns one general role code to most of the insiders in the time periods we consider. If an insider receives more than one general role code, we ignore that insider in the analysis. We store the dataset in a SQLite database for ease of analysis. The database contains both parts of the filings and has a size of 5.61 GB. The forms we analyze are publicly available through the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (The U.S. Security and Exchange Commission).

Figures 1 and 2 show the empirical cumulative distribution functions for the number of companies that insiders belong to and the number of transactions that insiders have, respectively. We observe that most insiders belong to a small number of companies and have a small number of transactions; however, there are a handful of insiders on the extremes, which are involved in many companies or actively trading their companies' stock. Figure 3 shows the geographical distribution of the transactions based on the zip codes of the corporate headquarters. The highest number of transactions occurs for companies headquartered in the state of California, followed by New York and Texas.

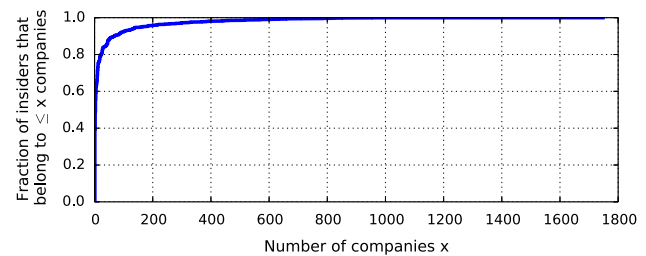


Fig. 1 Empirical cumulative distribution function for the number of companies that insiders belong to in our dataset. A majority of insiders belong to a small number of companies. (All figures best viewed in color)

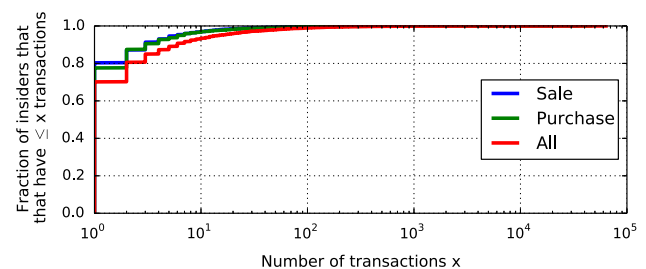


Fig. 2 Empirical cumulative distribution function for the number of transactions that insiders have in our dataset. Note that the *x*-axis is in log scale. A majority of insiders have a small number of transactions

Table 1 Summary statistics for our dataset

Insiders	370,627
Companies	15,598
Transactions	12,360,325
Sale transactions	3,206,175
Purchase transactions	1,206,038

We focus on open-market sale and purchase transactions

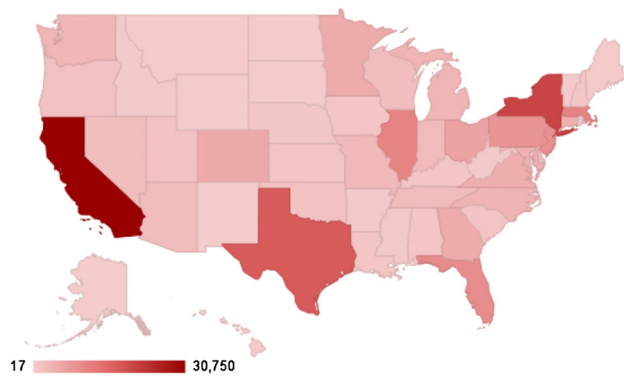


Fig. 3 Geographical distribution of the number of transactions based on the zip codes of the insiders' companies. *Darker color* indicates higher number. The highest number of transactions initiates from the state of California

3 Related work

This work intersects several research areas. We group the related work into different categories and overview previous work closely related to ours from each category. To the best of our knowledge, our work is the first academic study that extensively analyzes the SEC Form 4 data at scale.

3.1 Profiling insiders

In the finance domain, Cohen et al. (2012) characterize insiders into routine traders and opportunist traders. The authors show that the routine trades do not carry information in predicting future company events or achieving higher abnormal returns. In contrast, the irregular “opportunistic” activities carry significant information in the sense that strategies following such trades have a high abnormal return. Compared to their work, we explore insiders' trading behaviors from a network perspective.

Several studies find evidence that actively trading executives not only benefit from their insider knowledge, but also manipulate firm-related information by voluntary disclosures and then trade on that information. Cheng and Lo (2006) show that managers who intend to buy shares for their own accounts also tend to release abnormally negative news in the period just before their insider purchases to drive the prices down. Similarly, Brockman et al. (2010) find that managers release abnormally positive news before stock option exercises to obtain relatively high sales prices, and Aboody and Kasznik (2000) show that managers tend to release bad news before stock option grants to fix lower strike prices. Brockman et al. (2013) examine the relationship between the tone of conference calls presented by company executives and their subsequent insider trading behavior. The authors find that positive conference call

presentation tones predict net insider selling, whereas negative conference call tones predict net insider buying, and this discrepancy is stronger for CEOs than non-CEO executives. Our work is different than this line of research as we do not attempt to associate insider trades with events such as public news and conference calls.

Lorie and Niederhoffer (1968) explore several statistical properties of insider traders based on SEC filings. They find that insiders tend to buy more often before the stock prices increase and to sell more often before the prices decrease. The authors also determine that consecutive trades of the same type (purchase-then-purchase and sale-then-sale) are more likely than trades of opposite types. Lakonishok and Lee (2001) examine the information content of insiders' trades and the market's response to those trades. The authors draw an interesting conclusion that insiders tend to buy stocks with poor past performance, but sell those that performed well in the past. Furthermore, they demonstrate that the market underreacts to the signals from insiders' trades despite their high returns. In comparison with these works, we explore a significantly larger dataset both in terms of the number of companies and time span.

3.2 Detecting potential fraud and illegal trades

Goldberg et al. (2003) describe the Securities Observation, News, Analysis and Regulation (SONAR) system, which flags unusual price and volume movements in traded securities and identifies potential insider trading and fraud against investors. Compared to our approach, SONAR uses the SEC filings only for fraud detection, and it is not clear which particular filings are utilized by the system. Donoho (2004) focuses on options trading and adapts several data mining algorithms for the early detection of insider trading. The author concludes that volatility implied by the price is the best predictor of future news. Compared to this approach, we consider a larger dataset and focus on the more challenging stocks trading. Kirkos et al. (2007) evaluate the effectiveness of classification techniques, such as decision trees, neural networks, and bayesian networks, in discriminating firms that issue fraudulent financial statements, based on features extracted from the statements, such as debt information and inventory reports. Compared to this approach, our network-based analysis is insider-centric as opposed to firm-centric, and we do not question the credibility of the SEC filings. In Summers and Sweeney (1998), Summers et al. investigate the relationship between firms issuing fraudulent financial statements and the behavior of insiders of those firms. The authors find that insiders of fraudulent firms tend to sell their stocks to reduce their holdings, which is an indication of their knowledge of the fraud that is taking place. The work uses

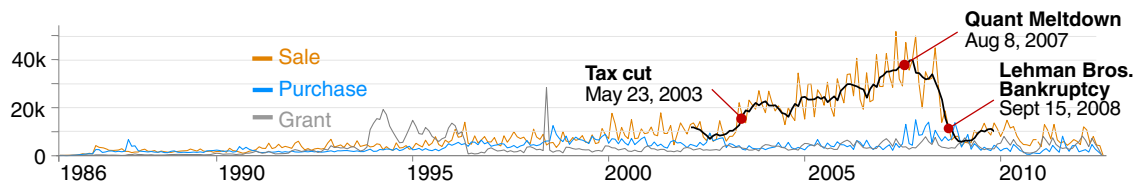


Fig. 4 The daily count of *Purchase*, *Sale*, and *Grant* transactions (the most common types) over 1986–2012, and 180-day centered moving average for *Sale* transactions shown in *black*. The change in the U.S.

tax law in 2003 (reduced capital gains taxes) boosted *Sale* transactions for following years. Financial crises like the “Quant Meltdown” in 2007 and the burst of “housing bubble” in 2008 suppressed them

SEC filings of around 50 firms mentioned in news reports as part of a fraud case. Compared to this work, we are interested in a larger span of SEC filings, and we do not seek to correlate public news with insider trades.

Other works that use data mining techniques for fraud detection include SNARE (McGlohon et al. 2009), which uses a network-based approach that adapts belief propagation (BP) to pinpoint misstated accounts in a sample of general ledger data. This work was inspired by the earlier NetProbe system that uses BP to detect collusion in online auctions (Pandit et al. 2007). A more general system, Sherlock (Bay et al. 2006) uses a suite of classic classification methods (naive Bayes, logistic regression, etc.) to identify suspicious accounts. The techniques we present in this work could form a basis for detecting suspicious and potentially illegal trades.

3.3 Mining financial data

Fan et al. (2004) present a data mining based automatic trading surveillance system for large data with skewed distribution using multiple classifiers. Bizjak et al. (2009) document the network structure in the interlocking board of directors to explain how inappropriately backdating compensation spreads. Adamic et al. (2010) construct and analyze a series of trading networks from transaction-level data and determine that properties of trading networks are strongly correlated with transaction prices, trading volume, inter-trade duration, and measures of market liquidity. The work uses audit trail, transaction-level data of E-mini S&P 500 futures contract from September 2009. Compared to the works above, we analyze a larger number of factors on a larger dataset spanning 26 years and focus on understanding the trading behaviors of insiders.

To the best of our knowledge, our work is the first in academia that extensively studies the Form 4 data at a large scale from a data mining perspective.

4 Patterns, observations and analysis

We hypothesize that two important factors reveal information from insiders’ transactions. The first factor is the

timings of transactions. If insiders place their transactions around major corporate events, it is likely that the transactions are based on information. Otherwise, if they trade routinely on the same month every year, it is more likely that the trades are for liquidity or diversification reasons (Cohen et al. 2012). The second factor is the connections between insiders. If a network of insiders consistently trade similarly, they are likely to share information with each other. Based on these assumptions, we present our analyses to extract temporal and network-based patterns from insiders’ transactions.

4.1 Time series in different facets

We first analyze trends in the time series of transactions. Since many factors contribute to the timings of transactions, we break down the data based on transaction types, role codes, and sectors of companies to examine the effect of each factor.

Analyzing transaction types reveals interesting patterns as shown in Fig. 4. In general, the number of sales is greater than that of purchases. This is especially significant during the period 2003–2008. Many insiders receive shares of stock as a part of their compensation via, for example, stock options. Only a small fraction of the shares are obtained through open-market purchases. Hence, sales are common as insiders rebalance their portfolios for better diversification and liquidate shares for consumption. Note that the increase in the frequency of sale transactions coincides with the 2003 change in the United States tax law¹ that reduced capital gains taxes. The sharp drop in sales occurs after the “Quant Meltdown” of August 2007² (Khandani and Lo 2011) but, interestingly, prior to the largest fall in market prices in late September and October 2008. The reduction in sales after the market drop is consistent with the behavioral (although not entirely rational) explanation that investors are less likely to sell at a loss (see Odean 1998). An alternative explanation for the drop in sales is that executive stock options, which are often

¹ Enacted May 23, 2003.

² A point identified, with hindsight, as the start of the financial crisis.

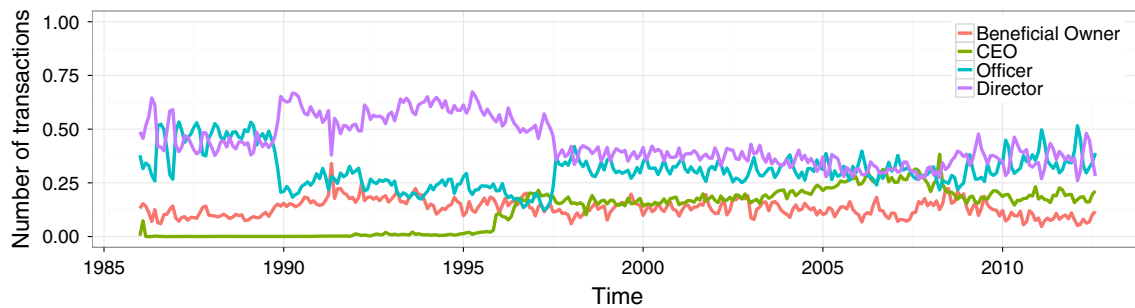


Fig. 5 Transactions break down by role codes. Only the most frequent four codes are shown. Beneficial owners behave differently than the other insiders

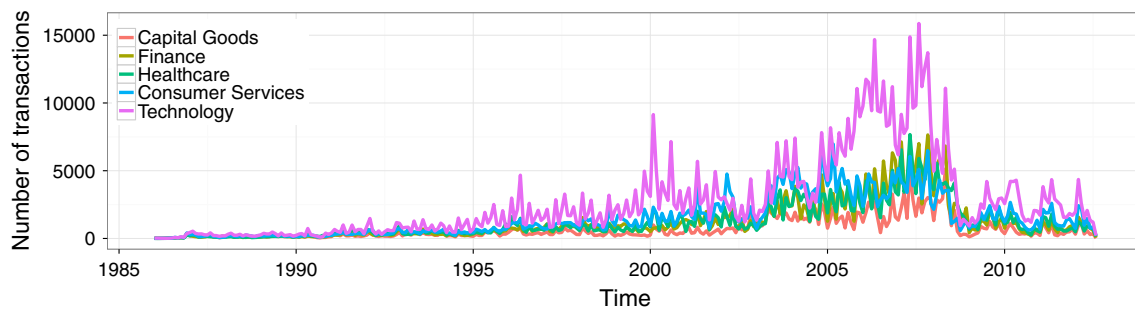


Fig. 6 Transactions break down by sectors. Only the most frequent five sectors are shown. Most activity comes from the technology sector

granted at-the-money, became worthless by the time they vested after 2008 and were never exercised.

Figure 5 illustrates that insiders with different roles have different trading patterns. Most transactions are made by directors and officers, mostly for the reason that they make up a large proportion of the insiders. The behaviors of CEOs are more volatile; they start selling aggressively after 2003 and stop doing so in late 2007. In contrast, the selling activity of beneficial owners increases only toward the eve of the financial crisis, and shortly after the crisis, their activity level decreases even though the transaction counts of other insiders fluctuate during the same period. The differences in the trading patterns could be due to the fact that beneficial owners do not have access to the same information as other insiders.

Figure 6 depicts trading activity in various sectors. In terms of the number of transactions, technology is the largest sector. Both the dot-com bubble and the subprime mortgage crisis appear in the plot as an increase around 2000 and a sharp drop around 2008, respectively. Another interesting observation is that the trend of the technology sector matches well with the sales trend in Fig. 4. Inspired by Ruiz et al. (2012), we compute the cross-correlation coefficient (CCF) between these two time series, with a lag parameter of 0 days. The resulting CCF value of 0.95 indicates that the trends are indeed similar ($p < 0.01$). This is likely due to technology companies compensating their employees with equity.

4.2 Analyzing transaction intervals

We next look at the patterns within the sequences of transactions. What fraction of insiders sell after a purchase and what fraction keep selling or purchasing? To answer these questions, we analyze the transaction intervals between consecutive trades.

Figures 7 and 8 depict the number of open-market sale and purchase transactions versus the interval in days between any two consecutive transactions, for all four combinations of the transaction types. If the insider has a sale transaction that is followed by a purchase transaction, we call this transaction pair a *sale-then-purchase* pair and denote it with the notation $S \rightarrow P$. The other three transaction pairs are *purchase-then-sale* ($P \rightarrow S$), *sale-then-sale* ($S \rightarrow S$), and *purchase-then-purchase* ($P \rightarrow P$). From Figs. 7 and 8, we see that, in general, $S \rightarrow P$ and $P \rightarrow S$ pairs are less common than $P \rightarrow P$ and $S \rightarrow S$ pairs. This could be due to a couple of factors. First, many insiders are employees who are compensated with equity grants. These insiders may choose to engage in periodic sales in order to liquidate or diversify their assets, which helps to explain the prevalence of the $S \rightarrow S$ pairs. Second, insiders may use 10b5-1 plans to accumulate shares by making periodic purchases, which helps to explain the prevalence of the $P \rightarrow P$ pairs. Another notable observation in Fig. 7 is that the pattern is strongly oscillatory, with a cycle of about 90 days.

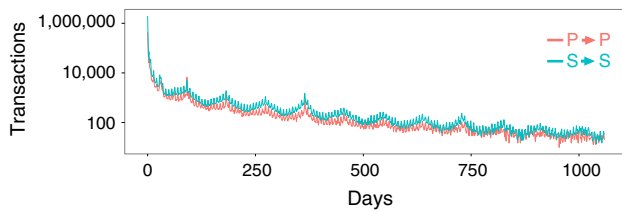


Fig. 7 Time between consecutive transactions of the same type: purchase-then-purchase (P→P) and sale-then-sale (S→S). The pattern is oscillatory, with a cycle of about 90 days

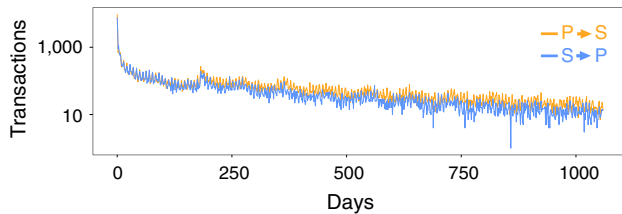


Fig. 8 Time between consecutive transactions of different types: purchase-then-sale (P→S) and sale-then-purchase (S→P). The highest peak for both distributions is around the point corresponding to 180 days

This could be due to corporate bylaws that prohibit transactions near quarterly earnings announcements.

The highest peak for both P→S and S→P distributions in Fig. 8 is around the point corresponding to 180 days. This appears to be a result of the short-swing profit rule, which is codified in Section 16(b) of the Securities Exchange Act of 1934.³ Essentially, the statute prevents insiders from realizing any trading profit resulting from a combined purchase and sale, or sale and purchase, of the firm's stock within a 6-month period. As a result of the rule, one might expect that round-trip transactions completed within a 6-month interval are rarely profitable.

To test this hypothesis, we consider each company C in the dataset and compute the profit earned from each of the S→P and P→S pairs of the company's insiders using the formula below. Assuming that the transactions in the pair occurred on dates t_k and t_ℓ ($t_k \leq t_\ell$), the profit earned is

³ The relevant portion of Section 16(b) reads:

For the purpose of preventing the unfair use of information which may have been obtained by [an insider] by reason of his relationship to the issuer, any profit realized by [an insider] from any purchase and sale, or any sale and purchase, of any equity security of such issuer...within any period of <6 months...shall inure to and be recoverable by the issuer, irrespective of any intention on the part of [the insider] in entering into such transaction of holding the security...purchased or of not repurchasing the security...sold for a period exceeding 6 months. Suit to recover such profit may be instituted...by the issuer, or by the owner of any security of the issuer in the name and in behalf of the issuer if the issuer shall fail or refuse to bring such suit within sixty days after request or shall fail diligently to prosecute the same thereafter[.].

$$\left(\log(P_{t_\ell}^C) - \log(P_{t_k}^C)\right) \times P_{t_k}^C \times \min(ST_{t_k}^C, ST_{t_\ell}^C), \quad (1)$$

where $P_{t_i}^C$ is the market closing price of company C 's stock at date t_i and $ST_{t_i}^C$ is the number of company C 's shares traded by the insider at date t_i . The first term in the formula is simply the log-return for the transaction pair. Because insiders may be compelled to disgorge only their realized trading profit, we multiply the log-return by the price of the first transaction and the smaller of the number of shares traded in the two transactions.⁴

Figure 9 shows the fraction of S→P and P→S pairs that are either profitable or unprofitable and which are at most 6 months apart (the rule above applies) or >6 but ≤7 months apart (the rule no longer applies).⁵ Interestingly, approximately 45 % of the pairs containing transactions that occur within 6 months of each other are profitable. In contrast, roughly 70 % of the pairs completed outside of the statutory holding period generate a profit.⁶

Two-tailed t tests with the alternative hypothesis $H_a : \mu_{\text{profit}} \neq 0$ indicate that the profit earned from such round-trip transactions is statistically significant ($p < 0.01$) in both samples. However, a one-tailed Welch's t test indicates that that the profit earned from the pairs completed outside of the statutory holding period is significantly ($p < 0.01$) greater than the profit earned from pairs completed within 6 months. While the data indicates that the short-swing profit rule may not completely deter insiders from making profitable short-swing trades, the rule seems to have an effect on the insiders' trading patterns.

To examine how insiders in different roles trade consecutively, we plot the transaction intervals for various role codes in Fig. 10. An interesting observation is that the beneficial owners as a group behave differently than the other insiders. The oscillatory pattern observed in the transaction intervals for other types of insiders is absent in the transaction intervals of beneficial owners. This might be explained by the fact that many beneficial owners are effectively "outsiders"—that is, they are not directly affiliated with the company and, consequently, may not be

⁴ Under *Smolowe v. Delendo Corp.*, 136 F.2d 231 (1943), when calculating the amount of short-swing profit realized by an insider, transactions should be matched to reach the maximum possible profit. Chin (1997) claims that a transportation algorithm should be used to compute the maximum possible profit when multiple transactions occur within rolling 6-month windows. Due to the sheer number of transactions, we only consider the consecutive transactions for simplicity.

⁵ We take into account the varying number of days in different months to get an accurate value for the number of months between the two transactions in a pair.

⁶ The Pearson's product-moment correlation coefficient value of 0.12 indicates positive correlation between profit and number of shares traded ($p < 0.01$).

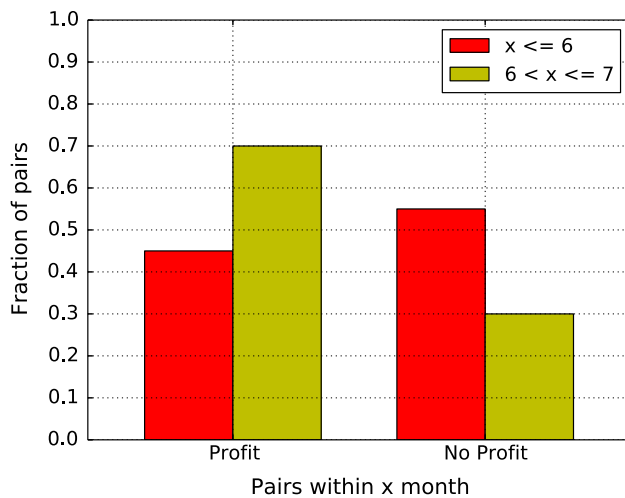


Fig. 9 Fraction of consecutive opposite transaction pairs ($P \rightarrow S$ and $S \rightarrow P$) that are profitable versus unprofitable. Forty-five percentage of the pairs that occur within a 6-month period are profitable despite the short-swing profit rule, which requires insiders to forfeit profit from trades that occur within 6 months of each other

subject to corporate bylaws—though some beneficial owners are other companies rather than individuals. We further observe that the patterns for the other types of insiders differ among themselves. For example, officers have significantly more $S \rightarrow S$ pairs than $P \rightarrow P$ pairs. This, again, is likely related to the stock options and grants given to the officers as part of their compensation packages. Directors are generally fewer in number and typically do not receive as much stock compensation.

Figure 11 illustrates that the companies' sectors also affect how insiders trade. For example, we observe that insiders in the technology sector consecutively sell more than they purchase, while in finance the number of consecutive purchase and sale transactions is more balanced. This may be attributed to how insiders are compensated in different sectors. For instance, the fact that employees in the technology sector are often compensated with stock or options implies that a large portion of their stock holdings are not derived from open-market purchases.

4.3 Correlational analysis of transaction and stock prices

Is it possible to assert that a certain set of insiders are likely to be making *informed* trades? Previous work looked at insiders' transactions before major company-related events, such as takeovers (Agrawal and Nasser 2011) and accounting scandals (Agrawal and Cooper 2008), and attempted to determine whether insiders might be trading in an informed manner by considering certain properties of the transactions, such as type, amount, etc. Instead of focusing on major events, we look at the complete

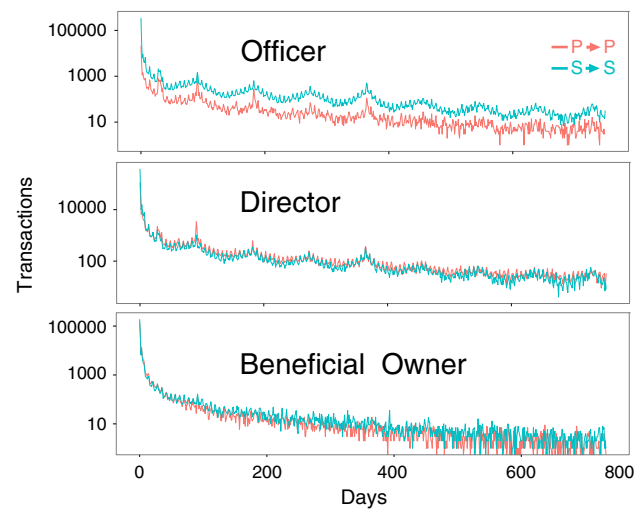


Fig. 10 Transaction intervals for different role codes. Insiders in different roles trade differently

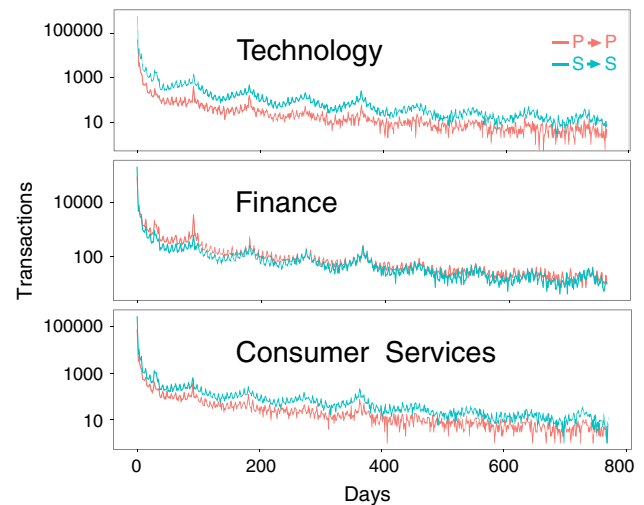


Fig. 11 Transaction intervals for different sectors. Insiders in different sectors trade differently

spectrum of trades with the same goal of unearthing suspicious trading activity.

Specifically, we consider all the open-market sale and purchase transactions of an insider, and for each transaction of the insider for company C , we compare the reported price of the transaction with the market closing price of company C 's stock on the date of the transaction. If an insider makes a purchase at price TP during the day and the market closing price, CP , of company C 's stock is strictly greater than TP ($CP > TP$), or if the insider makes a sale at price TP during the day, and we see that market closing price CP is strictly less than TP ($CP < TP$), then these trades might be information-based because the insider buys when the price is low or sells when the price is high in comparison to the market closing price.

An important question is how should we quantify the level of informedness of a particular transaction and, eventually, of an insider overall? In other words, how do we make sure that it is not only pure luck that is driving these trades? We propose the statistical procedure in Algorithm 1 as one possible approach.

transaction for which the number of shares traded is equal to the outcome of the summation (line 5). Subsequently, we retrieve the market closing price and *dollar volume*⁷ of the company's stock on the date of the transaction (lines 6–7).

Note that our goal here is to aggregate the “signals” from all the transactions of the insider, possibly for dif-

Algorithm 1 Correlational Analysis of Transaction and Stock Prices

Return: Insiders with a significant statistical result

```

1:  $T \leftarrow \{\}$ 
2: for each insider  $I$  do
3:    $S_I \leftarrow \{\}$ 
4:   for each transaction of insider  $I$  for company  $C$  do
5:      $TD, TT, TP, \Sigma ST \leftarrow$  transaction date, type, price, sum of shares traded in all the
       transactions with the same TD, TT, and TP
6:      $CP \leftarrow$  market closing price for company  $C$ 's stock on date  $TD$ 
7:      $DV \leftarrow$  dollar volume for company  $C$ 's stock on date  $TD$ 
8:      $R \leftarrow \frac{TP \times \Sigma ST}{DV}$ 
9:     if  $TT = \text{sale}$  then
10:      if  $CP < TP$  then
11:         $S_I \leftarrow S_I \cup R$ 
12:      else
13:         $S_I \leftarrow S_I \cup -R$ 
14:     if  $TT = \text{purchase}$  then
15:      if  $CP > TP$  then
16:         $S_I \leftarrow S_I \cup R$ 
17:      else
18:         $S_I \leftarrow S_I \cup -R$ 
19:    $T \leftarrow T \cup \{S_I\}$ 
20:  $\alpha_{\text{Bonferroni}} \leftarrow \frac{0.01}{|T|}$ 
21: for each sample  $S_I$  in  $T$  do
22:    $a \leftarrow$  p value from one tailed t test with  $H_a : \mu_{S_I} > 0$ 
23:   if  $a < \alpha_{\text{Bonferroni}}$  then
24:     return  $I$ 

```

In Algorithm 1, we first create an empty set T into which we will later insert separate sets consisting of values related to the insiders' transactions (line 1). The procedure then starts to consider each insider one by one (lines 2–19). Specifically, we first create a sample S_I for each insider I (line 3) and for each company that the insider has a transaction for, we consider the non-split transactions of the insider (lines 4–18). We say that a set of transactions are split transactions if they occur on the same date, are of the same type (sale or purchase), and have the same transaction price. We sum the number of shares traded in such transactions and consider them only once as a single

ferent companies. It is therefore important to somehow normalize each transaction of the insider so that a strong signal from one transaction does not affect the overall results. To do so, we obtain a normalized dollar amount for each transaction by multiplying the number of shares traded in the transaction with the transaction price, and dividing the outcome with the dollar volume for the stock

⁷ The dollar volume of a stock is a measure of its liquidity on a given day, and it is computed by multiplying the volume of the stock (i.e., total number of shares traded) on a day with the market closing price of the stock on the same day.

(line 8). Note that this ratio is >0 and almost always upper-bounded by 1,⁸ and it denotes the “magnitude” of the transaction in dollars relative to the other transactions on the same date. After obtaining this ratio, we compare the transaction price with the market closing price depending on the transaction type, as mentioned above. If the insider buys when the price is low or sells when the price is high in comparison with the market closing price, we add the actual value of the ratio to the sample S_I , otherwise we add the negative of the ratio to the sample (lines 9–18). We call the value included to the sample the *signed* normalized dollar amount for the transaction.

A suspicious case occurs when there are many positive observations in the sample. While at this point we could perform a one-tailed t test with the alternative hypothesis $H_a : \mu_{S_I} > 0$, we would face the multiple testing problem⁹ since the procedure needs to perform a hypothesis test for each insider in the dataset. Therefore, we store each S_I in set T (line 19) and later perform the Bonferroni correction to our predetermined original significance level of 0.01 (line 20). Briefly put, the Bonferroni correction controls the number of erroneous significant results by dividing (thus reducing) the original significance level with the number of hypothesis tests to be performed (Witte and Witte 2009). After obtaining the adjusted significance level, we return to set T and for each sample S_I in set T (lines 21–24), we compute the p value from a one-tailed t test with the alternative hypothesis $H_a : \mu_{S_I} > 0$ (line 22). If the p value is smaller than the adjusted significance level, the procedure returns the insider associated with the sample in consideration (lines 23–24).

We now discuss the results we obtain after applying the procedure to the dataset. We should note that all the transactions we consider occur on dates that are prior to their Form 4 filing dates (i.e., the dates on which the Form 4s become public). It is therefore unlikely that the stock prices on the dates of the transactions are affected by the public's reactions to the insiders' trade disclosures. We retrieve the market closing prices and the volumes of the stocks from the Center for Research in Security Prices (CRSP).¹⁰ We exclude the small number of transactions ($<0.01\%$) that have a normalized dollar amount greater than 0.5, as they might be subject to data entry errors. After eliminating these transactions and the transactions with a missing transaction date, type, price, or number of shares

Table 2 The insiders with a significant statistical result from Algorithm 1, ranked in descending order by the number of transactions they have

Insider	Transactions	Individual	Sectors and roles
1	1233	No	T-B, E-B, 2xCS-B
2	970	Yes	CS-D, CG-D, CD-D
3	501	No	H-B
4	433	No	12xH-B, CND-B
5	373	No	F-B, T-B
6	352	No	CG-B
7	213	Yes	CG-CEO
8	206	Yes	E-CEO
9	175	No	CND-B
10	162	Yes	CG-D, T-D
11	155	Yes	CG-D, CD-D
12	110	No	T-B
13	110	No	3xH-B, 2xF-B, 2xT-B, 1xCS-B
14	101	Yes	F-CEO
15	94	No	7xT-B
16	90	Yes	CS-CEO
17	71	Yes	E-CEO
18	54	Yes	CS-D
19	49	Yes	F-CEO
20	47	Yes	H-OO
21	46	Yes	F-OO
22	41	Yes	E-OO
23	31	Yes	CG-OO
24	27	Yes	CD-CFO
25	26	Yes	H-CFO
26	26	Yes	BI-OO
27	23	Yes	BI-B
28	18	Yes	CND-OO
29	18	Yes	CND-OO

traded value, the remaining sample consists of transactions for roughly 48k insiders. This means that our adjusted significance level is close to 10^{-7} .

Table 2 lists the 29 insiders returned from the procedure with significant statistical results. The list is ranked in descending order according to the number of transactions. We also report whether the insider is an individual or a company, the insider's companies' sectors, and the insider's roles in the companies. Recall that a company can be an insider of another company as a beneficial owner if it holds more than 10 % of the company's stock. The possible sectors for the companies are technology (T), energy (E), consumer services (CS), capital goods (CG), consumer durables (CD), healthcare (H), consumer non-durable (CND), finance (F), transportation (T), and basic industries (BI). We report the sectors and role codes in pairs, e.g., T-B means that the sector of the insider's company is

⁸ The scenarios leading to a ratio >1 are very unrealistic, e.g., on a given day all the trades for a company's stock should be performed by a single insider; the dataset confirms our belief.

⁹ The multiple testing problem arises when testing multiple hypotheses simultaneously. In this setting, the likelihood of observing an erroneous significant result purely by chance increases with the number of tests performed (Witte and Witte 2009).

¹⁰ <http://www.crsp.uchicago.edu/>.

technology, and the role of the insider in the company is beneficial owner. If a pair appears more than once, we use the $N \times P$ notation to denote that pair P occurs N times.

The procedure returns more individual insiders than institutional insiders. However, institutional insiders conduct more transactions. The institutional insiders are all beneficial owners, whereas the individual insiders vary in terms of their roles—interestingly CFOs constitute the minority. We see that the institutional insiders are mostly from the healthcare sector, whereas there is more heterogeneity in the sectors represented by individual insiders.

To better illustrate the behavior captured by the procedure, Fig. 12 zooms in and shows the time series of the signed normalized dollar amounts for the transactions of the top-2 insiders in Table 2. Notice that the bulk of the transactions in both time series has positive normalized dollar amounts. This is particularly obvious for Insider 2, who almost consistently times her transactions correctly starting from 2009. While we do not imply that these 29 insiders are earning profits, our results show that certain insiders come very close to doing so by taking the first step and correctly predicting the price movements during the course of a day.

4.4 Constructing networks of insiders

We now study insider behavior from a network perspective. We conjecture that insiders within and across companies may share non-public inside information with each other. We build insider networks—graphs in which insiders

(nodes) with similar trading behaviors are connected (edges)—to identify insiders who might be exchanging information with each other.

We aim to link together insiders who consistently trade on similar dates. But, how can we determine if two insiders are similar enough in terms of trading behavior? The challenge here is to define a similarity function, which takes as input the transaction times of two traders who are insiders of the same company and returns a value denoting the similarity between the timings of the transactions. In this paper, we consider the transactions that occur on the same dates.

We represent the transactions of trader T who is an insider of company C in a set denoted by $T_C = \{t_1, \dots, t_m\}$, where t_j is the date of a transaction. Note that trader T can be an insider of more than one company; however, T_C contains the dates of the transactions only related to company C . We focus on the distinct transaction dates by defining T_C as a set to avoid split transactions of insiders affecting the results.

Our network generation procedure is illustrated in Algorithm 2. We start by forming an empty network G . We then perform a firm-by-firm comparison of the transaction dates of every possible pair of insiders of a firm. That is, for every company C , we compare the sets of transaction dates X_C and Y_C for every possible pair of traders X and Y who are insiders of company C . To avoid insiders having a small number of transactions affecting the results, we only consider the insiders with at least h_z distinct transactions. The similarity function, which we use to compute the similarity between X_C and Y_C , is defined as

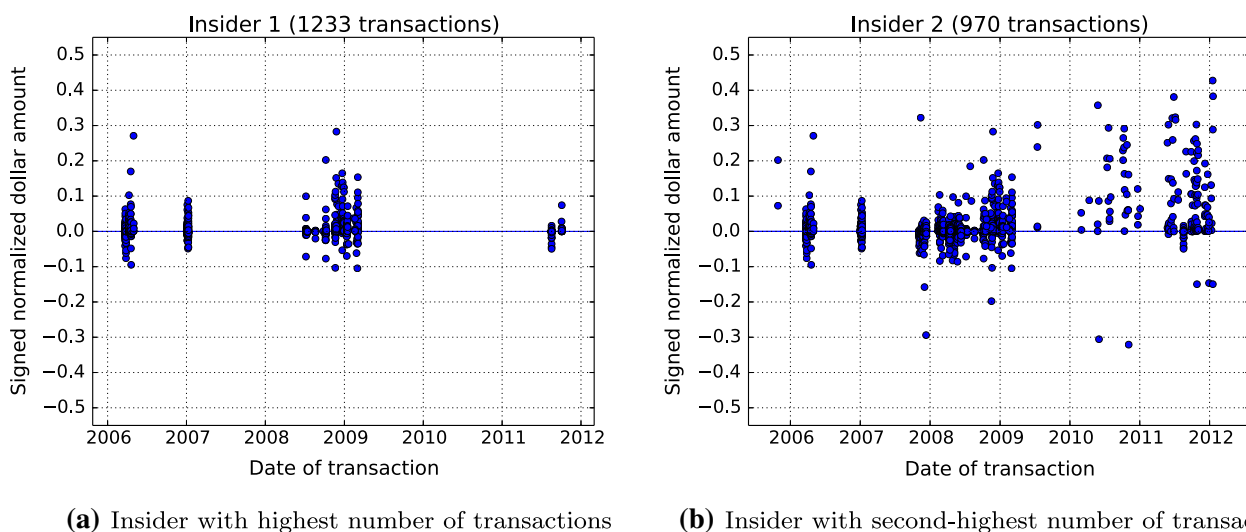


Fig. 12 Time series of the signed normalized dollar amounts for the transactions of the top-2 insiders in Table 2; if the transaction is above the *straight line*, the insider is buying when the price is low or selling

when the price is high in comparison with the market closing price. The bulk of the transactions is located above the *straight line* in both figures, illustrating that our approach can capture this trading behavior

$$S(X_C, Y_C) = \frac{\left(\sum_{i=1}^{|X_C|} \sum_{j=1}^{|Y_C|} I(x_i, y_j) \right)^2}{|X_C| \times |Y_C|}, \quad (2)$$

where $I(x, y)$ is a function that returns 1 if $x = y$ and 0 otherwise. Note that $S(X_C, Y_C)$ is equal to 1 if insiders X and Y always trade on the same date and 0 if insiders X and Y have no common transactions dates. If the similarity between X_C and Y_C is greater than a threshold h_m , we include a node for each of insiders X and Y to network G (if the nodes do not already exist) and form an edge between them.

Algorithm 2 Generate-Network

Return: Insider Network

```

1:  $G \leftarrow$  graph with node set  $N = \emptyset$  and edge set  $E = \emptyset$ 
2: for each company  $C$  do
3:   for each pair of  $X_C$  and  $Y_C$  do
4:     if  $|X_C| \geq h_z$  and  $|Y_C| \geq h_z$  then
5:       if  $S(X_C, Y_C) \geq h_m$  then
6:         if node for insider  $X$ ,  $n_X \notin N$  then
7:            $N \leftarrow N \cup n_x$ 
8:         if node for insider  $Y$ ,  $n_Y \notin N$  then
9:            $N \leftarrow N \cup n_y$ 
10:         $E \leftarrow E \cup$  edge connecting  $n_X$  and  $n_Y$ , labeled company  $C$ 
11: return  $G$ 
```

We now analyze two networks generated using the aforementioned process: the *Sale network* and the *Purchase network*. The first is generated using the sale transactions, whereas the second is generated using the purchase transactions. The reason we focus on sale and purchase transactions is because these transactions are insider-initiated, unlike other transactions in the dataset (e.g., option grants), and thus are more likely to reflect the information flow between the insiders. We do not combine the sale and purchase transactions together because these two types of transactions may have different implications, i.e., traders may purchase shares for different reasons than they sell (e.g., profit vs. diversification). We do not consider beneficial owners in this section because typically they are institutional insiders representing a business entity; our focus here is individual insiders and their relationships with each other. To generate the networks, we set h_z to 5 and h_m to 0.5 based on domain knowledge.

Table 3 shows the simple network parameters for the Sale and Purchase networks. Both networks have a similar number of nodes (insiders), but, as expected, the Purchase network has more edges (each generated due to similar trading behavior for a particular company) than the Sale

Table 3 Simple network parameters

Network	Nodes	Edges	Connected components
Sale	1630	1473	623
Purchase	1678	2656	489

network because an insider has, on average, more sale transactions than purchase transactions in the dataset and the likelihood that two insiders trade on the same dates decreases as they have more transactions overall. As we perform firm-by-firm analysis and not all traders are

insiders of the same single company, both networks are sparse and consist of isolated connected components, such as those in Fig. 13. The Sale network has more connected components than the Purchase network.

Next, we study the sizes of the connected components, i.e., the number of insiders in the components. In Fig. 14, we plot the distributions of the fraction of connected components with a particular size. We observe that most of the connected components in the networks are of size 2, indicating that most insiders of a company do not tend to trade on the same dates. In some sense, this is encouraging as it illustrates that the transaction times can be used as a discriminating factor between insiders, enabling us to extract interesting patterns more easily. Note, however, that there are several components that are considerably large in size, such as the one shown in Fig. 15, which is the largest connected component in the Purchase network.

A trader can be an insider of multiple companies and have similar trading behavior with insiders from each of these companies. When this happens, we observe multiple companies in a connected component, such as the middle triangle in Fig. 13. Table 4 specifies the percent of connected components including a particular number of companies. Note that most connected components in the

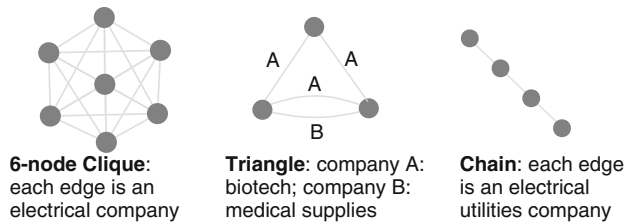


Fig. 13 Examples of connected components from the Sale network. The insiders form different clusters in terms of shape

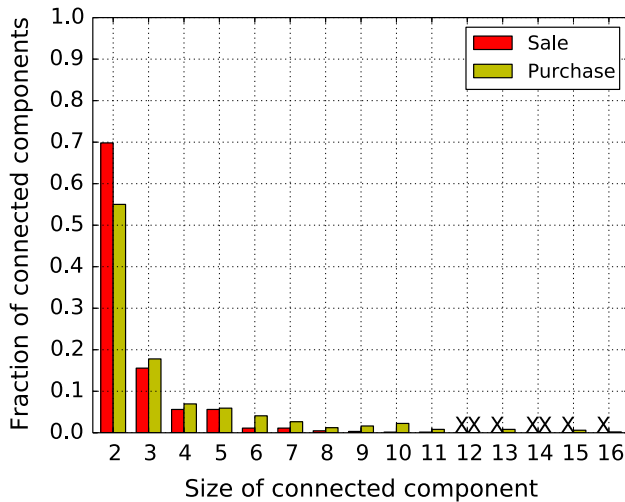


Fig. 14 Distributions of the fraction of connected components with size of a particular value. “X” is used for values that are not applicable. Some insiders form large clusters in which trade-related information might propagate

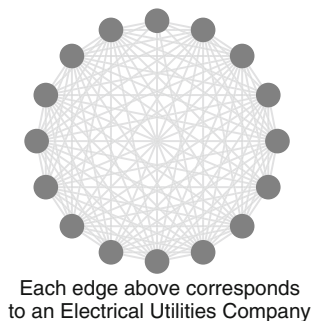


Fig. 15 Largest connected component in the Purchase network: 16 insiders form a “trading clique”

Table 4 Percent of connected components including a particular number of companies. The connected components are homogeneous in terms of the companies of the insiders

	Number of companies						
	1	2	3	4	5	6	7
Sale	96.8 %	2.7 %	-	0.3 %	-	-	0.2 %
Purchase	97.5 %	2.5 %	-	-	-	-	-

networks are homogeneous in the sense that we observe only one company in them. This suggests it is unlikely that there is trade-related information flow about multiple companies between the insiders.

Next, we ask, in a connected component, do insiders with similar or different roles tend to be connected? Fig. 16 shows the *counts* for all combinations of *role pairs* observed in the components (e.g., an edge between CEO-CFO). For instance, in both networks, we observe that, given that an insider is a CEO, it is more likely that she is connected to an OO in the networks, indicating similar trading behavior between CEOs and OOs in general. Assuming that the CEOs are at the top of the corporate hierarchy, followed by CFOs, Ds, and OOs, the interesting observation is that higher-level insiders are more likely to be connected to lower-level insiders, whereas lower-level insider insiders are more likely to be connected to each other. This suggests that there may be both **vertical** (between higher and lower levels) and **horizontal** (between only lower levels) information flow between insiders.

Next, we explore the persistence of the similar trading behaviors of the insiders. Specifically, for each pair of directly connected insiders, we compute the difference in days between their last and first common transactions. Recall that we set h_z to 5; thus, the insiders have at least 5 transactions. We plot the result in Fig. 17. For most of the insiders, we do not observe a common transaction after 1000 days. There are, however, some pairs of insiders who trade similarly in an interval of at least 3000 days. We observe that in general, similar trading behaviors are more persistent with respect to purchase transactions in comparison with sale transactions.

We finally study the collective trading behaviors between the insiders and their neighbors in the networks. We ask, given that all the neighbors of an insider trade on a set of dates, on what fraction of these dates does the insider also trade? Specifically, we consider the connected components in which we observe only one company, say company C , and for each insider X in the connected component, we first retrieve insider X 's neighbors' sets of transaction dates for company C , say $Y_C^1, Y_C^2, \dots, Y_C^n$. We then take intersection of these n sets, $I = Y_C^1 \cap Y_C^2 \cap \dots \cap Y_C^n$, to determine the transaction dates that are common to all the n neighbors of insider X . Subsequently, we retrieve insider X 's set of transaction

dates for company C , X_C , and compute the fraction $\frac{|X_C \cap I|}{|I|}$, which is the fraction of transaction dates of insider X that are common with all the common transaction dates of her neighbors. If $|I| = 0$, we assume that the fraction is 0. We compute a fraction for each insider and take the average of the fractions of the insiders with the same number of neighbors.

Figure 18 shows the results for both the Sale and Purchase networks. Interestingly, we observe an increasing trend that eventually reaches the value 1 in both networks, showing that an insider is likely to trade on a date given that all of her neighbors also trade on that date. Note that our networks contain only the insiders with similar trading behaviors by construction. However, the similarity function we use to construct the networks is defined for only a pair of insiders, i.e., it compares the transaction dates of an insider with those of another insider, and therefore does not ensure collective trading behaviors between the insider and

her neighbors. A partial, mathematical explanation for the increasing trend is that as the number of neighbors increases, the value of the denominator in the fraction decreases. We should note, however, that the lowest positive denominator we obtain is 5 for an insider with 15 neighbors, which is still a high value considering the large number of neighbors.

Some possible reasons for the collective trading behavior are the following. First, there might be information flow from the neighbors to the insiders. In other words, as the number of signals the insider receives increases, she is more willing to trade on a particular date. Second, the insider and her neighbors might have the same internal source of information. For instance, if both the insider and her neighbors are aware of an important company-related event that will soon happen (e.g., merger/acquisition), they are likely to trade on the same dates. Third, the insider and her neighbors might be expected to trade on certain dates, e.g., due to regulations or laws. Again, in this case, it is very likely that they trade on the same dates. We should emphasize, once more, that these are some possible reasons for the collective trading behaviors between the insiders.

4.5 Network-based anomaly detection

To further analyze the Purchase and Sale networks, we would ideally like to examine each node (insider) and evaluate the way it is connected to other nodes in the networks. However, having over one thousand nodes in each of the two networks makes it too tedious for such an exhaustive examination. To conduct such an in-depth

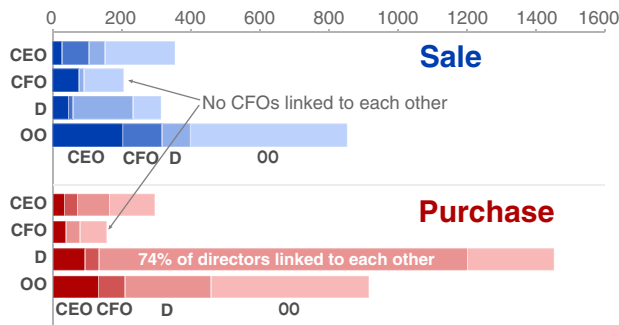


Fig. 16 Counts for all combinations of *role pairs* (e.g., CEO-CFO, D-D), where D is *Director*, OO is *Other Officer*. High-level insiders (e.g., CEO, CFO) are more likely to be linked to low-level insiders (e.g., Director)

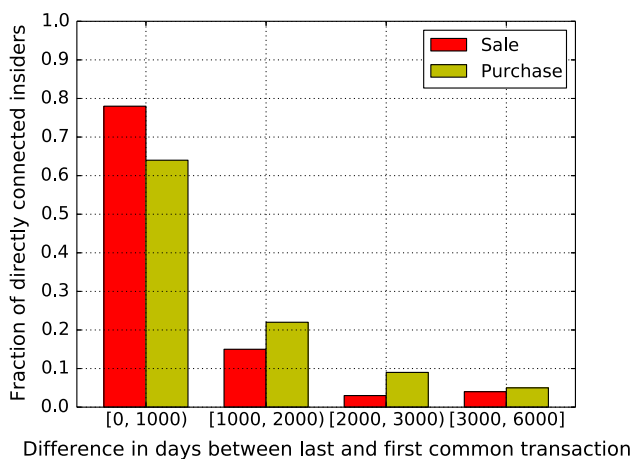


Fig. 17 A comparison of the persistence of the similar trading behaviors of the insiders. The persistence is greater for purchase transactions

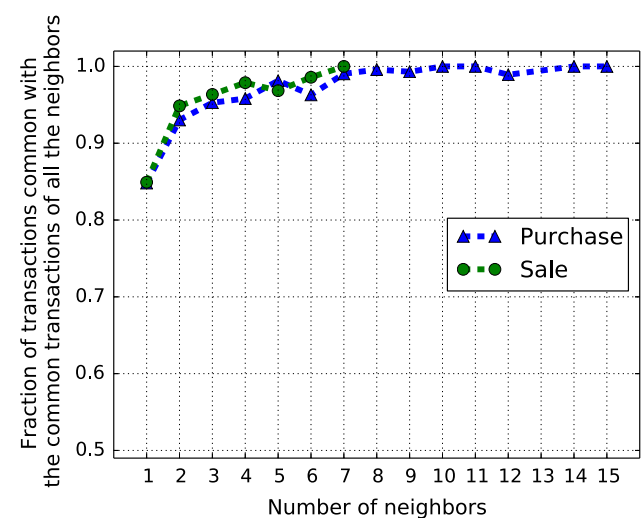


Fig. 18 Collective trading behavior between the insiders and their neighbors: Given that all the neighbors of an insider trade on a date, the insider is likely to trade on the same date

analysis, we seek to flag a small number of nodes as “interesting,” based on some criteria that distinguishes them from the other nodes.

In this section, we seek to detect anomalous nodes in the networks. However, a formal definition of an “anomaly” in the context of networks is elusive: How do we define the *norm*, or the characteristic metrics of a non-anomalous node? Then, how do we quantify the deviation of a given node, relative to this *norm*? Existing work on anomaly detection in graph data has mainly focused on using minimum description length, an information-theoretic principle, to detect anomalous nodes (Eberle and Holder 2007) or edges (Chakrabarti 2004). Alternatively, random walk-based methods have been suggested for identifying outliers in object similarity graphs (Moonesinghe and Tan 2008), or bipartite graphs (Sun et al. 2005). However, these methods exhibit some limitations: While we are interested in detecting anomalous nodes, i.e., insiders, Chakrabarti (2004) focuses on edges; the algorithm of Sun et al. (2005) is designed for bipartite graphs, which does not apply to our networks; Eberle and Holder (2007) assumes some entity-relationship model among the nodes in order to detect anomalies, an assumption that may not be satisfied in our data, and the approach in Moonesinghe and Tan (2008) is difficult to evaluate, given that it requires parameter tuning, which can highly affect the results.

Akoglu et al. (2010) attempt to overcome these difficulties by analyzing the network at the level of *egonets*, where an *ego* is a given node in the network, and its corresponding *egonet* is the subgraph induced from the ego and all its direct neighbors. Their approach is advantageous in that (1) it detects anomalous nodes in *general weighted* graphs, (2) it does not assume any labels on the nodes, (3) it yields results that are easy to interpret, and (4) it is scalable, with linear-time complexity in the size of the network. In what follows, we extract two metrics for each *egonet* in our networks: the number of neighbors (degree) of the ego V_u , and the number of edges in the *egonet* E_u , where u is the ego.

Motivated by the finding in Akoglu et al. (2010) that for many real networks, there exists a power-law relationship between V_u and E_u , we examine the relationship between the two metrics for our networks. Surprisingly, both the Sale and Purchase networks exhibit power-laws for the relationship between V_u and E_u , as illustrated in Fig. 19. The power-law (red line in the figures) is the least-squares fit on the median values of each bucket of points. This line is considered as the *norm* against which we will compare nodes in the networks in order to detect anomalies. More precisely, if y_u is the number of edges in the *egonet* of ego u , and $f(x_u)$ is the *expected* number of such edges according to the power-law fit, when *egonet* u has x_u nodes, we define the distance of a node u relative to the norm, as:

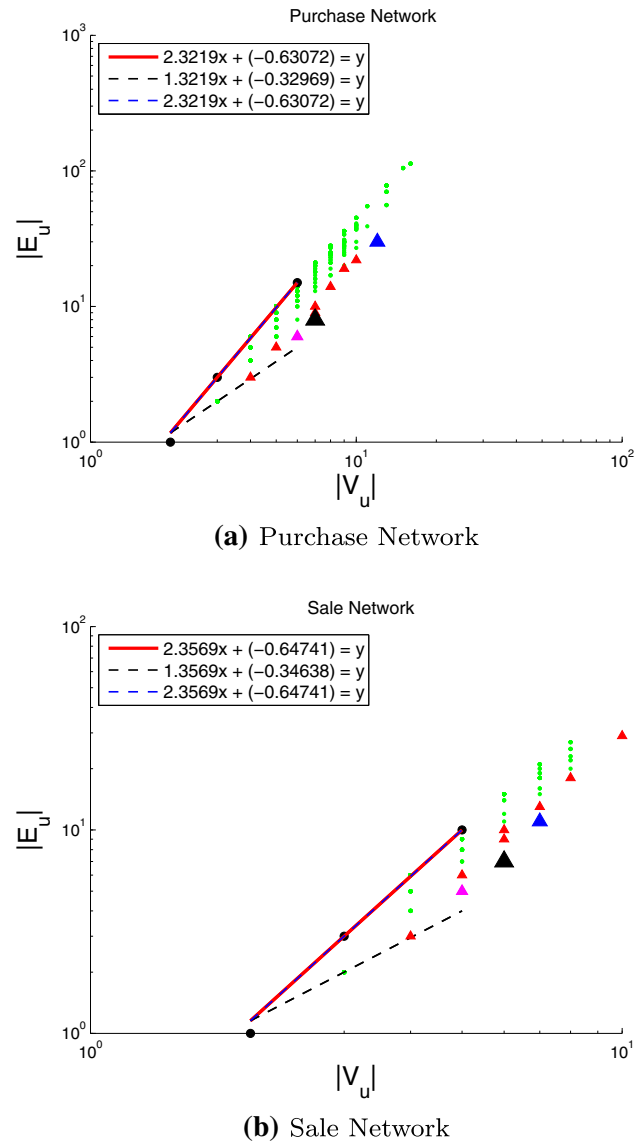


Fig. 19 Distribution of the number of neighbors of each ego (insider), V_u , and the number of edges inside V_u 's *egonet*, E_u , in the networks. The distributions exhibit a power-law relationship. The outlierness of an insider is determined based on the deviations from the power-laws

$$\text{out-distance}(u) = \frac{\max(y_u, f(x_u))}{\min(y_u, f(x_u))} \cdot \log(|y_u - f(x_u)| + 1) \quad (3)$$

The value of $\text{out-distance}(u)$ is zero when (x_u, y_u) is on the power-law line fit and grows with the deviation of (x_u, y_u) from the line. The final outlierness score for u is then its out-distance combined with another outlierness measure used in Akoglu et al. (2010), the Local Outlier Factor (LOF) score of u , which is a density-based measure that flags outliers when they are in a relatively sparse area of the graph. Once we compute the outlierness score of each ego, we simply sort the values in descending order of that score



Fig. 20 Insiders from several companies in different sectors/industries form a long chain in the Sale network

and look at some of the egos with the highest outlieriness scores. In Fig. 19, the ten most anomalous egos in each network are designated with larger triangles indicating higher outlieriness scores. We discuss the interesting findings from this analysis in Sect. 5.

5 Notable observations

In this section, we discuss interesting findings from the network-based analysis and point out directions for future work. The network-based analysis of the insiders' trades reveals some interesting, hidden facts that would otherwise be difficult to discover if we were to analyze the Form 4 filings alone (i.e., the text).

For instance, consider the long chain of insiders in Fig. 20 from the Sale network that was found by our technique. At first glance, one may think that these insiders are from different, unrelated companies. However, with closer look, we find that all of these insiders actually belong to the same investment firm, who may be acting on behalf of the firm. This shows that our approach can indeed extract hidden relationships between insiders from the Form 4 filings.

Second, we find that insiders from the same family tend to trade similarly. Specifically, about 7 % of the directly connected insiders in the networks share the same last names. Manual validation of a subset of these insiders suggests that many are indeed related.

Third, we present an interesting anomalous structure discovered by the method described in Sect. 4.5. Recall that this method flags nodes (or egos) whose neighborhoods' (or egonets) structures deviate from the general pattern across all nodes. In Fig. 21, one such node from the Purchase network and its neighborhood are visualized. Each edge in the figure corresponds to similar trading behavior for the same insurance company. The ego is the middle node in red, which is connected to all of the other nodes as they are its direct neighbors. The thickness of the edges is proportional to the value of the similarity function defined in Eq. 2, which we use to construct the networks. Hence, the thicker the edges, the more similar the two corresponding insiders are in terms of their trading behaviors. What we observe in this instance of the anomaly detection results is an insider (in red) that is connected to three cliques: at the top, a clique formed of six nodes, at the bottom right a clique of three nodes (or triangle), and at the bottom left a clique of two nodes (any two nodes connected

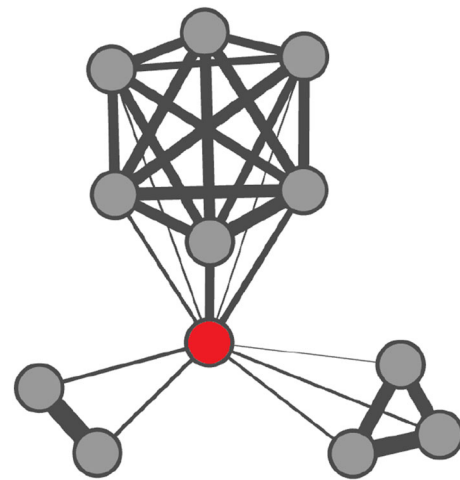


Fig. 21 A visualization of the egonet of the middle node, flagged as anomalous by the method described in Sect. 4: The ego is connected to three cliques, which deviates from the pattern of the power-law fit for the Purchase network in Fig. 19

by an edge form a clique). Even more interestingly, the three cliques are strictly not connected directly among each other. Also, the within-clique similarity is high as highlighted by the thick edges. While we cannot directly assess the reasons behind such a structure, all of the properties of this egonet suggest that the ego (in red) has some intermediary function: the insider trades similar to three distinct mutually exclusive groups of insiders of the same company. This one example highlights the importance of adopting automated anomaly detection methods to facilitate the process of exploratory data analysis and reducing the complexity in a large networked dataset.

6 Conclusions

This work presents the first academic, large-scale exploratory study of the complete insider trading data from SEC. We study the trades by insiders from the temporal and network perspectives. For the former, we explore how the trading behaviors of insiders differ based on their roles in their companies, the types of their transactions, and the sectors of their companies. For the latter, we construct insider networks in which insiders with similar trading behaviors are connected and study the various characteristics of the networks. Additionally, we perform a correlational analysis of prices of insiders' transactions and market closing prices of their companies' stocks, and using

a statistical approach, we determine the insiders who time their transactions well. As future work, we plan to take into account transactions that occur within a time window to capture more patterns in our network-based analysis. Additionally, we intend to incorporate the geographical location information of the insiders' companies into our analyses. We believe our work raises exciting research questions, opens up many opportunities for future studies, and has taken a major step toward helping financial regulators and policymakers understand the dynamics behind insider trading.

References

- Aboody D, Kasznik R (2000) CEO stock option awards and the timing of voluntary corporate disclosures. *J. Account. Econ.* 29:73–100
- Adamic, L, Brunetti C, Harris JH, Kirilenko AA (2010) Trading networks. Social Science Research Network Working Paper Series. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1361184
- Agrawal A., Cooper T (2008) Insider trading before accounting scandals. Social Science Research Network Working Paper Series. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=929413
- Agrawal A, Nasser T (2011) Insider trading in takeover targets. Social Science Research Network Working Paper Series. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1517373
- Akoglu L, McGlohon M, Faloutsos C (2010) Oddball: spotting anomalies in weighted graphs. In: Proceedings of the Pacific-Asia conference on advances in knowledge discovery and data mining
- Bay S, Kumaraswamy K, Anderle MG, Kumar R, Steier DM (2006) Large scale detection of irregularities in accounting data. In: Proceedings of the IEEE international conference on data mining
- Bernhardt D, Hollifield B, Hughson E (1995) Investment and insider trading. *Rev Financ Stud* 8(2):501–543
- Bizjak J, Lemmon M, Whitby R (2009) Option backdating and board interlocks. *Rev Financ Stud* 22(11):4821–4847
- Brockman P, Li X, Price SM (2013) Do managers put their money where their mouths are? Evidence from insider trading after conference calls. Social Science Research Network Working Paper Series. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2200639
- Brockman P, Martin X, Puckett A (2010) Voluntary disclosures around CEO stock option exercises. *J Corp Financ* 16:120–136
- Chakrabarti D (2004) Autopart: parameter-free graph partitioning and outlier detection. In: Proceedings of the European conference on principles of, data mining and knowledge discovery, pp 112–124
- Cheng Q, Lo K (2006) Insider trading and voluntary disclosure. *J Account Res* 44:815–848
- Chin A (1997) Accurate calculation of short-swing profits under section 16(b) of the Securities Exchange Act of 1934. *Del J Corp Law* 22(32):587–599
- Cohen L, Malloy C, Pomorski L (2012) Decoding inside information. *J Financ* 67(3):1009–1044
- Donoho S (2004) Early detection of insider trading in option markets. In: Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining
- Eberle W, Holder L (2007): Discovering structural anomalies in graph-based data. In: Proceedings of the IEEE international conference on data mining workshops, pp 393–398
- Edelson R, Whisenant S (2009) A study of companies with abnormally favorable patterns of executive stock option grant timing. Social Science Research Network Working Paper Series. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1326517
- Fan W, Yu PS, Wang H (2004) Mining extremely skewed trading anomalies. In: Proceedings of the extending database technology conference
- Goldberg HG, Kirkland JD, Lee D, Shyr P, Thakker D (2003) The NASD securities observation, new analysis and regulation system (SONAR). In: Proceedings of the conference on innovative applications of artificial intelligence
- Hamilton J (2011) The SEC's new approach to fraud. Accessed March 22, 2013. <http://www.businessweek.com/magazine/the-secs-new-approach-to-fraud-12082011.html>
- Khandani AE, Lo AW (2011) What happened to the quants in August 2007? Evidence from factors and transactions data. *J Financ Mark* 14(1):1–46
- Kirkos E, Spathis C, Manolopoulos Y (2007) Data mining techniques for the detection of fraudulent financial statements. *Expert Syst Appl* 32(4):995–1003
- Lakonishok J, Lee I (2001) Are insider trades informative? *Rev Financ Stud* 14(1):79–111
- Lorie JH, Niederhoffer V (1968) Predictive and statistical properties of insider trading. *J Law Econ* 11(1):35–53
- McGlohon M, Bay S, Anderle MG, Steier DM, Faloutsos C (2009) SNARE: a link analytic system for graph labeling and risk detection. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining
- Moonesinghe H, Tan PN (2008) Outrank: a graph-based outlier detection framework using random walk. *Int J Artif Intell Tools* 17(01):19–36
- Odean T (1998) Are investors reluctant to realize their losses? *J Financ* 53(5):1775–1798
- Pandit S, Chau DH, Wang S, Faloutsos C (2007) NetProbe: a fast and scalable system for fraud detection in online auction networks. In: Proceedings of the international conference on world wide web
- Ruiz EJ, Hristidis V, Castillo C, Gionis A, Jaimes A (2012) Correlating financial time series with micro-blogging activity. In: Proceedings of the ACM international conference on web search and data mining, pp 513–522
- Summers SL, Sweeney JT (1998) Fraudulently misstated financial statements and insider trading: an empirical analysis. *Account Rev* 73(1):131–146
- Sun J, Qu H, Chakrabarti D, Faloutsos C (2005) Neighborhood formation and anomaly detection in bipartite graphs. In: Proceedings of the IEEE international conference on data mining
- Tamersoy A, Xie B, Lenkey SL, Routledge BR, Chau DH, Navathe SB (2013) Inside insider trading: Patterns & discoveries from a large scale exploratory analysis. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining
- The U.S. Security and Exchange Commission Electronic Data Gathering, Analysis, and Retrieval (EDGAR) System: <http://www.sec.gov/edgar.shtml>
- Witte RS, Witte JS (2009) Statistics, 9 edn. Wiley