2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

# Inside Insider Trading: Patterns & Discoveries from a Large Scale Exploratory Analysis

Acar Tamersoy
College of Computing
Georgia Institute of Technology
tamersoy@gatech.edu

Bo Xie
College of Computing
Georgia Institute of Technology
bo.xie@gatech.edu

Stephen L. Lenkey
Division of Economic and Risk Analysis
U.S. Securities and Exchange Commission
LenkeyS@sec.gov

Bryan R. Routledge
Tepper School of Business
Carnegie Mellon University
routledge@cmu.edu

Duen Horng Chau
College of Computing
Georgia Institute of Technology
polo@gatech.edu

Shamkant B. Navathe
College of Computing
Georgia Institute of Technology
sham@cc.gatech.edu

***Abstract*—How do company insiders trade? Do their trading behaviors differ based on their roles (e.g., CEO vs. CFO)? 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 fillings from the U.S. Securities and Exchange Commission (SEC). We analyzed 12 million transactions by 370 thousand insiders spanning 1986 to 2012, the largest reported in academia. We explore the temporal and network-centric aspects of the trading behaviors of insiders, and make surprising and counter-intuitive discoveries. We study how the trading behaviors of insiders differ based on their roles in their companies, the transaction types, the company 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 towards these dynamics.**

## I. 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 [1]). As many recent high profile cases highlight, illegal insider trading is actively prosecuted.

Most trades by insiders, however, are not illegal. Insiders are defined as corporate officers, directors and 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.

**Opportunities for Data Mining.** Government regulators are increasingly interested in applying data mining techniques to detect fraud and illegal insider trading [2]. 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 their roles, company sectors, and how their connections with other insiders affect their trades. This knowledge could eventually help detect potential illegal activities at a large scale. Here, we focus on two broad classes of techniques: *temporal analysis* and *network analysis*. First, tools that analyze the time series of insiders' trades are important because, as we show, insiders' trading behaviors were 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-centric analysis is crucial for detecting illegal insider trading since insiders often share information through their social networks. Through network-based techniques, we could uncover the hidden communication channels through which the insider information flows, and better understand how insiders operate collectively.

**Exploring Insider Filings.** This work explores a large dataset of the *SEC Form 4 filings*, which describe every change in the ownership interests of insiders in their firms. Details about the dataset are described in Section II.

We examine transactions over time from various perspectives. Profiling various aspects of the trades, such as the company's sector, individual's role within the corporation and the type of transaction, we find distinct patterns along

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.

*ASONAM'13*, August 25-29, 2013, Niagara, Ontario, CAN

each dimension, suggesting a highly multi-modal nature of underlying factors. At the same time, insiders' trades seem to be influenced by common factors, such as market cycles and regulations, independent of the aforementioned characteristics.

**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 non-parametric way. As such, our results could form a basis to initiate an examination of a particular set of insiders' transactions that are suspicious. We envision use by financial regulators and policymakers as the most likely avenue for deploying our research. Our analysis also has the potential to spur future research by economists and legal scholars.

**Contributions.** We conduct an extensive large-scale analysis of insider trading data using the SEC Form 4 filings. Our analysis consists of two major components. The first is the temporal analysis, where we discover patterns in the data by partitioning on corporate roles, sectors and transaction types. The second is the network-centric analysis. In particular, we construct networks of insiders based on the similarity in insiders' transaction timing. Our main contributions include:

- We perform the first academic, large-scale exploratory study of insider filings and related data from SEC;
- We discover distinctive temporal patterns in insiders' trades that may be explained by government regulations, corporate policies, role differences (e.g., CEO vs. board member), and company sectors;
- We find strong evidence that insiders form tightly connected clusters; trade-related information propagate both **vertically** (between higher and lower level insiders) and **horizontally** (among lower level insiders).

Our work takes a computational and statistical modeling approach towards the challenging problem of uncovering correlations among insiders. As we will 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. In this paper, we will replace the names of insiders and companies with generic symbols (e.g., company A).

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

## II. Dataset

United States federal law requires corporate insiders to report their open-market transactions and other ownership changes to SEC within 2 business days via Form 4. The filing period for Form 4 was originally 10 days, but changed to 2 days effective August 29, 2002. A Form 4 consists of two parts, namely Part 1 and Part 2. Part 1 contains 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 to SEC.

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

TABLE I. Summary statistics for our dataset. We focus on sale and purchase transactions.

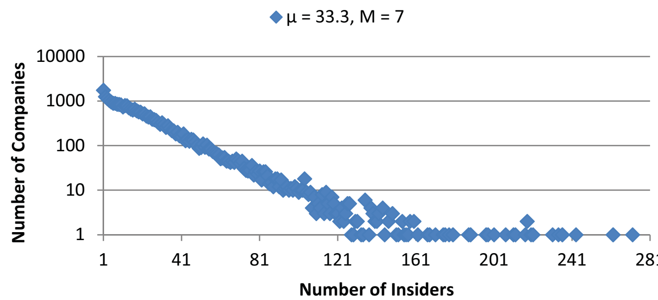


Fig. 1. Distribution of the number of companies having a particular number of insiders. $\mu$ and M stand for the mean and median of the distribution, respectively. Note that the y-axis is in log-scale. The data follows exponential distribution with $\lambda = 0.035$ (p-value $< 0.05$).

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 U.S. Table I provides a set of summary statistics for the dataset. Each record in the dataset contains information about a transaction by an insider. Some informative fields include the name and company of the insider, transaction date and type, number of shares traded, transaction price, and role of the insider in the company. Unfortunately, several fields, in particular the price and number of shares traded, are sometimes either empty or have invalid values. Figs. 1 and 2 show the distributions of the number insiders per company and the number of transactions per insider, respectively. We observe that both distributions have relatively heavy tails, indicating that most companies (insiders) have a few number of insiders (transactions), however there are few companies (insiders) with a significant number of insiders (transactions).

We store the dataset in a SQLite database for ease of analysis. The database contains both parts of the SEC Form 4 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 [3].

## III. Related Work

This work intersects several research areas. We group the related work into the following 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 analyze the SEC Form 4 data at scale.

**Profiling Insiders.** In the finance domain, Cohen et al. [4] characterize insiders into routine traders and opportunist traders. The authors demonstrate that the routine trades do not carry information in predicting future company events or

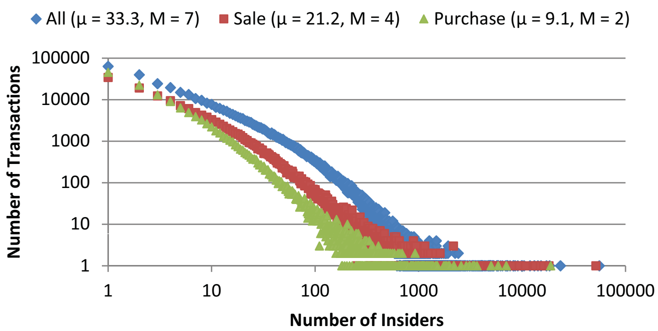


Fig. 2. Distributions of the number all/sale/purchase transactions made by a particular number of insiders. $\mu$ and M stand for the mean and median of the distribution, respectively. Note that both axes are in log-scale. The data for all/sale/purchase transactions follow power-law distribution with $\alpha_{all} = 1.71, \alpha_{sale} = 1.31$, and $\alpha_{purchase} = 1.46$ (p-values $< 0.05$).

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-centric 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 et al. [5] 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. [6] find that managers release abnormally positive news before stock option exercises to obtain relatively high sales prices and Aboody et al. [7] show that managers tend to release bad news before stock option grants to fix lower strike prices. Brockman et al. [8] examine the relation 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.

**Detecting Fraud and Illegal Trades.** Goldberg et al. [9] describe the Securities Observation, News, Analysis and Regulation (SONAR) system, which flags unusual price and volume movement 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 [10] 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 focusing on the more challenging stocks trading and utilize the SEC filings to protect the market and its investors from potential losses.

Other works that use data mining techniques for fraud detection include SNARE [11], 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 [12]. A more general system, Sherlock [13] 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.

**Mining Financial Data.** Fan et al. [14] presents a data mining based automatic trading surveillance system for large data with skewed distribution using multiple classifiers. Bizjak et al. [15] document the network structure in the interlocking board of directors to explain how inappropriately backdating compensation spreads. Adamic et al. [16] 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 these works above, we analyze more 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 study, via time series and network-centric analyses, the Form 4 data, at its largest scale available.

## IV. Patterns, Observations & Analysis

We hypothesize that two important factors reveal information from insiders' transactions. One is the timing of a transaction. If insiders place their transactions in around major corporate events, it is likely that the transactions are based on information. Otherwise, if they just trade routinely on the same month every year, it is more likely the trades are for liquidity or diversification reasons [4]. The second factor is the connection between insiders. If a sub-group of insiders trade similarly, they are more likely to be sharing information with each other.

Based on these assumptions, we analyze the insider trading data from two aspects, namely, time series and network-centric analyses. Next, we present the results of these analyses.

### A. Time Series in Different Facets

We analyze trends in the time series of insider transactions. Since many factors contribute to the timing of transactions, we break down the data according to transaction types, role codes and sectors of companies to examine the importance of each factor. For each facet, we plot only the most common categories. The analysis provides insights into the specific behaviors of different insiders and companies.

Analyzing transaction types reveals interesting patterns as shown in Fig. 3. In general, the number of sales is greater than that of purchases. This is especially significant during 2003-2008. Many insiders receive shares of stock as 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

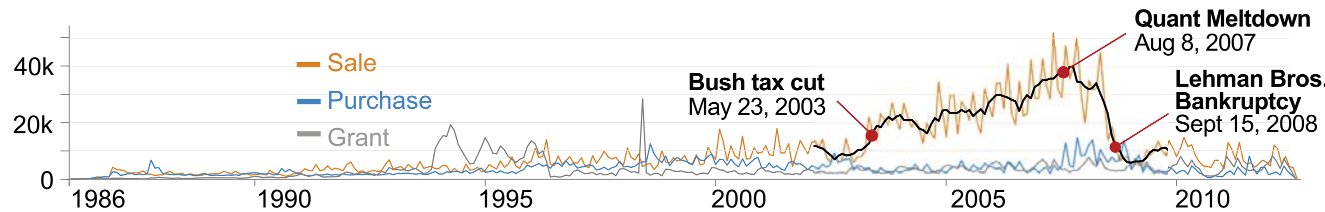


Fig. 3. The daily count of *Purchase*, *Sale*, and *Grant* transactions (the most common types) over 1986-2012. 180-day centered moving average for Sale transactions shown in black. The Bush tax cut in 2003 (reduced capital gains taxes) boosted Sale transactions for following years. Financial crises like the "Quant Meltdown" in 2007 and the burst of "housing bubble' in 2008 suppressed them.

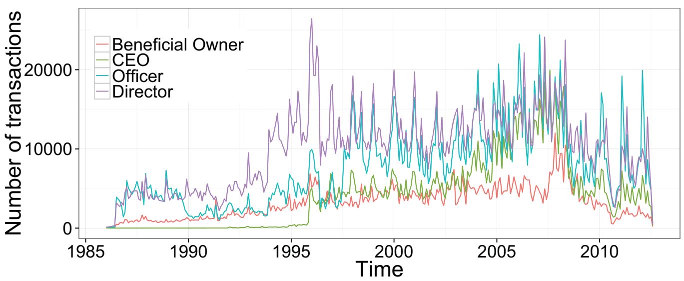


Fig. 4. Transactions break down by role codes. Only the most frequent four codes are shown. Beneficial owners behave differently than the other insiders.

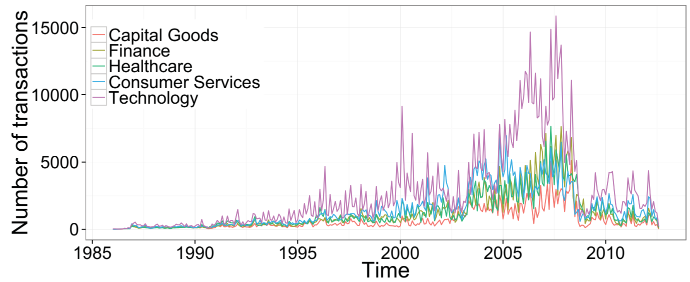


Fig. 5. Transactions break down by sectors. Only the most frequent five sectors are shown. Most activity comes from the technology sector.

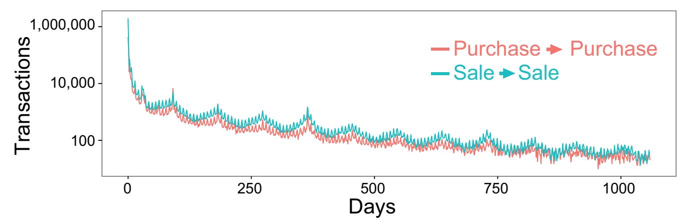


Fig. 6. Time between consecutive transactions of the same type: Purchase-after-Purchase (Purchase→Purchase) and Sale-after-Sale (Sale→Sale). The pattern is oscillatory, with a cycle of about 90 days.

sale transactions coincides with the 2003 change in the United States tax law (enacted May 23, 2003) that reduced capital gains taxes, as illustrated in Fig. 3. The sharp drop in sales occurs after the "Quant Meltdown" of August 2007 [17] (a point identified, with hindsight, as the start of the financial crisis), 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 [18]). An alternative explanation for the drop in sales is that executive stock options, which are often granted at-the-money, became worthless by the time they vested after 2008 and were never exercised.

Fig. 4 illustrates that people with different jobs have different trading patterns. Most transactions are made by directors and officers for the simple 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. These individuals are more of standard insiders, while beneficial owners are not really insiders in the sense that they usually do not have access to the detailed operations of a company. Such an information gap is observed in their trading patterns. Their selling activity is increased only towards the eve of the financial crisis. Shortly after the crisis, their activity level keeps decreasing even though the transactions of other insiders fluctuate during the same period.

Examining the data from a different perspective, Fig. 5 depicts trading activities 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 show up 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. 3. This is likely due to technology companies preferring to compensate their executives with equity.

### B. *Analyzing Transaction Intervals*

While exact transaction dates inform our understanding of trends through time, they do not reflect the patterns within the sequence of transactions. What fraction of insiders sell after purchase and what fraction keep selling or keep purchasing? To answer these questions, we look at the distribution of transaction intervals between two consecutive trades.

Here, we only analyze open-market purchase and sale transactions. We present the number of transactions versus the interval between the transactions in Figs. 6 and 7. In general, Purchase-after-Sales (S→P) and Sales-after-Purchase (P→S) transactions are less common than Purchase-after-Purchase (P→P) and Sales-after-Sales (S→S) transactions. This can be attributed to the phenomenon within technology companies, where most people obtain stocks not from open market purchases, but from stock grants. Insiders keep selling to liquidate or diversify their assets. As a result, the number of S→S transactions is greater than that of P→P transactions. Another notable phenomenon is that the pattern is strongly oscillatory, with a cycle of about 90 days. This is probably due to the corporate bylaws that prohibit transactions near quarterly earnings announcements. For P→S and S→P, the

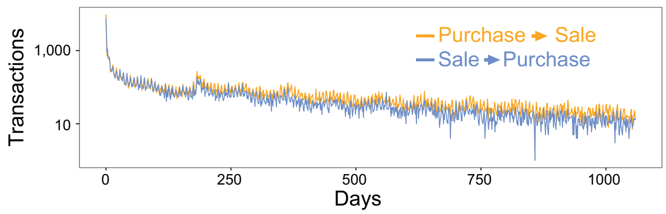


Fig. 7. Time between consecutive transactions of different types: Purchase-after-Sale (Sale→Purchase) and Sale-after-Purchase (Purchase→Sale). The highest peaks are at the point corresponding to six months.

highest peaks are seen at the point corresponding to 6 months. This is probably caused by the short-swing profit rule codified in §16(b) of the Securities Exchange Act of 1934 (i.e., certain insiders may be required to disgorge any profit realized from a combined purchase and sale, or sale and purchase, that occur within six months of each other). For S→S and P→P, the highest peaks are seen at the point corresponding to 3 months (quarterly). These are likely to be routine trades.

To examine how insiders in different roles trade consecutively, we plot the transaction intervals for various role codes in Fig. 8. An interesting observation is that the beneficial owners behave differently from the other individuals. Their number of transactions decreases almost monotonically, without the oscillatory pattern observed in the other types of insiders. This might be explained by the fact that the beneficial owners are effectively "outsiders" – they may not be directly affiliated with the company and, consequently, may be exempt from the corporate bylaws. We further observe that the patterns for the "real insiders" differ among each other. For example, the officers have significantly more S→S sequences than P→P sequences. This, again, is related to the stock options and grants given to them as part of their compensation package. The directors are generally fewer in number and typically do not receive as much stock compensation.

Fig. 9 illustrates that the sectors of the companies also affect how the insiders affiliated with them trade consecutively. For example, we observe that the insiders in the technology sector sell more than they purchase, while in finance, the number of purchase and sale transactions are more balanced. This may be attributed to how the insiders are compensated in different sectors. For instance, the employees in the technology sector often receive company shares or options as part of their compensation, hence most of their stock holdings are not derived from open-market purchases.

### C. Constructing Networks of Insiders

We now study insider behavior from a network-centric perspective. We conjecture that insiders within and across companies 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


Fig. 8. Transaction interval by different role codes. Insiders in different roles trade differently.

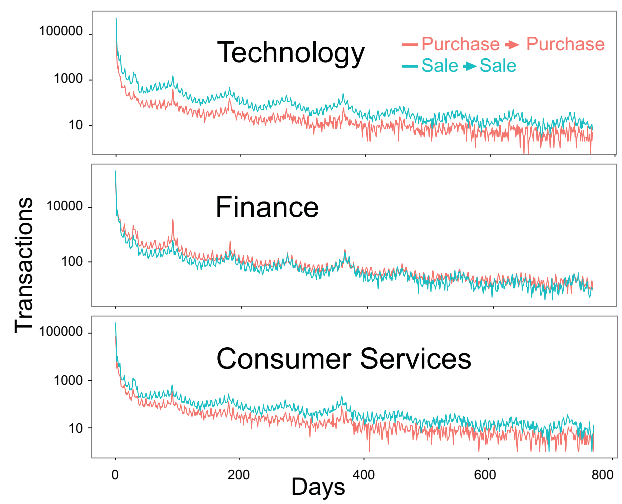


Fig. 9. Transaction interval by different sectors. Insiders in different sectors trade differently.

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 $T_C = \{t_1, ..., t_m\}$, where $t_j$ is the date of a transaction. $|T_C|$ denotes the *size* of $T_C$, defined as the number of transaction dates in $T_C$. Note that trader $T$ can be an insider of more than one company, however $T_C$ contains the dates of those 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.

The network generation process is illustrated in Algorithm 1. 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 possible company $C$, we compare the sets of transaction dates $X_C$ and $Y_C$ for every possible pair of traders $X$ and $Y$ who

**Algorithm 1** Generate-Network

**Return:** Insider Network
$G \leftarrow$ graph with node set $N = \varnothing$ and edge set $E = \varnothing$
**for each** company $C$ **do**
  **for each** pair of $X_C$ and $Y_C$ **do**
    **if** $|X_C| \geq h_z$ and $|Y_C| \geq h_z$ **then**
      **if** $S(X_C, Y_C) \geq h_m$ **then**
        **if** node for insider X, $n_X \notin N$ **then**
          $N \leftarrow N \cup n_x$
        **if** node for insider Y, $n_Y \notin N$ **then**
          $N \leftarrow N \cup n_y$
        $E \leftarrow E \cup$ edge connecting $n_X$ and $n_Y$, labeled company $C$
**return** $G$

TABLE II. SIMPLE NETWORK PARAMETERS

| Network | Nodes | Edges | Connected Components |
|---|---|---|---|
| Sale | 1630 | 1473 | 623 |
| Purchase | 1678 | 2656 | 489 |

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 follows:

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

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 $Y$ trade on the exact same dates 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 insiders $X$ and $Y$ to network $G$ (if the nodes do not already exist) and form an edge between them.

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). To generate the networks, we set $h_z$ to 5 and $h_m$ to 0.5. We obtain results that are qualitatively similar for various values of the aforementioned threshold parameters.

Table II 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 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 a single company, both networks consist of isolated connected components, such as those in Fig. 10. The Sale network has more connected components than the Purchase network (see Table II).

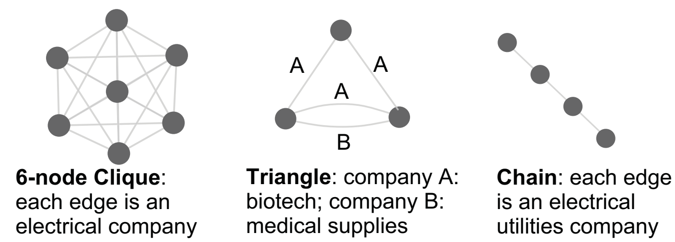


Fig. 10. Examples of connected components from the Sale network. The insiders form different clusters in terms of shape.

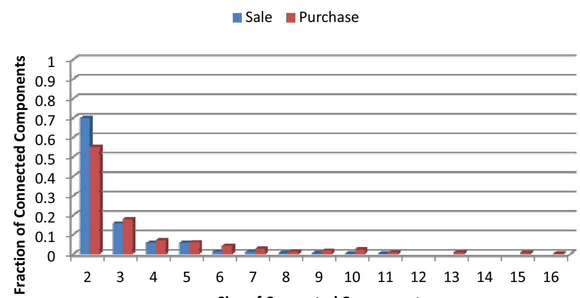


Fig. 11. Distributions of the fraction of connected components with size of a particular value. Some insiders form large clusters in which trade-related information might propagate.

Next, we study the sizes of the connected components, that is, the number of insiders in the components. In Fig. 11, we plot the distributions of the fraction of connected components with size of a particular value. 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. 12, which is the largest connected component in the Purchase network.

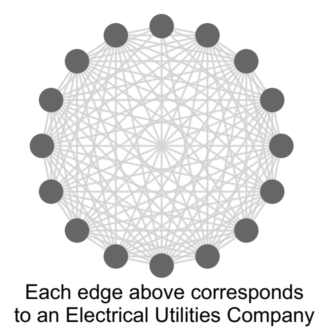


Fig. 12. Largest connected component in the Purchase network: 16 insiders form a "trading clique".

We then study how tightly connected the insiders are in the connected components. Do the insiders form dense clusters in which an insider's neighbors are also connected, such as the left clique in Fig. 10, or are they sparsely connected such that an insiders' neighbors are not connected, such as

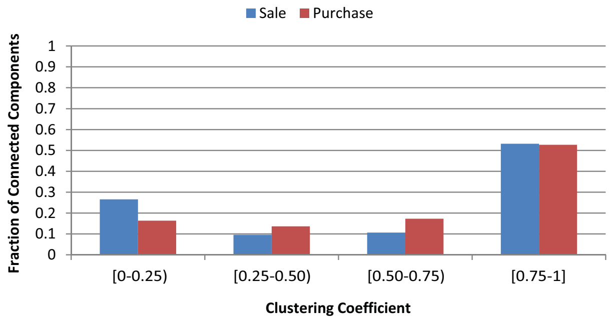


Fig. 13. Distributions of the fraction of connected components with clustering coefficients in a particular interval. The insiders form tightly connected clusters.

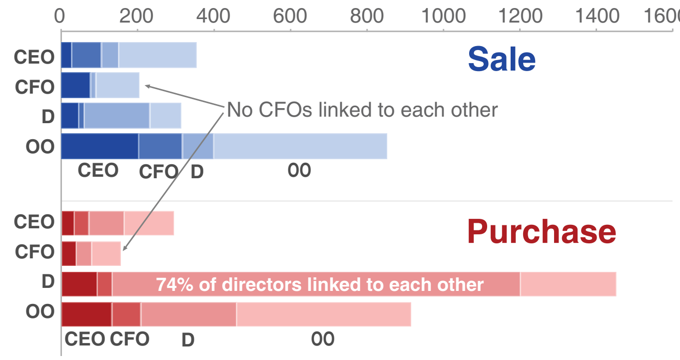


Fig. 14. Counts for all combinations of *role pairs* (e.g., CEO-CFO, D-D), where D is *Director*, OO is *Other Officer*. High-level insiders (e.g., CEO, CFO) more likely to be linked to low-level insiders (e.g., Director).

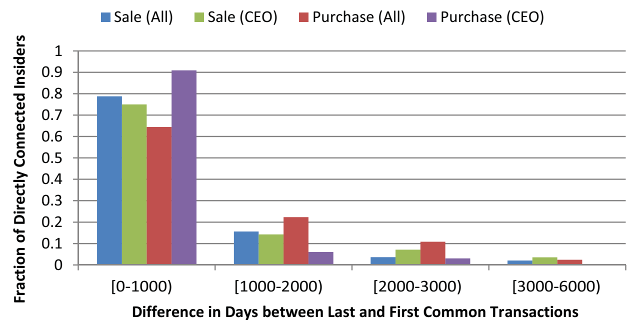


Fig. 15. A comparison of the persistence of the similar trading behaviors of the insiders. The persistence is greater for purchase transactions.

the right chain in Fig. 10? To answer this question, we use the clustering coefficient measure [19]. The local clustering coefficient is a measure of how well connected are the nodes around a given node. The clustering coefficient is then the mean of the local clustering coefficients for all the nodes in a subgraph/graph. In Fig. 13, we plot the distributions of the fraction of connected components with clustering coefficients in a particular interval. Note that the clustering coefficient is undefined for subgraphs/graphs of size 2, thus we ignore them in the analysis. We observe that, in both networks, a significant fraction of the components have large clustering coefficients, indicating that the insiders are tightly connected in the components. This suggests trade-related information may propagate very easily between the insiders.

TABLE III. 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% | - | - | - | - | - |

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. 10. Table III specifies the percent of connected components including a particular number of companies. Note that most connected components in the 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? Each insider reports at least 1 and at most 4 role codes when a Form 4 is filed. There are over 50 possible roles, ranging from Chairman of the Board to Retired. Unfortunately, there is no strict standards as to when an insider should use a particular code (i.e., a role code's job nature is only loosely defined). Previous work has proposed heuristics to map specific role codes to more general ones. Here, we use the mapping from [20], 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). For each insider, we obtain a single role code from each Form 4 and we here consider the pairs of insiders that receive only one general role code after the mapping. Fig. 14 shows the *counts* for all combinations of *role pairs* (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, then come CFOs, Ds, and OOs, respectively, 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 it is likely that there is both **vertical** (between higher and lower levels) and **horizontal** (between only lower levels) information flow between the 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 that occur on the same dates. We plot the result in Fig. 15. 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. Even though we observe more persistent behavior in the Purchase network than the Sale network, it is interesting to see that the CEOs are less persistent in terms of the purchase transactions than the sale transactions.


Fig. 16. Insiders from several companies in different sectors/industries form a long chain in the Sale network.

## V. Discussion of Case Studies & Future Work

We now discuss interesting findings from the network-based analysis as case studies and point out directions for future work. The network-centric analysis of the insiders' trades reveals some interesting, hidden facts, that would otherwise be hard 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. 16 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 would 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. The manual validation of a subset of these insiders suggests that many are indeed related.

So far, we have considered the transactions that occur on the same dates when building the insider networks. We plan to take into account transactions that occur within a time window $w$ to capture more patterns. Furthermore, we intend to explore additional dimensions along which we can generate the insider networks. One possible dimension would be the performance of the insiders in terms of making trade-related decisions, such as when did they buy or sell stocks. As another dimension, we plan to experiment with other similarity functions.

## VI. 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-centric 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. Our work raises exciting research questions and opens up many opportunities for future studies. We believe our work has taken a major step towards helping financial regulators and policymakers understand the dynamics behind insider trading.

## References

[1] D. Bernhardt, B. Hollifield, and E. Hughson, "Investment and insider trading," *The Review of Financial Studies*, vol. 8, no. 2, pp. pp. 501–543, 1995. [Online]. Available: http://www.jstor.org/stable/2962280

[2] J. Hamilton, *The SEC's New Approach to Fraud*, 2011 (accessed March 22, 2013), http://www.businessweek.com/magazine/the-secs-new-approach-to-fraud-12082011.html.

[3] The U.S. Security and Exchange Comission Electronic Data Gathering, Analysis, and Retrieval (EDGAR) System, http://www.sec.gov/edgar.shtml.

[4] L. Cohen, C. Malloy, and L. Pomorski, "Decoding inside information," *The Journal of Finance*, vol. 67, no. 3, pp. 1009–1044, May 2012.

[5] Q. Cheng and K. Lo, "Insider trading and voluntary disclosure," *Journal of Accounting Research*, vol. 44, pp. 815–848, 2006.

[6] P. Brockman, X. Martin, and A. Puckett, "Voluntary disclosures around ceo stock option exercises," *Journal of Corporate Finance*, vol. 16, pp. 120–136, 2010.

[7] D. Aboody and R. Kasznik, "Ceo stock option awards and the timing of voluntary corporate disclosures," *Journal of Accounting and Economics*, vol. 29, pp. 73–100, 2000.

[8] P. Brockman, X. Li, and S. M. Price, "Do managers put their money where their mouths are? evidence from insider trading after conference calls," *Social Science Research Network Working Paper Series*, 2013.

[9] H. G. Goldberg, J. D. Kirkland, D. Lee, P. Shyr, and D. Thakker, "The nasd securities observation, new analysis and regulation system (sonar)," in *Proceedings of the Conference on Innovative Applications of Artificial Intelligence*, 2003.

[10] S. Donoho, "Early detection of insider trading in option markets," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[11] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos, "Snare: a link analytic system for graph labeling and risk detection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 1265–1274.

[12] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos, "Netprobe: a fast and scalable system for fraud detection in online auction networks," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 201–210.

[13] S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar, and D. M. Steier, "Large scale detection of irregularities in accounting data," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 75–86.

[14] W. Fan, P. S. Yu, and H. Wang, "Mining extremely skewed trading anomalies," in *Proceedings of the Extending Database Technology Conference*, 2004, pp. 801–810.

[15] J. Bizjak, M. Lemmon, and R. Whitby, "Option backdating and board interlocks," *The Review of Financial Studies*, vol. 22, no. 11, pp. pp. 4821–4847, 2009. [Online]. Available: http://www.jstor.org/stable/40468375

[16] L. Adamic, C. Brunetti, J. H. Harris, and A. A. Kirilenko, "Trading networks," *Social Science Research Network Working Paper Series*, 2010.

[17] A. E. Khandani and A. W. Lo, "What happened to the quants in august 2007? evidence from factors and transactions data," *Journal of Financial Markets*, vol. 14, no. 1, pp. 1–46, 2011.

[18] T. Odean, "Are investors reluctant to realize their losses?" *The Journal of Finance*, vol. 53, no. 5, pp. pp. 1775–1798, 1998. [Online]. Available: http://www.jstor.org/stable/117424

[19] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[20] R. Edelson and S. Whisenant, "A study of companies with abnormally favorable patterns of executive stock option grant timing," *Social Science Research Network Working Paper Series*, 2009.