# GRAPH-BASED ALGORITHMS AND MODELS FOR SECURITY, HEALTHCARE, AND FINANCE

A Thesis
Presented to
The Academic Faculty

by

Acar Tamersoy

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Computer Science

School of Computational Science and Engineering
Georgia Institute of Technology
May 2016

# GRAPH-BASED ALGORITHMS AND MODELS FOR SECURITY, HEALTHCARE, AND FINANCE

Approved by:

Professor Duen H. Chau, Advisor and Committee Chair
School of Computational Science and Engineering
*Georgia Institute of Technology*

Professor Rahul C. Basole
School of Interactive Computing
*Georgia Institute of Technology*

Professor Shamkant B. Navathe, Co-advisor
School of Computer Science
*Georgia Institute of Technology*

Dr. Kevin A. Roundy
*Symantec Research Labs*

Professor Munmun De Choudhury
School of Interactive Computing
*Georgia Institute of Technology*

Date Approved: 11 March 2016

# ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Duen H. Chau, and my co-advisor, Professor Shamkant B. Navathe, for their continuous support, guidance, and motivation throughout this work. I am grateful to the other members of my committee, Professor Munmun De Choudhury, Professor Rahul C. Basole, and Dr. Kevin A. Roundy, for their valuable comments and feedback on this thesis.

I would like to also thank my family for their never-ending encouragement. This work would not have been possible without their support.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# SUMMARY

Graphs (or networks) are now omnipresent, infusing into many aspects of society. This dissertation contributes unified graph-based algorithms and models to help solve large-scale societal problems affecting millions of individuals' daily lives, from cyber-attacks involving malware to tobacco and alcohol addiction. The main thrusts of our research are:

(1) Propagation-based Graph Mining Algorithms: We develop graph mining algorithms to propagate information between the nodes to infer important details about the unknown nodes. We present three examples: AESOP (patented) unearths malware lurking in people's computers with 99.61% true positive rate at 0.01% false positive rate; our application of ADAGE on malware detection (patent-pending) enables to detect malware in a streaming setting; and EDOCS (patent-pending) flags comment spammers among 197 thousand users on a social media platform accurately and preemptively.

(2) Graph-induced Behavior Characterization: We derive new insights and knowledge that characterize certain behavior from graphs using statistical and algorithmic techniques. We present two examples: a study on identifying attributes of smoking and drinking abstinence and relapse from an addiction cessation social media community; and an exploratory analysis of how company insiders trade.

Our work has already made impact to society: deployed by Symantec, AESOP is protecting over 120 million people worldwide from malware; EDOCS has been deployed by Yahoo and it guards multiple online communities from comment spammers.

# CHAPTER I

# INTRODUCTION

This thesis is concerned with graph-based algorithms and models to solve large-scale societal problems in the security, healthcare, and finance domains, which are deemed to be among the strategic and high-impact areas in the United States [126]. Graphs (or networks) provide a powerful machinery to model many types of relationships and they offer a convenient abstraction to reason about important problems. The first paper on the subject is considered to be the formulation of the historical seven bridges of Königsberg problem, written by Leonhard Euler in 1736. This problem established what is known as graph theory, and graph-based approaches have since been increasingly developed and applied in many disciplines to solve real-world problems of practical interest [138].

Graphs are omnipresent in today's big data era, infusing into many aspects of our society. This thesis is motivated by the recent calls and efforts towards harnessing big data for social good. The Executive Office of the President in the United States, for instance, recently published a report that encourages the use of big data towards the betterment of society, particularly where existing policies or institutions do not otherwise support such progress [172]. Other examples include the annual Data Science for Social Good programs at the University of Chicago[1] and the Georgia Institute of Technology[2], which bring together data scientists to work on projects with social impact, and IBM's recent Big Data for Social Good Challenge[3] that invited developers to build applications for social benefit. In a similar vein, this thesis leverages graphs

[1] `dssg.uchicago.edu`
[2] `dssg-atl.io`
[3] `ibmhadoop.devpost.com`

from security, healthcare, and finance to benefit societies at large, by helping solve real-world problems affecting millions of individuals' daily lives, from cyber-attacks involving malware to tobacco and alcohol addiction. Our overarching goal is to help solve large-scale societal problems; in doing so, we take a graph-based perspective such that we represent the relationships between the entities central to the problems as well as information about the entities in the form of graphs (with the entities as the nodes, the relationships between the entities as the edges, and information about the entities as the node or edge attributes), based on which we design and develop algorithms and models that contribute towards solving these problems. As an example, a large-scale societal problem we tackle is detecting malware lurking in people's computers. The files are the central entities in this problem and our AESOP algorithm leverages a graph that captures *goodness* information about the files—denoting whether they are malicious, benign, or unknown—and the relationships between the files that tend appear together on people's computers to detect malware with very high accuracy based on the guilt-by-association principle (i.e., an unknown file that consistently appears together with the malicious files is deemed to be malicious). Another large-scale societal problem we tackle is characterizing abstinence from tobacco and alcohol addiction from social media. The social media users who are also abstainers are the central entities in this problem and our supervised learning-based statistical models leverage a graph that captures the interactions between the users on the social media platform to identify the key characteristics of short-term and long-term abstainers, and examine the use of these characteristics in predicting the abstinence status of the individuals.

**Why use graphs and take a graph-based perspective?**

The advantages of using graphs and taking a graph-based perspective to tackle large-scale societal problems are multifold. First, graphs provide a natural representation

of the data in many domains, including those that we consider in this thesis. For instance, in the security domain, the spread of malware between computers naturally forms a graph, with nodes being the computers and the edges corresponding to the transmission of malware from one computer to another. In these cases, this natural representation of the data as a graph helps us more easily understand and explore the data, form hypotheses and verify them, and communicate our findings with other users in the domain. Second, graphs enable us to build powerful and scalable algorithms and models that can incorporate or leverage information about how the entities are related to or associated with each other in the broader context. This includes information about direct relationships between the entities as well as indirect relationships involving additional entities in-between. If entity $A$ is directly related to entity $B$ that is directly related to entity $C$, oftentimes the information about the indirect relationship between entities $A$ and $C$ is also important (e.g., when entity $A$ is a malicious file and entity $C$ is an unknown file, and the principle of guilt-by-association is used to detect malware). These relationships can be captured in a scalable way by the algorithms and models using optimized storage techniques established for graphs, such as adjacency lists, compressed row storage (CRS), and so on [51]. This is particularly beneficial for large datasets containing a significant number of entities, such as those that we consider in this thesis.

In summary, graphs have the ability to naturally capture the relationships between the entities in a structure or topology that can be exploited computationally, which gives an edge in tackling large-scale societal problems as we demonstrate in this thesis. In malware detection, for instance, graphs enable us to capture our novel observation that some files tend to appear together on people's computers. Our Aesop algorithm operates on a graph that represents these co-occurrence relationships. By doing so, it detects malware with much higher accuracy than the existing approaches that treat the files as independent of one another (see [92] for a survey). As another

example, in characterizing smoking and drinking abstinence from social media, graphs enable us to capture the interactions between the social media users who are also abstainers. Specifically, our supervised learning-based statistical models leverage a graph that represents which user provides support to whom by writing comments on their posts on the social media platform. This way, we extend the existing body of research [124, 110] by examining the additional role of interaction in characterizing abstinence.

## 1.1 Thesis Overview and Main Ideas

Next, we provide an overview of the thesis, listing the problems we address and presenting a summary of our contributions. Our research groups into two interrelated topics, which form the main thrusts of the thesis.

### 1.1.1 Part I: Propagation-based Graph Mining Algorithms

In the first part of the thesis, we design and develop graph mining algorithms to propagate the information we possess about the entities (e.g., goodness information about the files in malware detection) between the nodes of our graphs based on the graph structure. Propagation-based algorithms that operate on graphs are useful as it is often the case that we do not possess the same level of information for all the entities in the graph. That is, we might have accurate and certain information for some of the entities, and limited or no information for the others. As an example, in malware detection, we might know for certain that some files are malicious or benign, but there might also be files that we do not know much about, hence are treated as unknown. Then, the careful and systematic propagation of the information we already possess for some of the entities from those entities to the others that we know less about in the graph can reveal important details about the latter entities and enable us to learn more about them. Returning to the previous example, we might assign goodness scores to the files and propagate them from the malicious and benign

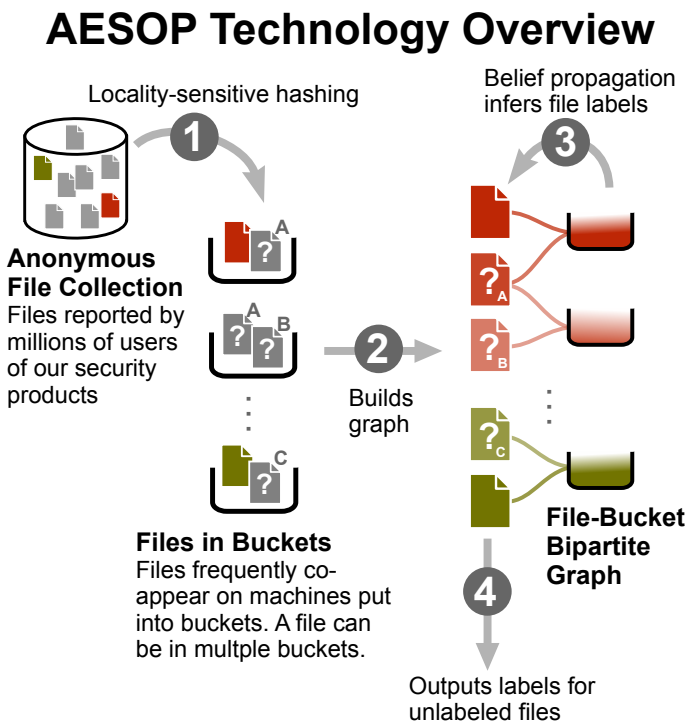



Figure 1: Overview of our AESOP algorithm. AESOP detected malware across over 43.3 million files with 99.61% true positive rate at 0.01% false positive rate.

files to the unknown files in the graph to determine the nature of the unknown files based on, e.g., the guilt-by-association principle (i.e., if an unknown file is related to many malicious files, it would receive low goodness scores from its neighbors in the graph, producing a low goodness score for the file itself). In this part of the thesis, we describe several propagation-based graph mining algorithms.

**AESOP for Malware Detection (Chapter 2).** Detecting malware lurking in people's computers is an important problem because cyber-attacks involving malware have been causing great damage to individuals, organizations, and governments. The majority of the existing techniques either consider each file independently and check if it fits existing profiles of known malware (see [92] for a survey), or leverage the machine-file relationships, denoting which file appears on which machine [38].

We made the novel observation that some files tend to appear together on people's

computers (e.g., multiple files used by the same software) and these co-occurrence relationships can be exploited to detect malware based on the guilt-by-association principle. The idea is that an unknown file that consistently co-occurs with the malicious files might also be malicious, as it might be needed by the latter files to perform certain actions (e.g., communicating with the command and control server). Graphs enable us to capture the co-occurrence relationships between the files. This differentiates us from the existing techniques as they do not consider these relationships. Our Aesop algorithm (Figure 1) leverages a graph that represents such co-occurrence relationships, on which it performs large-scale inference by propagating goodness scores from the malicious and benign files to the unknown files in the graph to determine the nature of the unknown files based on the guilt-by-association principle. As an example, Aesop would assign a low goodness score to an unknown file that consistently co-occurs with the malicious files as it would receive low goodness scores from the neighboring malicious files in the graph.

Aesop detected malware across over 43.3 million files both more accurately (achieving 99.61% true positive rate at 0.01% false positive rate vs. 76.74% true positive rate at 0.01% false positive rate) and sooner (flagging them at least one week sooner) than the state-of-the-art technique [38]. Aesop is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.

**Application of ADAGE to Malware Detection (Chapter 3).** ADAGE is an algorithm that systematically determines the appropriate intervals to construct a sequence of graph snapshots from streaming edges. ADAGE was developed in a joint effort led by our collaborators; we contributed mainly with an extensive case study on malware detection using a propagation-based algorithm to demonstrate the usefulness of ADAGE in practice.

Consider a social network of people, which represents the friendship relationships between the individuals. Assume that the relationships are dynamic (or time-evolving) in that a relationship between two individuals can be formed at any time in the network (hence the streaming relationships or edges). In this setting, analysts often want to grab longitudinal snapshots of the network to study topics such as network growth or evolution of the communities. The current practice in generating the snapshots is to use a single fixed-length interval, whose length is often arbitrarily selected. ADAGE provides a systematic way to determine the appropriate intervals to generate the snapshots.

In the context of malware detection, prior work [38] used a machine-file graph that captures the relationships between machines and files, denoting which file appears on which machine. The prior work infers the nature of the unknown files by propagating goodness scores between the files and the machines in the graph. Assume a setting with a finite stream of time-stamped machine-file relationships. In this case, the prior work would consider the final, full graph that includes all the relationships. We made the novel observation that leveraging the smaller snapshots of the graph generated from the intervals determined by ADAGE can enable us to detect malware more accurately—by propagating goodness scores between the files and the machines as the prior work does—in comparison to using the final graph. This is because it is often the case that infected machines receive a short burst of malicious files over a time-span of minutes, therefore longer snapshots destroy the purity of the graph's connected components by polluting these bursty malware clusters with increasing numbers of benign files. Effectively, longer snapshots lose the finer granularity needed to detect short-lived trends in the data by increasing the graph's density.

We validated our observation with an extensive case study over 574 thousand files, achieving an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots in comparison to 43% true positive rate at 0.01% false positive rate

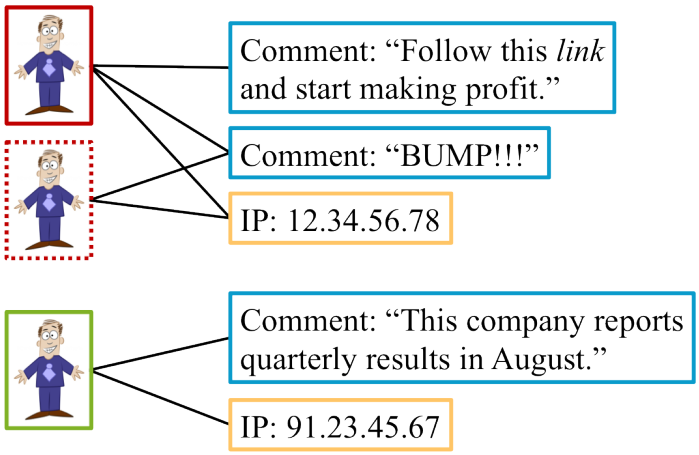



Figure 2: Our EDOCS algorithm leverages a graph that captures the relationships between the social media users and the effort-requiring resources of comment messages and IP addresses to detect comment spammers. In this toy graph, the users in the red and green rectangles are spammers and a legitimate user, respectively. (Cartoon image from `wikihow.com`)

with the final graph. This observation we made is patent-pending.

**EDOCS for Comment Spammer Detection (Chapter 4).** Detecting comment spammers that use comment threads on social media platforms to post spam content is an important problem because spam comment messages have become prevalent [3] and dangerous, with some containing links to malware sites [95]. The majority of the existing techniques consider each comment message independently and attempt to determine if it is spam or not by examining the properties of the comment and its sender [118, 3, 95, 151, 50].

We made the novel observation that comment spammers tend to be lazy and put limited *effort* towards preparing and disseminating their comments, therefore it might be possible to detect the comment spammers if we can quantify the effort scores of the social media users (i.e., the users with low effort scores are expected to be spammers). For instance, we observed that some spammers recycle the comment messages and share the same IP addresses with other spammers, as each message is time-consuming to craft and obtaining unique IP addresses is costly. Assuming that the comment

messages and the IP addresses are the two effort-requiring resources, graphs enable us to capture the relationships between the users and these resources, denoting which user posted a particular comment message and had a specific IP address (see Figure 2 for an example). By doing so, we differ from the existing techniques as we consider all the comment messages in relation to each other in the broader context. Our EDOCS algorithm leverages a graph that represents such effort-related relationships, on which it performs message propagation to quantify the effort scores of the users, and it then flags the users with low effort scores as spammers.

EDOCS detected comment spammers across over 197 thousand users accurately with 95% true positive rate at 3% false positive rate as well as preemptively (i.e., it detected spammers early on), and it outperformed the existing technique used by Yahoo (exact performance details proprietary). EDOCS is patent-pending, has been integrated into Yahoo's anti-abuse technology for their social media platforms, and guards multiple online communities from comment spammers.

### 1.1.2 Part II: Graph-induced Behavior Characterization

In the second part of the thesis, we derive new insights and knowledge that characterize certain behavior of the entities (e.g., activity of smoking or drinking abstainers in an addiction cessation social media community) using statistical and algorithmic techniques that incorporate information from our graphs (e.g., network features extracted from a graph that reflects access to social support among the abstainers in the community, which is known to help individuals fight addiction urges [153, 72]) as well as other useful information about the entities that might be captured externally (e.g., linguistic cues gleaned from the abstainers' posts and comments in the community). Behavior characterization is important because it is an essential first step for analytical tasks such as forecasting (i.e., estimating the likelihood of future events based on the past behavior) and anomaly detection (i.e., revealing activities

that deviate from the behaviors of the majority). As a forecasting-related example, by characterizing the behaviors of short-term and long-term smoking or drinking abstainers from social media, we could design early warning systems that analyze the activities of the abstainers on the social media platform and engage appropriately if a long-term abstainer starts to exhibit the characteristics of the short-term abstainers, as the latter abstainers are more vulnerable to a relapse. In this part of the thesis, we describe several graph-induced behavior characterizations.

**Characterizing Smoking/Drinking Abstinence and Relapse from Social Media (Chapters 5 and 6).** Alcohol and tobacco are among the top causes of preventable deaths in the United States [120]. Achieving long-term abstinence of tobacco or alcohol is difficult [175]—most abstainers are known to relapse within one to three months of cessation. Prior work examining addiction behavior manifested on social media investigates mainly the role of linguistic attributes in characterizing health challenges related to addiction [124, 110]. Also, these pieces of research use crowdsourcing to obtain information about the abstinence status of the individuals. However, simply looking at social media posts may not always allow third-party judges to reliably capture abstinence status.

In our work, which consists of two parts, we focused on two prominent smoking and drinking cessation communities on the social media site Reddit: StopSmoking and StopDrinking. These communities are identified as "self-improvement communities" on Reddit and are geared toward providing support and motivation to smoking and drinking addiction sufferers. A unique aspect of these communities is that they allow the users to acquire "badges" (see Figure 3). Badges are a mechanism by which the users can self-report the duration of their abstinence. We collected data on the users' badges, posts, comments, and associated metadata from these communities, and developed statistical models to analyze the role of social media language,


Figure 3: Screenshots from the StopSmoking and StopDrinking subreddits, showing example post topics and abstinence badges. The badge icon contains the abstinence stage (e.g., star-shaped smiley face for "one year and beyond"), while the actual number of days of abstinence is reported next to it (e.g., 365 days). The usernames are blurred for anonymity.

interactions, and engagement in characterizing smoking/drinking abstinence and relapse. Addiction literature indicates social support to act as an important mediator of stress during smoking/drinking urges [153, 72]. In this context, graphs enable us to capture the interactions and engagement between the users, which reflect access to social support. Specifically, our models leverage a graph that represents which user provides social support to whom by writing comments on their posts in the communities. In summary, through our work, we extend the existing body of research by using self-reported abstinence information on smoking and drinking, and examining the additional role of interaction and engagement in characterizing these addiction-related health challenges.

The first part of our work (Chapter 5) focuses on characterizing abstinence from smoking and drinking. We used the badges of 1,168 users to construct ground truth information on short-term (<40 days) and long-term (>one year) abstainers, and we

formulated and identified the key linguistic and interaction characteristics of these abstainers based on activity in the communities spanning eight years, from 2006 to 2014. We developed supervised learning-based statistical models based on these characteristics to distinguish long-term abstinence from short-term abstinence with over 85% accuracy. We found linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence.

The second part of our work (Chapter 6) focuses on characterizing relapse to smoking and drinking. Here, we used longitudinal data on the badges of 5,991 users to determine their abstinence or relapse status, and we formulated and identified the key engagement and linguistic characteristics of the abstainers and relapsers based on activity in the communities spanning almost nine years, from 2006 to 2015. We developed a robust statistical methodology based on survival analysis to examine how participation in the communities and the characteristics above relate to the risk of relapse. Our results show that although participation in the communities is not linked to high likelihood of smoking/drinking abstinence during the one/two months post-cessation, it shows a stable trend of heightened chance of abstinence beyond three years, suggesting the efficacy of the communities in preventing relapse in the long term. Furthermore, we found positive affect and increased engagement to be predictors of abstinence.

The two parts of our work differ from each other in terms of the problem statement, the statistical method, and the dataset as follows. (1) The first part focuses on characterizing attributes of short-term and long-term abstinence from smoking/drinking. The second part focuses on modeling relapse events self-reported by individuals, and how they, collectively, might indicate the effectiveness of the communities in preventing relapse. (2) The first part uses a supervised learning-based statistical technique. The second part identifies the limitations of such supervised learning techniques in

analyzing relapse events, and employs techniques from the survival analysis literature. (3) The first part considers a dataset with one badge per user. The second part expands this dataset with a unique method to obtain daily badges, and considers a dataset with multiple badges per user to determine the relapse events of the users.

**Analysis of Trading Behaviors of Company Insiders (Chapter 7).** The insiders of a company are corporate officers, directors, or beneficial owners who own more than 10% of the company's stock. While the insiders can legally trade their companies' stock in financial markets, some insiders exploit their roles and use *non-public* information about their companies as a basis for trade. This is called illegal insider trading and it is actively prosecuted by the Securities and Exchange Commission (SEC). To monitor trades by the insiders, SEC requires these trades to be disclosed via a form called *Form 4*. To the best of our knowledge, very little published research is available that uses computational techniques to help financial regulators and policymakers better understand the dynamics behind how the insiders trade.

We performed the first academic, large-scale exploratory study of the complete Form 4 filings from SEC, and made surprising and counterintuitive discoveries. We analyzed over 12 million transactions by around 370 thousand insiders spanning years 1986 to 2012, the largest reported in academia. Our analysis consists of two major components. The first explores the trading behaviors of the insiders from a temporal perspective. By analyzing the time series of the transactions, we discovered distinctive temporal patterns in the insiders' trades that may be explained by government regulations, corporate policies, and macroeconomic factors. For instance, we determined that a significant portion of the insiders makes short-swing profits (i.e., profit resulting from a combined purchase and sale, or sale and purchase, of the company's stock within a 6-month period) despite the existence of a rule designed to prevent short-swing trading.

The other main component of our analysis explores the trading behaviors of the insiders from a graph-based perspective. Specifically, it focuses on the insiders who consistently trade on similar dates, and therefore, might be sharing nonpublic inside information with each other. Graphs enable us to capture such relationships between all the insiders in the broader context. By constructing insider networks that represent these relationships and studying the characteristics of the networks, we found strong evidence that insiders form small clusters in which trade-related information might propagate both vertically (between higher-level and lower-level insiders) and horizontally (among lower-level insiders).

We believe this work could form the basis of novel tools for financial regulators and policymakers to detect suspicious trades based on our characterization of how the insiders trade. The results of this work were presented to SEC.

## 1.2 Scope of Thesis

In this thesis, our overarching goal is to help solve large-scale societal problems; in doing so, we take a graph-based perspective such that we represent the relationships between the entities central to the problems as well as information about the entities in the form of graphs (with the entities as the nodes, the relationships between the entities as the edges, and information about the entities as the node or edge attributes), based on which we design and develop algorithms and models that contribute towards solving these problems. Table 1 provides an overview of the graphs we consider in our work. We harness a variety of graphs with different semantics from the security, healthcare, and finance domains. We deal with different types of graphs ranging from static to dynamic, unipartite to bipartite, undirected to directed, and unweighted to weighted graphs. We note that our graph mining algorithms and statistical models are designed and developed for these particular types of graphs. As such, other types of graphs, such as probabilistic graphs used to model uncertain relationships

Table 1: Overview of the graphs analyzed in this thesis. We consider a variety of graphs from the security, healthcare, and finance domains.

| Graph | Domain | Type | Semantics | Nodes | Edges |
|---|---|---|---|---|---|
| File co-occurrence (Chapter 2) | Security | Static, bipartite, undirected, unweighted | Nodes represent files and buckets containing co-occurring files. Edges denote which file appears on which bucket, hence they indirectly capture the co-occurrence relationships between the files. | 6M | 19.1M |
| Machine-file (Chapter 3) | Security | Dynamic, bipartite, undirected, unweighted | Nodes are files and machines. Time-stamped edges represent when a particular file is observed on a particular machine. | 627K | 3.3M |
| Effort on social media (Chapter 4) | Security | Static, bipartite, undirected, unweighted | Nodes represent social media users and the effort-requiring resources of comment messages and IP addresses. Edges denote which user posted a particular comment message and had a particular IP address. | 1.4M | 1.6M |
| Social support among abstainers (Chapters 5 and 6) | Healthcare | Static, unipartite, directed, weighted | Nodes represent social media users who are smoking/drinking abstainers. Edges denote which user provides social support to whom by writing comments on their posts on the social media platform. Edge weights indicate the extent of support provided or received. | 5.6K | 47K |
| Insider collaboration (Chapter 7) | Finance | Static, unipartite, undirected, weighted | Nodes represent insiders. Edges denote the likely collaborations between the insiders who consistently trade on similar dates. Edge weights indicate how similar insiders' timings of their trades are. | 1.6K | 2.6K |

among the entities and constrained graphs typical in operations research where they represent capacity or flow constraints between the entities, and the application of our algorithms and models to such graphs are beyond the scope of this thesis.

## 1.3 Thesis Statement

*Large-scale societal problems in diverse domains such as security, healthcare, and finance can be addressed from a graph-based perspective via propagation-based algorithms and by characterizing the key behaviors in these domains.*

## 1.4 Research Contributions and Impacts

Our research contributes in multiple facets and has made the following impacts to society.

**New Observations:**

- We made the novel observation that some files tend to appear together on people's computers and these co-occurrence relationships can be exploited to detect malware based on the guilt-by-association principle (i.e., an unknown file that consistently co-occurs with the malicious files is deemed to be malicious).

- We made the novel observation that leveraging the smaller snapshots of a machine-file graph generated from the intervals determined by the ADAGE algorithm can enable us to detect malware more accurately in comparison to using the final, full graph that includes all the machine-file relationships. We validated our observation with an extensive case study over 574 thousand files, achieving an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots in comparison to 43% true positive rate at 0.01% false positive rate with the final graph. This observation we made is patent-pending.

- We made the novel observation that comment spammers tend to be lazy and put limited effort towards preparing and disseminating their comments, therefore it might be possible to detect the comment spammers if we can quantify the effort scores of the social media users (i.e., the users with low effort scores are expected to be spammers).

**New Algorithms:**

- Our AESOP algorithm for malware detection leverages the co-occurrence relationships between the files. AESOP detected malware across over 43 million files

both more accurately (achieving 99.61% true positive rate at 0.01% false positive rate vs. 76.74% true positive rate at 0.01% false positive rate) and sooner (flagging them at least one week sooner) than the state-of-the-art technique [38].

- Our EDOCS algorithm for comment spammer detection quantifies the effort scores of the social media users. EDOCS detected comment spammers across over 197 thousand users accurately with 95% true positive rate at 3% false positive rate as well as preemptively (i.e., it detected spammers early on), and it outperformed the existing technique used by Yahoo (exact performance details proprietary).

**New Characterization-based Insights and Knowledge:**

- We are among the first to understand the smoking/drinking abstinence and relapse experiences of individuals from social media, and provide quantitative insights into evaluating the effectiveness of social media support communities in promoting cessation. By leveraging self-reported abstinence information, we developed statistical models to analyze the role of social media language, interactions, and engagement in characterizing smoking/drinking abstinence and relapse. As an example, we found linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence. Based on participation to the communities we study, we determined that individuals who continue to abstain beyond three years tend to maintain high likelihood of sustained abstinence, suggesting the efficacy of the communities in preventing relapse in the long term. We also found positive affect and increased engagement to be predictors of abstinence.

- We performed the first academic, large-scale exploratory study of the complete insider filings from SEC, and made surprising and counterintuitive discoveries.

As an example, by analyzing the time series of the transactions, we determined that a significant portion of the insiders makes short-swing profits (i.e., profit resulting from a combined purchase and sale, or sale and purchase, of the company's stock within a 6-month period) despite the existence of a rule designed to prevent short-swing trading. Also, in our graph-based analysis, we found strong evidence that insiders form small clusters in which trade-related information might propagate both vertically (between higher-level and lower-level insiders) and horizontally (among lower-level insiders). The results of this work were presented to SEC.

**Impact:**

- Our AESOP algorithm is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.
- Our EDOCS algorithm is patent-pending, has been integrated into Yahoo's anti-abuse technology for their social media platforms, and guards multiple online communities from comment spammers.

# CHAPTER II

# AESOP: LARGE-SCALE MALWARE DETECTION VIA GUILT-BY-ASSOCIATION

Detecting malware lurking in people's computers is an important problem because cyber-attacks involving malware have been causing great damage to individuals, organizations, and governments. The majority of the existing techniques either consider each file independently and check if it fits existing profiles of known malware (see [92] for a survey), or leverage the machine-file relationships, denoting which file appears on which machine [38].

We made the novel observation that some files tend to appear together on people's computers (e.g., multiple files used by the same software) and these co-occurrence relationships can be exploited to detect malware based on the guilt-by-association principle. The idea is that an unknown file that consistently co-occurs with the malicious files might also be malicious, as it might be needed by the latter files to perform certain actions (e.g., communicating with the command and control server). Graphs enable us to capture the co-occurrence relationships between the files. This differentiates us from the existing techniques as they do not consider these relationships. Our Aesop algorithm leverages a graph that represents such co-occurrence relationships, on which it performs large-scale inference by propagating goodness scores from the malicious and benign files to the unknown files in the graph to determine the nature of the unknown files based on the guilt-by-association principle. As an example, Aesop would assign a low goodness score to an unknown file that consistently co-occurs with the malicious files as it would receive low goodness scores from the neighboring

Material adapted from work appeared at ACM KDD 2014 [167].

malicious files in the graph.

Aesop detected malware across over 43.3 million files both more accurately (achieving 99.61% true positive rate at 0.01% false positive rate vs. 76.74% true positive rate at 0.01% false positive rate) and sooner (flagging them at least one week sooner) than the state-of-the-art technique [38]. Aesop is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.

## 2.1 Introduction

Protection against novel malware attacks, also known as 0-day malware, is becoming increasingly important as the cost of these attacks increases. For individuals, the dollars and cents cost is rising due to the increasing prevalence of financial fraud and the increasing viciousness of malware, such as the CryptoLocker ransomware program that encrypts personal data files and holds them for a ransom of 300 dollars [28]. Emotional and professional costs can be much higher, as when attacks result in the loss of privacy. The situation is arguably worse for governments and businesses, which find themselves under siege by well-funded attackers that routinely create devastating financial losses, and perhaps even more impactful losses of intellectual property and operational secrets [163].

Computer security providers recognize the need to respond with better protection against novel threats. The goal of these 0-day threat protections is to limit the malware's window of effectiveness, so that malicious files are detected as soon as possible after their first appearance. Another critical measure of success is a vanishingly small false positive rate, as labeling a benign file as malicious can have devastating consequences, particularly if it is a popular file or one that is essential to the stability of the system, as in the case of operating system and driver files.

We present Aesop (Figure 4), a novel approach to detecting malicious executable

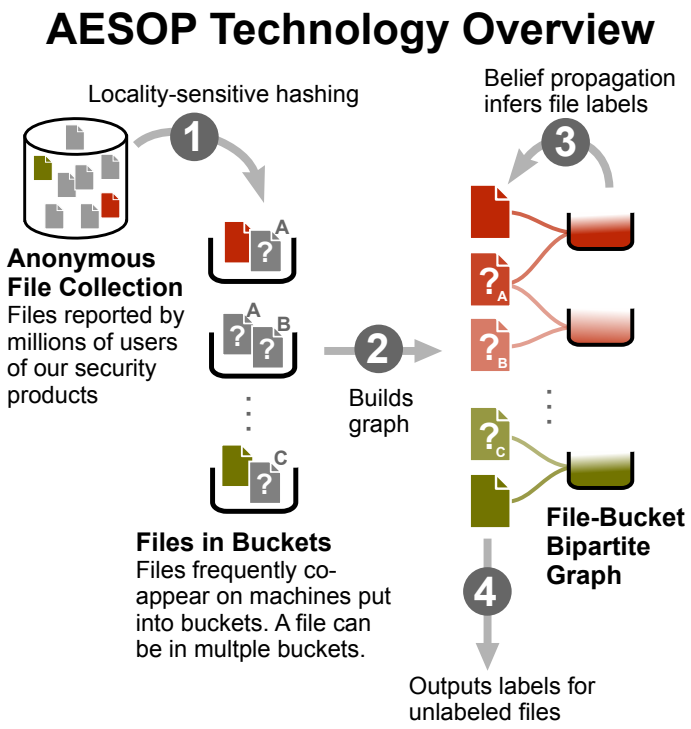



Figure 4: Overview of the AESOP algorithm.

files by applying the well-known aphorism that "a person is known by the company he or she keeps," and in our case, a file's goodness may be judged by the other files that often appear with it on users' machines based on the guilt-by-association principle (i.e., an unknown file that consistently co-occurs with the malicious files is deemed to be malicious). In contrast with most other malware detection techniques, we set individuals files into a broader context and infer unlabeled files' reputation (or goodness) by analyzing their relations with labeled peers.

AESOP is not the first attempt to detect malware by establishing file reputation scores. A representative work in this space is Polonium [38], which leverages the insight that some computer users have poor internet hygiene in that they attract many more malicious files than users that follow security best practices. Polonium constructs a bipartite graph between files and machines, in which a file-machine edge represents the existence of a particular file on a particular machine. This approach

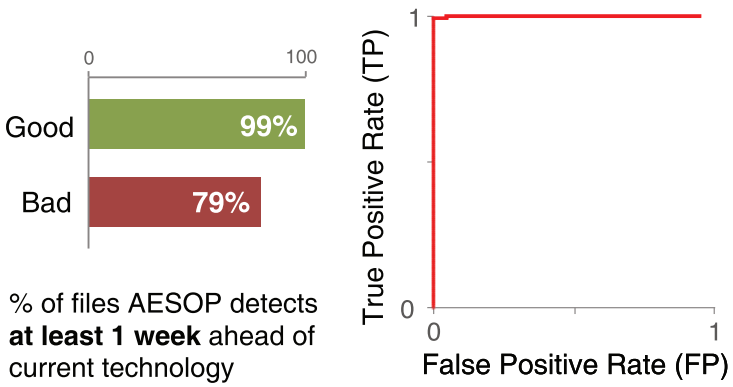



Figure 5: *Left*: 99% of the known good files and 79% of known bad files detected by AESOP were labeled *at least 1 week* ahead of Symantec's current technology. *Right*: AESOP achieves almost perfect detection for malware, with few false alarms (0.9961 TP rate at 0.0001 FP rate).

proved to be successful; Symantec has deployed Polonium. However, Polonium misses many malicious files as it can only observe malware's file-to-file relationships indirectly through the lens of low-hygiene machines. By contrast, AESOP leverages a graph that enables to directly capture file-to-file affinity and it can therefore identify malicious files that co-occur with one another, even when they do not appear on heavily infected machines. As we shall demonstrate, AESOP is able to detect many malicious files over a week before they are labeled by Symantec's existing Polonium-based technology, with a 0.0001 false positive rate (see Figure 5).

We leverage Symantec's Norton Community Watch data, the most important elements of which are unique file and machine identifiers. File identifiers are SHA-256 or MD5 cryptographic hash values that are computed over the file's raw bytes. Symantec's proxy for a true machine identifier is based on the serial number of Norton security products, which is an adequate but imperfect fit because product re-installation on a single machine may result in a serial number change, and a single serial number can be carried from one machine to another. The scale of this dataset is impressive, comprising 119 million machines and 10.1 billion files.

Our work makes the following contributions:

- We formulate the malware detection problem as a large-scale graph mining and inference problem, where our goal is to identify an unknown file's relations with other files so that we can establish guilt or innocence by its association with files that are known to be benign or malicious.

- We present the Aesop algorithm that leverages locality-sensitive hashing to efficiently compute file similarity values to construct a file-relation graph for inferring file goodness based on belief propagation.

- Aesop achieved early detection of 99% of benign files and 79% of malicious files that remained unlabeled by Symantec for over a week before they were eventually labeled, with exceptionally low error rates (see Figure 5).

- Aesop is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.

The remainder of this chapter proceeds as follows. We first survey related work. We then describe the notation we will use throughout the chapter. Afterwards, we proceed to a description of Aesop and its various components, followed by the experiments we conducted to demonstrate its effectiveness. Finally, we end by presenting our conclusions.

## 2.2 Prior Work and Our Differences

The exceptional depth and breadth of related work in the malware detection space is a testament to the importance and difficulty of the problem. Idika and Mathur [92] survey 45 different malware detection techniques that are divided into two categories: (i) *signature-based detection*, which detects malware that fits certain profiles (or signatures), and (ii) *anomaly-based detection*, which detects malware's deviation from some presumed "normal" behavior. Broadly, these techniques consider each file individually and assume that the files are independent of one another. In contrast, Aesop

considers the files in relation to each other in the broader context by leveraging a graph that captures the co-occurrence relationships between the files. More closely related to AESOP'S malware detection approach are reputation-based techniques and techniques that exploit similarities between files for detection.

There exist reputation systems that have been developed to address security-related problems, such as reputation scoring for IP addresses [13] and DNS entries [14, 25]. The most closely related work to ours is Polonium [38], one of Symantec's current malware detection technologies. Polonium also takes a graph-based approach to infer file reputation, however with important differences. First, AESOP infers files' goodness by directly considering file-to-file relations, which is different than Polonium's indirect approach of analyzing file-to-machine relations. Second, Polonium was not designed to pick out related files that frequently co-appear, while AESOP does; leveraging this relational information, AESOP is able to accurately label many files at least one week before the current technologies (as discussed in Section 2.5).

As the number of unique malware executable files has exploded due to their use of polymorphic and metamorphic techniques, security researchers are increasingly turning to techniques that identify clusters of related malware files rather than attempt to detect files individually. Symantec's MutantX-S [87] system clusters executables according to their static and dynamic properties. This approach works with low-level malware features such as sequences of machine-language opcodes, making it largely orthogonal to our approach. Similarly, Hu et al. [88] propose system called SMIT that implements a malware conviction approach which casts the problem of determining if a new binary sample is malicious into one of locating the samples nearest neighbors in the malware database. Their approach converts each malware program into its function-call graph representation derived from the malware's source code, and performs nearest neighbor search based on this graph representation using an approximate graph-edit distance metric for improved scalability. As this approach

also leverages low-level malware features, it is orthogonal to ours.

Karampatziakis et al. [96] use file placement as the primary component of its malware detection technique, by leveraging unique properties of file containers that would not generalize to machines, such as the idea that the presence of any malicious file in an archive is sufficient evidence to label all files in that archive as malicious. In addition, rather than performing inference as AESOP does with belief propagation, their logistic regression classifier only looks at a file's immediate neighbors in the archive to which it belongs.

In summary, not only does AESOP demonstrate the independent value of calculating file-to-file similarity scores, it also provides an algorithm that addresses scalability problems while achieving impressive results compared to the existing techniques. Furthermore, AESOP's belief propagation approach provides a reputation-based system with nuanced scores that are ideally suited for integrating and improving existing malware detection technologies.

## 2.3 Background

### 2.3.1 Problem Formulation

We consider a dataset $D$ consisting of records of the form $< f, M_f >$, where $f$ is a file and $M_f$ is the set of machines that file $f$ appears on, i.e., assuming that $M$ is the set of all the machines, $M_f = \{m_1, m_2, ...\}$ where $m_i \in M$. Each file is either *labeled* or *unlabeled*. The possible labels for a labeled file are *good* and *bad*, indicating the nature of the file, i.e., whether it is purely benign or malicious, respectively. We refer to a labeled file with the label *good* as a *good file* and with the label *bad* as a *bad file*. The good and bad files comprise the ground-truth set. Our informal high-level problem statement can be stated as follows: *Given a dataset as defined above, assign a label (i.e., good or bad) to unlabeled files based on their co-occurrence with the labeled files.* Table 2 lists the symbols used throughout the chapter.

Table 2: Main symbols used throughout the chapter. LSH stands for locality-sensitive hashing.

| Symbol | Meaning |
|---|---|
| $f$ | File (a.k.a. executable, software, application) |
| $m$ | Machine (or computer) |
| $M$ | Set of all machines; $m \in M$ |
| $M_f$ | Set of machines that file $f$ appears on |
| $D$ | Input dataset; records consist of $< f, M_f >$ |
| $J(M_{f_i}, M_{f_j})$ | Jaccard similarity between $M_{f_i}$ and $M_{f_j}$ |
| $h$ | Random permutation function in MinHashing |
| $b$ | Number of bands in LSH |
| $r$ | Number of MinHash values in each band in LSH |
| $n$ | Total number of MinHash values in LSH; $n = b \times r$ |
| $s$ | Jaccard similarity between a pair of files |
| $TP$ | True positive; a malware instance correctly identified as bad |
| $FP$ | False positive; a benign file incorrectly identified as bad |

### 2.3.2 File Co-occurence Strength

We define the *strength* of co-occurrence between files $f_i$ and $f_j$ based on the overlap between sets $M_{f_i}$ and $M_{f_j}$, and employ the Jaccard similarity measure given by the formula $J(M_{f_i}, M_{f_j}) = \frac{|M_{f_i} \cap M_{f_j}|}{|M_{f_i} \cup M_{f_j}|}$. This measure takes a value between 0 and 1 (inclusive); the former indicates a nonexistent co-occurrence relationship and the latter indicates a perfect co-occurrence relationship. Based on domain knowledge, we assume that if $J(M_{f_i}, M_{f_j}) > 0.5$, this indicates a *strong* co-occurrence between files $f_i$ and $f_j$. AESOP leverages the strong co-occurrence relationships between the files to label them. To quantify these relationships, AESOP uses Jaccard similarity because it can be efficiently computed and well-approximated for large-scale datasets through locality-sensitive hashing, which we describe below.

## 2.4 Proposed Method: The AESOP Algorithm

In this section, we describe the design rationale behind AESOP so that it can scale to a large number of files and machines. Figure 4 provides an overview of the AESOP approach. We begin by describing our use of MinHashing, which allows us to

Table 3: An example dataset $D$ and random permutation function $h$.

| File | Set of machines containing the file |
|---|---|
| $f_1$ | $M_{f_1} = \{m_2, m_4, m_5, m_8\}$ |
| $f_2$ | $M_{f_2} = \{m_3, m_5, m_7\}$ |
| $f_3$ | $M_{f_3} = \{m_1, m_3, m_5, m_6, m_7\}$ |

$$h(m_1) = 3 \quad h(m_2) = 6 \quad h(m_3) = 2 \quad h(m_4) = 4$$
$$h(m_5) = 8 \quad h(m_6) = 7 \quad h(m_7) = 1 \quad h(m_8) = 5$$

approximate the Jaccard similarity between two sets efficiently. Next, we explain our adaptation of locality-sensitive hashing to efficiently identify peer-groups of co-occurring files. Finally, we describe how we propagate information from labeled files to their unlabeled peers using belief propagation.

### 2.4.1 MinHashing for Co-occurence Strength Estimation

It is not efficient to compute the Jaccard similarity between large sets due to the expensive set intersection and union operations involved. *MinHashing* [31], which is short for *Minwise Independent Permutation Hashing*, is a popular technique to efficiently estimate the Jaccard similarity between two sets. MinHashing has been proven to work well for large-scale real-world applications, such as detecting duplicate images [42] and clustering users on Google news [53]. We will explain MinHashing using dataset $D$ in Table 3 as a running example. MinHashing randomly reorders the machines in $M$ using a bijective function $h$ that maps the machines in $M$ to the set $\{1, ..., |M|\}$ in a random fashion. We call function $h$ a *random permutation function*. An example function $h$ for $M = \{m_1, ..., m_8\}$ is given in Table 3. Notice that if we rearrange the machines in $M_f \in D$ in ascending order of the machines' values retrieved from function $h$, we obtain a random permutation of $M_f$, which we refer to as $M_f^h$. For instance, $M_{f_2}$ in Table 3 is permuted as $M_{f_2}^h = (m_7, m_3, m_5)$ since $h(m_7) = 1 < h(m_3) = 2 < h(m_5) = 8$. The MinHash value of $M_f$ under function $h$, which we refer to as $h_{min}(M_f)$, is defined as $h_{min}(M_f) = \arg\min_{m_i \in M_f} h(m_i)$.

Informally, $h_{min}(M_f)$ is the first element of $M_f^h$. For instance, $h_{min}(M_{f_1}) = m_4$ in Table 3.

The key property of MinHashing is that the probability of the MinHash values of two sets being equal is equal to the Jaccard similarity between the sets. Formally, $\Pr(h_{min}(M_{f_i}) = h_{min}(M_{f_j})) = J(M_{f_i}, M_{f_j})$ (see Cohen et al. [46] or Rajaraman and Ullman [142] for a proof). As an example, in Table 3, $h_{min}(M_{f_1}) = m_4$, $h_{min}(M_{f_2}) = m_7$, and $J(M_{f_1}, M_{f_2}) = 0.17$. This property provides a probabilistic estimation of the Jaccard similarity.

### 2.4.2 Clustering Co-occurring Files

Despite the use of MinHashing, for large datasets the number of file pairs that need to be considered to capture the co-occurrence relationships between all the files remains very large. It is also possible that two sets may not receive the same MinHash value but in fact have a high Jaccard similarity, or vice versa. Hence, a single MinHash value is typically not sufficient to deduce whether two sets have a high Jaccard similarity. Locality-sensitive hashing (LSH) addresses these points; it allows us to identify peer-groups of co-occurring files efficiently (with one pass over the dataset) and accurately.

LSH is a technique for approximate clustering and near-neighbor search in high dimensional spaces [93, 74]. Its main idea is to use multiple hash functions to map items into buckets such that similar items are more likely to be hashed to the same bucket. LSH uses *locality-sensitive function families* to achieve this goal.[1] At a high-level, each individual function in a locality-sensitive function family should be able to provide lower and upper bounds on the probability of whether two items with a pairwise similarity (or distance) in a particular interval will receive the same hash value from the function. Therefore, locality-sensitive function families are defined for particular similarity or distance measures, such as Hamming distance [74], $L_p$

[1] A function family is a group of functions that share certain characteristics.

Table 4: Hypothetical inputs and outputs for locality-sensitive hashing (LSH). The inputs are the MinHash values for each file. The outputs are buckets containing files. This LSH scheme uses three bands, each consisting of two MinHash values.

| | $h_{min}$ | $M_{f_4}$ | $M_{f_5}$ | $M_{f_6}$ | Buckets |
|---|---|---|---|---|---|
| Band 1 | $h^1_{min}$ | $m_1$ | $m_1$ | $m_1$ | $[f_4, f_5]$ $[f_6]$ |
| | $h^2_{min}$ | $m_1$ | $m_1$ | $m_2$ | |
| Band 2 | $h^3_{min}$ | $m_5$ | $m_5$ | $m_3$ | $[f_4, f_5]$ $[f_6]$ |
| | $h^4_{min}$ | $m_8$ | $m_8$ | $m_4$ | |
| Band 3 | $h^5_{min}$ | $m_1$ | $m_7$ | $m_7$ | $[f_4][f_5, f_6]$ |
| | $h^6_{min}$ | $m_6$ | $m_6$ | $m_6$ | |

norms [74, 54], and earth mover's distance [37]. The random permutation functions used in MinHashing (see Section 2.4.1) form a locality-sensitive function family for the Jaccard similarity measure [46].

A useful property of the locality-sensitive function families is that they can be amplified by combining values returned from multiple functions via logical AND and/or OR [142]. In our context, this means that we can compute $n$ MinHash values (using $n$ different random permutation functions) for each $M_f \in D$. Subsequently, these $n$ MinHash values can be combined in multiple ways. One way is to partition $n$ MinHash values into $b$ bands, each consisting of $r$ values, such that $n = b \times r$.

As an example, consider Table 4, which lists six MinHash values for $M_{f_4}$, $M_{f_5}$, and $M_{f_6}$, obtained from six different random permutation functions $h^1, ..., h^6$. These six MinHash values are partitioned into three bands, each consisting of two values. For instance, $M_{f_4}$'s MinHash values for Band 2 are $(m_5, m_8)$. Assume that we use a cryptographic hash function, such as SHA-256, to assign files to buckets based on their MinHash values in a band. Then, the files will appear in the same bucket if all of their $r$ MinHash values in that band are the same. For instance, in Band 2, files $f_4$ and $f_5$ appear in the same bucket because their MinHash values for this band, denoted by $h^3_{min}$ and $h^4_{min}$, are both $(m_5, m_8)$, whereas file $f_6$ appears in a separate bucket because its MinHash values are $(m_3, m_4)$. In this scheme, the files have $b$ chances of appearing

in the same bucket. This type of amplification is called an AND-construction with $r$ rows followed by an OR-construction with $b$ bands [142]. This is because files will hash to the same bucket at least once if *all* of their $r$ MinHash values (logical AND operation) in *any* of the $b$ bands are the same (logical OR operation).

Based on this scheme, we can derive the probability that files $f_i$ and $f_j$ will appear in at least one bucket given their true Jaccard similarity, $J(M_{f_i}, M_{f_j}) = s$. As discussed in Section 2.4.1, the probability that one MinHash value of $M_{f_i}$ and $M_{f_j}$ being equal is $s$. Therefore, the probability of $r$ MinHash values of $M_{f_i}$ and $M_{f_j}$ being the same is $s^r$. Notice that $s^r$ is the probability that files $f_i$ and $f_j$ will hash to the same bucket in a particular band. Therefore, the probability that files $f_i$ and $f_j$ will not hash to the same bucket in a particular band is $1 - s^r$. Then, the probability that files $f_i$ and $f_j$ will not hash to the same bucket in all of the $b$ bands is $(1 - s^r)^b$. Finally, the probability that files $f_i$ and $f_j$ will hash to the same bucket in at least one of the $b$ bands is $1 - (1 - s^r)^b$. In Section 2.5.1, we discuss how we set the values of the LSH parameters based on this probability.

### 2.4.3 Labeling Files Based on Co-occurrence

The output of LSH on a dataset is multiple bands, each consisting of a varying number of buckets that contain co-occurring labeled and unlabeled files. A file appears at most once in a band, inside one of the buckets of the band. Across different bands, the file might appear with a different set of files. For instance, in Table 4, file $f_5$ appears with file $f_4$ in the first two bands and with file $f_6$ in the last band. In this section, we discuss how we combine the buckets from different bands into a single structure and assign labels to the unlabeled files using it.

**Unipartite File Graph.** Graphs provide a powerful representation of relationships between objects, hence one approach to combine the buckets could be to construct an undirected unipartite file graph by considering every pair of files in the buckets. In

this graph, the files are represented as nodes and they are connected with an edge if they appear in the same bucket. This graph can then be used in a way that goodness and badness information is propagated from the labeled files to the unlabeled files in the graph. Our preliminary analyses showed that constructing such a unipartite file graph is not feasible. The main reason is that some buckets contain a large number of files, which contribute dense subgraphs to the graph. In turn, the number of edges in the graph increases dramatically, making it infeasible to operate on the graph. This is most likely a property of the domain; there are intrinsic dependency relationships between files (e.g., the files under the "\Windows\System32" folder in the Windows operating systems).

**Bipartite File-Bucket Graph.** For improved scalability, AESOP operates on an undirected bipartite file-bucket graph, which we refer to as a *file-relation graph*. In this graph, the files and buckets are represented as nodes, and there is an edge connecting a file to a bucket if the file appears in that bucket. Notice that the number of edges that would be included to the unipartite file graph from a bucket of $N$ co-occurring files is $\mathcal{O}(N^2)$; in contrast, the same number is $\mathcal{O}(N)$ for the file-relation graph. The bipartite graph contains more nodes than the unipartite graph due to the additional nodes for the buckets, however this is less of a concern for information propagation purposes, as we will discuss. The file-relation graph captures all the information needed to assign labels to the unlabeled files; its difference from the unipartite file graph is that the files are now indirectly connected through the buckets, therefore goodness information shall be first propagated from the labeled files to the buckets and then from the buckets to the unlabeled files.

**Remarks.** A useful property of the file-relation graph is that it intrinsically captures the notion of a weight between the files. To illustrate this, consider files $f_i$ and $f_j$ with a Jaccard similarity $J(M_{f_i}, M_{f_j}) = s$. If we use a LSH scheme with $b$ bands and

$r$ MinHash values, the probability that files $f_i$ and $f_j$ appear in the same bucket in a band is $s^r$. Then, the number of bands files $f_i$ and $f_j$ appear together in a bucket is a random variable $X$ that follows the Binomial distribution with parameters $b$ and $s^r$, i.e., $X \sim \mathrm{B}(b, s^r)$. Thus, the larger the value of $s$, the more bands in which files $f_i$ and $f_j$ will appear together inside a bucket. In the file-relation graph, this results in a larger number of paths between files $f_i$ and $f_j$ that go through the buckets, thereby allowing files $f_i$ and $f_j$ to influence each other more than the other files do.

Also, after the file-relation graph is constructed, it is possible that some of its connected components consist of one file or only unlabeled files. These components do not contribute to solving the problem of assigning labels to unlabeled files, therefore AESOP excludes them from the graph to retain only the useful information.

**Belief Propagation.** Next, we describe our approach to assign labels to unlabeled files using the file-relation graph. Our goal is to label the nodes corresponding to unlabeled files as *good* or *bad*, along with a measure of confidence. We adapt a probabilistic approach and treat each file as a random variable $X \in \{x_g, x_b\}$, where $x_g$ is the good label and $x_b$ is the bad label. The file's goodness and badness can then be expressed by the probabilities $\Pr(x_g)$ and $\Pr(x_b)$, respectively, such that $\Pr(x_g) + \Pr(x_b) = 1$. Based on this formulation, for an unlabeled file $f_i$, our goal is to determine the marginal probabilities $\Pr(X_{f_i} = x_g)$ and $\Pr(X_{f_i} = x_b)$.

An undirected graph whose nodes are expressed probabilistically as specified above is a pairwise Markov random field (MRF) [101]. The task of inferring the marginal distribution of the nodes in a pairwise MRF is NP-complete [177]. The belief propagation (BP) algorithm [177] is a successful approximation technique for solving this problem. BP has been adapted to various domains, such as image restoration [68] and fraud detection [116]. The algorithm is also scalable; it takes time linear in the number of edges in the graph.

At a high level, BP infers the marginal distribution of a node using some prior knowledge about the node and the messages arriving from the node's neighbors. The idea is to iteratively pass messages between every pair of connected nodes $i$ and $j$. Typically, $m_{ij}(x_k)$ represents the message sent from node $i$ to node $j$, which denotes node $i$'s belief that node $j$ is in state $x_k$. The prior knowledge, or simply the *prior*, for node $i$ is denoted by the node potential function $\phi_i$ that specifies the prior probabilities that node $i$ is in each of the possible states. The message passing procedure stops when the messages converge or a maximum number of iterations is reached. The final, inferred marginal probabilities are called the final beliefs. The symbol $b_i(x_j)$ denotes the final belief that node $i$ is in state $x_j$.

The BP algorithm is carried out as follows in practice. An edge between nodes $i$ and $j$ passes a message towards each direction for each possible state. The order of the transmission can be arbitrary if all the messages are passed in every iteration. The set of beliefs that a node has for each of its neighbors is kept normalized to sum to 1. This prevents any numerical underflow, i.e., a certain belief reaching 0 due to limited precision. A message from node $i$ to its neighbor node $j$ is generated based on node $i$'s neighbors' messages about node $i$. Formally, the message update equation is:

$$m_{ij}(x_k) \leftarrow \sum_{x_\ell \in X} \phi_i(x_\ell)\ \psi_{ij}(x_\ell, x_k)\ \frac{\prod_{p \in N(i)} m_{pi}(x_\ell)}{m_{ji}(x_\ell)} \tag{1}$$

where $N(i)$ is the set of nodes neighboring node $i$, and $\psi_{ij}(x_\ell, x_k)$ is the *edge potential*, which specifies the probability that node $i$ is in state $x_\ell$ and node $j$ is in state $x_k$.

Although BP is not theoretically guaranteed to converge in general graphs, in practice the algorithm usually converges quickly. After the message passing procedure stops and the algorithm ends, the final beliefs are computed as:

$$b_i(x_j) \leftarrow k \times \phi_i(x_j) \times \prod_{p \in N(i)} m_{pi}(x_j) \tag{2}$$

where $k$ is a normalizing constant.

Table 5: Edge potential function indicating that files with similar nature tend to co-occur on the users' machines.

| $\psi_{ij}(x_\ell, x_k)$ | $x_\ell = \text{good}$ | $x_\ell = \text{bad}$ |
|---|---|---|
| $x_k = \text{good}$ | 0.99 | 0.01 |
| $x_k = \text{bad}$ | 0.01 | 0.99 |

We tailor BP to our context as follows. Recall that there are two types of nodes in the file-relation graph: files and buckets. For brevity, we only mention the priors for the good state. We set the priors of the buckets to 0.5. This way, the buckets are initially neutral and are influenced only by the files to which they are connected. We set the priors of the good files to 0.99 and of the bad files to 0.01. We set the priors of the unlabeled files to 0.5 so that they are also initially neutral and their final beliefs are indirectly determined by the labeled files with which they co-occur. We set the edge potential function so that it reflects the guilt-by-association assumption that a good file is more likely to be associated with a bucket consisting of other good files than a bucket consisting of bad files (similar reasoning for the bad files), as shown in Table 5.

### 2.4.4 Time Complexity of AESOP

AESOP has two main components: (i) the clustering component with LSH and (ii) the labeling component with BP. We analyze the time complexity of each component and obtain an overall time complexity.

At a high level, LSH considers each file in dataset $D$ one by one; specifically, it maintains a MinHash value with respect to each permutation function while iterating over the set of machines the file appears on. Assume that we compute a total of $n$ MinHash values for each file, using $n$ different random permutation functions. Also, assume that the dataset contains $|D|$ files, and recall that $M$ denotes the set of all the machines. Then, a file can appear on at most $|M|$ machines. As a result, the time complexity for computing the MinHash values is $\mathcal{O}(|D| \cdot |M| \cdot n)$. The random

permutation functions can be determined in advance by randomly shuffling $M$ using the Fisher-Yates shuffle [70], which takes time linear in the number of elements to be shuffled [152], and obtaining a mapping from the machines to their positions in the permutation. For $n$ random permutation functions, this approach has a time complexity of $\mathcal{O}(|M| \cdot n)$. Also, additional work is needed to form the buckets containing the files as part of LSH, however this involves iterating over the files' MinHash values in each band and, for $b$ bands, has a time complexity of $\mathcal{O}(|D| \cdot b)$. Putting these together, we obtain that the time complexity of the clustering component with LSH is $\mathcal{O}(|D| \cdot |M| \cdot n)$ since $b \leq n$.

The BP algorithm iterates over each edge in the graph a constant amount of times if it is set to run up to a maximum number of iterations [177], which is the case in practice [38]. Assume that $E$ is the set of the edges and $|E|$ is the number of edges in the file-relation graph. Then, the time complexity for the labeling component with BP is $\mathcal{O}(|E|)$.

The overall time complexity of Aesop is therefore $\mathcal{O}(|D| \cdot |M| \cdot n + |E|)$.

## 2.5 Experiments

This section presents an experimental evaluation of Aesop. We measure its effectiveness in detecting labeled benign and malicious files as well as discovering labels for unlabeled files. We conducted our experiments on a 64-bit Linux machine (RedHat Enterprise Linux Server 5.7) with 8 Opteron 2350 quad core processors running at 2.0 GHz, 64GB of RAM, and 100GB disk-quota per user.

### 2.5.1 Setting LSH Parameters

We first discuss how we set the values of the LSH parameters for our experiments. Recall from Section 2.4.2 that LSH has three parameters: $n$, $b$, and $r$, with the constraint that $n = b \times r$. To improve the clustering accuracy, we set $n$ to the largest possible value supported by our computing resources, which was determined to be

100.

Assuming that $n = 100$, we set the values of $b$ and $r$ as follows. Recall again from Section 2.4.2 that the probability that two files $f_i$ and $f_j$ with a Jaccard similarity $J(M_{f_i}, M_{f_j}) = s$ will be sent to the same bucket in at least one of the $b$ bands is $P(s) = 1 - (1 - s^r)^b$. Consider the general case where we want files $f_i$ and $f_j$ to appear together in at least one bucket only if their Jaccard similarity $J(M_{f_i}, M_{f_j}) = s$ is greater than a Jaccard similarity threshold $t$. Then, the goal with LSH is that files $f_i$ and $f_j$ with $J(M_{f_i}, M_{f_j}) = s \leq t$ have a very small $P(s)$ value so that they are unlikely to appear together in a bucket in any of the bands, and files $f_i$ and $f_j$ with $J(M_{f_i}, M_{f_j}) = s > t$ have the largest possible $P(s)$ value so that they are highly likely to appear together in a bucket in at least one of the bands. Formally, we formulate the following problem: *Given a threshold $t$, find the $b$ and $r$ values such that $P(t) \leq 0.01$ and the area under the curve formed by $P(s \in [0, 1])$ is maximal.* Note that $P(s \in [0, 1])$ is a monotonically increasing S-shaped function with any choice of $b$ and $r$ [142], therefore $P(t) \leq 0.01$ ensures that for some $t' < t$, $P(t') \leq 0.01$. The procedure by which we determine the desired $b$ and $r$ values for threshold $t$ is as follows: For any $b$ and $r$ pair such that $n = 100 = b \times r$, (i) test if $P(t) \leq 0.01$, (ii) consider 10,000 Jaccard similarity values equidistant in the range $[0, 1]$ and generate a discrete $P(s \in [0, 1])$ curve by computing their $P(s)$ values, (iii) compute the area under the $P(s \in [0, 1])$ curve using the trapezoidal method [15], and (iv) return the $b$ and $r$ pair that maximizes the area under the curve.

Our procedure returned $b = 10$ and $r = 10$ for $t = 0.5$ that captures the notion of strong co-occurrence between the files (see Section 2.3.2). We considered this combination of $b$ and $r$ values in our experiments.

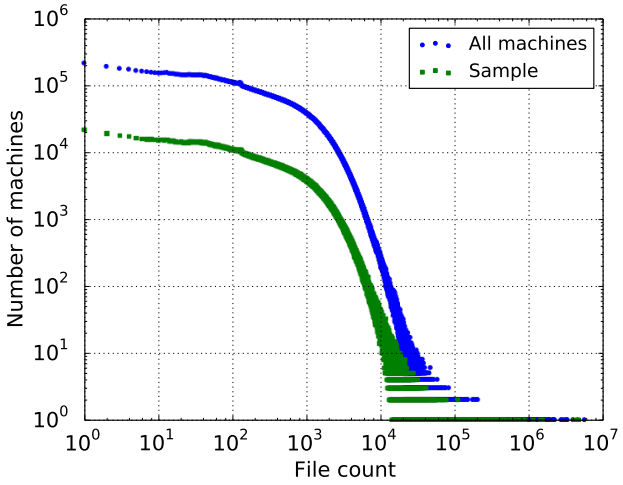



Figure 6: Distributions of the number of machines (vertical axis) with a particular file count (horizontal axis) for the original dataset (higher blue curve with circles) and the sample (lower green curve with rectangles). Our sampling strategy preserves the overall shape of the original distribution.

### 2.5.2 Sampling Norton Community Watch

We leverage Symantec's Norton Community Watch data, the most important elements of which are unique file and machine identifiers. This terabyte-scale dataset contains more than 119 million machines and over 10.1 billion files. Due to the limited disk space budget, we obtained a sample from this data as follows.

Symantec's Worldwide Intelligence Network Environment (WINE) samples and aggregates datasets that Symantec uses in its day-to-day operations to share them with the research community [61]. The WINE sampling scheme selects machines uniformly at random and retrieves any data for the sampled machines from the production systems. Previous work showed that the uniform sampling of the machines is effective in terms of estimating or extrapolating crucial attributes of the original datasets from the samples [136].

Motivated by this result, we employed a similar technique to sample machines from the Norton Community Watch data. The set of files appearing on each sampled machine was retrieved completely. Figure 6 shows the distributions of the number of

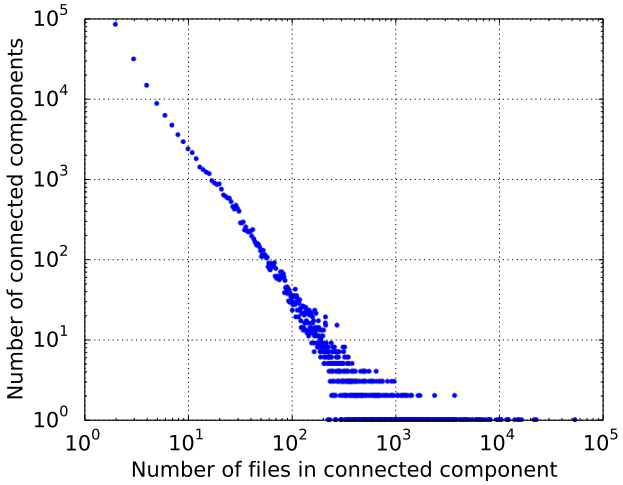



Figure 7: Distribution of the number of connected components (vertical axis) containing a particular number of files (horizontal axis) in the file-relation graph. Smaller components are less likely to contain a mix of good and bad files. The distribution is heavy tailed, indicating that most files appear in small-sized connected components.

machines containing a particular number of files for the original dataset and a 10% sample (i.e., the number of machines in the sample is 10% of the total number of machines in the original dataset). The uniform random sampling approach preserves the overall shape of the original distribution; both distributions are heavy-tailed with few machines containing a large number of files and a large number of machines containing few files.

We obtained the sample on November 6, 2013. It contains 11,939,429 machines and 43,353,581 files, with labels for 7% of the files in the sample, and it occupies 120GB of space on disk. Each file in the sample occurs on at least 5 sampled machines.

### 2.5.3 File-relation Graph

From the sample, AESOP generated a file-relation graph of 6,056,802 nodes and 19,103,825 edges. The graph contains 1,663,506 good files, 47,956 bad files, and 1,085,937 unlabeled files, and 3,259,403 nodes that correspond to buckets. The number of buckets is large because AESOP uses 10 bands in LSH; each band contributes a similar set of files but a distinct set of buckets.

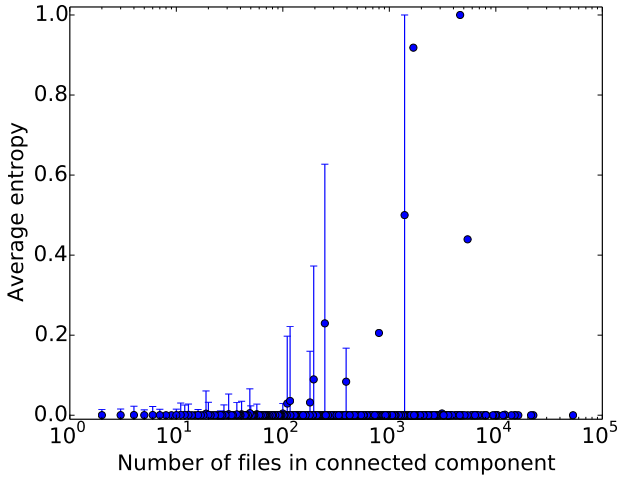



Figure 8: Average entropy for the connected components (vertical axis) containing a particular number of files (horizontal axis) in the file-relation graph. The error bars correspond to one standard deviation. A significant fraction of the connected components have a zero entropy, indicating that they consist of files with identical labels.

### 2.5.4 Sizes of Connected Components

Aesop is expected to perform better if the files form small, disconnected clusters in the file-relation graph. This is because large groups of files are likely to contain a mix of good and bad files that are difficult to classify accurately. The connected components of a graph are its largest clusters, so in Figure 7 we show the graph's distribution of connected component sizes in terms of the number of files they contain. Note that the distribution is heavy tailed, indicating that most files appear in small-sized connected components. The graph's connected components that contain a very large number of files justify our selection of operating on a bipartite file-bucket graph instead of a unipartite file graph (see Section 2.4.3).

### 2.5.5 Purity of Connected Components

It is also important that the file-relation graph's connected components are pure, i.e., they consist of files with identical labels. To test this, we turned to *entropy*, a widely

used measure for determining the uncertainty or irregularity of a system [99]. We computed the entropy of a connected component as $(-\frac{e_g}{e_g+e_b}\log_2\frac{e_g}{e_g+e_b} - \frac{e_b}{e_g+e_b}\log_2\frac{e_b}{e_g+e_b})$, where $e_g$ and $e_b$ are the number of good and bad files in the component, respectively. Note that a smaller entropy denotes a purer connected component. Figure 8 shows the average entropy for the connected components containing a particular number of files. The error bars correspond to one standard deviation. We observe that a significant fraction of the connected components have entropies close to zero, indicating that they are pure regardless of their sizes.

### 2.5.6 Performance Evaluation with Cross-validation

Next, we evaluate the effectiveness of AESOP in detecting benign and malicious files. Our evaluation scheme used 10-fold cross-validation. We treated the files in the test set as unlabeled files by setting their priors for the good state to 0.5. The files in the training set were assigned priors as described in Section 2.4.3. For each fold, we ran the BP component of AESOP for 10 iterations and reported the true positive (TP) rate at a fixed 0.0001 false positive (FP) rate. Recall that, in our context, a TP is a malware instance that is correctly identified as malicious and an FP is a benign file incorrectly identified as malicious.

Figure 9 shows the overall and zoomed-in receiver operating characteristic (ROC) curves for this experiment. To obtain the ROC curve, we sorted the final beliefs of all the files in ascending order and considered each value as a threshold; all files with final beliefs above that value were classified as good, or bad otherwise. Then, the TP rate and FP rate were computed using these classifications. We observe that AESOP achieved an impressive 0.9983 TP rate at 0.0001 FP rate while labeling over 1.6 million files.

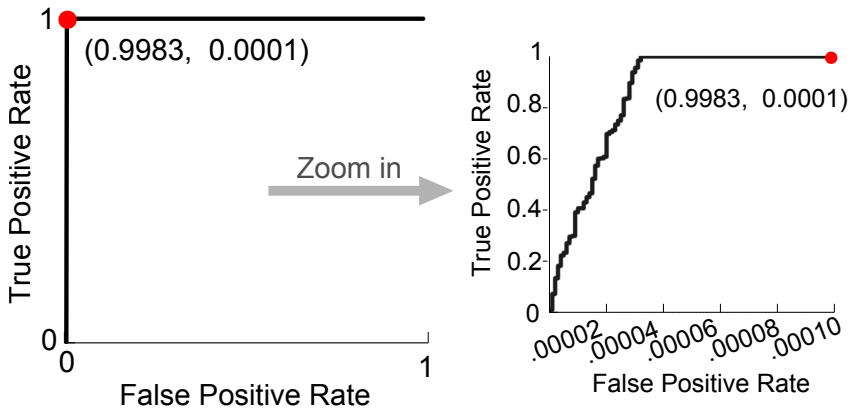



Figure 9: *Left*: ROC curve for the cross-validation experiment. AESOP achieved 0.9983 true positive rate in detecting malware at 0.0001 false positive rate while labeling over 1.6 million files. *Right*: Zoomed-in view.

### 2.5.7 Early Discovery of Unlabeled Benign and Malicious Files

Next, we test the effectiveness of AESOP in assigning labels to unlabeled files. To this end, we retrieved updated file label information on November 13, 2013 and also on February 1, 2014. We first focused on the files that were unlabeled on November 6 and become labeled as of February 1 (we refer to these files as converted files), and we examined if AESOP could predict the labels of the converted files accurately using the label information we had originally. There were 774 unlabeled-to-bad and 17,997 unlabeled-to-good converted files. Here, BP was set to run for 10 iterations.

Figure 10 shows the overall and zoomed-in ROC curves for this experiment. We obtained the ROC curves with an approach similar to that described in Section 2.5.6; the main difference is that we used the updated file label information from February 1 when computing the TP rate and FP rate values. We observe that AESOP achieved an impressive 0.9961 TP rate at 0.0001 FP rate while labeling over 18 thousand originally unlabeled files.

To examine if AESOP can label files ahead of Symantec's existing Polonium-based

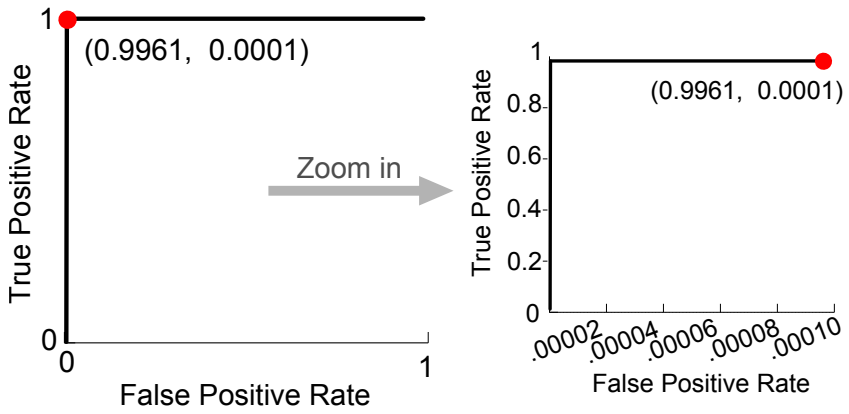



Figure 10: *Left*: ROC curve for the early discovery experiment. AESOP achieved 0.9961 true positive rate in detecting malware at 0.0001 false positive rate while labeling over 18 thousand originally unlabeled files. *Right*: Zoomed-in view.

technology [38], we considered the file label information from November 13 and computed how many converted files were labeled as of this date. From Figure 11, we observe that only a small number of conversions happened within the first week, showing that AESOP could label the converted files *at least* one week ahead of the Polonium-based technology in this case.

### 2.5.8 Performance Comparison with Polonium

Next, we present a direct comparison of AESOP with the state-of-the-art Polonium algorithm [38] in terms of file labeling effectiveness. Here, we considered the setting and the experiment in Section 2.5.7 again, with Polonium configured as described in [38]. Figure 12 shows the overall and zoomed-in ROC curves for this experiment. We obtained the ROC curves as described in Section 2.5.6. We observe that AESOP outperformed Polonium by achieving higher TP rate values across the whole spectrum of FP rate values. Specifically, at 0.0001 FP rate, AESOP achieved 0.9961 TP rate, whereas Polonium achieved 0.7674 TP rate.

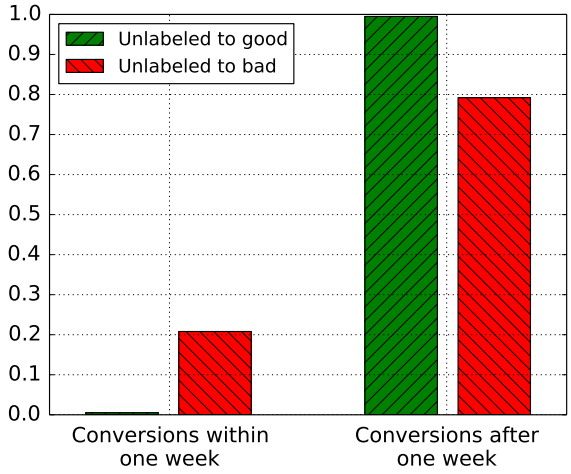



Figure 11: Fraction of unlabeled files that were and were not assigned labels within a week of the sample generation date. AESOP could provide at least a week's advantage in assigning labels to a significant amount of unlabeled files.

### 2.5.9 Scalability

Finally, we evaluate the scalability of AESOP by studying how much time it needs to cluster and label the files. Here, we considered the setting and the experiment in Section 2.5.7 again, recording the number of seconds taken by LSH configured with seven threads, and BP configured with a single thread. We observed that LSH took 5,751 seconds and BP took 282 seconds. LSH required the most amount of time; this is expected because it performs the initial processing of the input dataset for BP (recall that our dataset was 120GB in size). To demonstrate the scalability of LSH, we also ran it on smaller datasets generated by randomly sampling 10%, 20%, ..., 90% of the files. Figure 13 shows the results, which empirically validate that LSH scales linearly with the number of files to be clustered.

## 2.6 Conclusions

This chapter presents AESOP, an algorithm that uses the principle of guilt by association to establish nuanced reputation scores for executable files based on the company they keep. We use a large dataset voluntarily contributed by the members of Norton

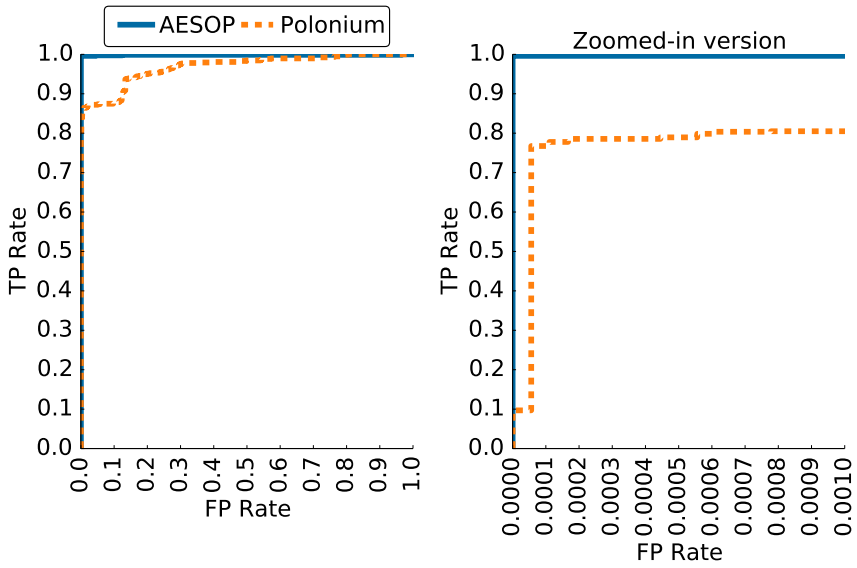



Figure 12: ROC curves for the comparison with Polonium experiment. AESOP outperformed Polonium by achieving higher true positive (TP) rate values across the whole spectrum of false positive (FP) rate values.

Community Watch, consisting of partial lists of the files that exist on their machines. AESOP leverages locality-sensitive hashing to efficiently compute file similarity values to construct a file-relation graph for inferring file goodness based on belief propagation. Our experiments show that AESOP achieved early detection of unlabeled files with exceptionally low error rates. AESOP is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.

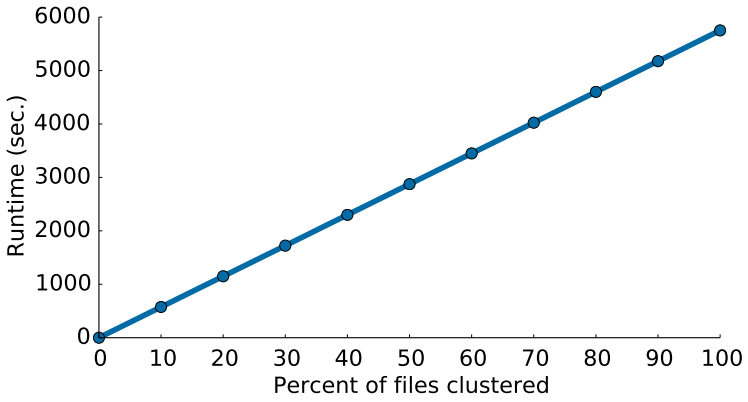



Figure 13: Scalability of Aesop. The runtime to cluster files is linear in the number of files in the dataset.

# CHAPTER III

# APPLICATION OF ADAGE TO MALWARE DETECTION

ADAGE is an algorithm that systematically determines the appropriate intervals to construct a sequence of graph snapshots from streaming edges. ADAGE was developed in a joint effort led by our collaborators; we contributed mainly with an extensive case study on malware detection using a propagation-based algorithm to demonstrate the usefulness of ADAGE in practice.

Consider a social network of people, which represents the friendship relationships between the individuals. Assume that the relationships are dynamic (or time-evolving) in that a relationship between two individuals can be formed at any time in the network (hence the streaming relationships or edges). In this setting, analysts often want to grab longitudinal snapshots of the network to study topics such as network growth or evolution of the communities. The current practice in generating the snapshots is to use a single fixed-length interval, whose length is often arbitrarily selected. ADAGE provides a systematic way to determine the appropriate intervals to generate the snapshots.

In the context of malware detection, prior work [38] used a machine-file graph that captures the relationships between machines and files, denoting which file appears on which machine. The prior work infers the nature of the unknown files by propagating goodness scores between the files and the machines in the graph. Assume a setting with a finite stream of time-stamped machine-file relationships. In this case, the prior work would consider the final, full graph that includes all the relationships. We made the novel observation that leveraging the smaller snapshots of the graph generated

Material adapted from work appeared at WWW 2016 [157].

from the intervals determined by ADAGE can enable us to detect malware more accurately—by propagating goodness scores between the files and the machines as the prior work does—in comparison to using the final graph. This is because it is often the case that infected machines receive a short burst of malicious files over a time-span of minutes, therefore longer snapshots destroy the purity of the graph's connected components by polluting these bursty malware clusters with increasing numbers of benign files. Effectively, longer snapshots lose the finer granularity needed to detect short-lived trends in the data by increasing the graph's density.

We validated our observation with an extensive case study over 574 thousand files, achieving an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots in comparison to 43% true positive rate at 0.01% false positive rate with the final graph. This observation we made is patent-pending.

## 3.1 Introduction

This work addresses the problem of determining the *proper* intervals for aggregating a stream of time-stamped edges into a sequence of *structurally mature* networks. Despite its importance, this problem has received very little attention from the research community. Existing approaches frequently select a single fixed-length interval, whose length is often arbitrarily selected. For instance, they group all of the edges that arrive during the same day into a single graph.

To identify the appropriate aggregation intervals, this work introduces ADAGE, short for *Adaptable Graph Edge Interval Framework*. ADAGE was developed in a joint effort led by our collaborators; we contributed mainly with an extensive case study on malware detection. ADAGE partitions a timeline of streaming time-stamped edges into disjoint, variable-length intervals, each giving rise to a single structurally mature graph snapshot. This work is inspired by the following observations.

**Observation 1.** Graph mining algorithms require their input graphs to possess some amount of *structure* (such as a large connected component). Without such structure, meaningful graph analysis is impossible. For example, belief propagation (BP) will perform poorly on a graph that is a collection of dyads.

**Observation 2.** One should use as *short* an interval as possible to produce a graph snapshot possessing the necessary structure. First, if one is given a finite timeline and wishes to understand network change, it makes sense to obtain as many structured snapshots as possible within that timeline. This allows for more fine-grained understanding of graph dynamics. Moreover, some applications (like BP) can perform poorly on dense graphs.

**Observation 3.** Intervals should be of *variable lengths*. Data can stream at very different rates during the observation timeline. For example, in the famed Enron Email dataset [33], some days contain tens of emails, while others contain hundreds. A fixed-length interval would not be suitable in such cases.

To identify structural maturity, ADAGE utilizes characteristics of real-world graphs, such as the existence of a large connected component. It postulates that a network is structurally mature when it has stabilized with respect to such a characteristic. To apply ADAGE, a user selects a graph statistic based on phenomenon under study, such as the size of the largest connected component. Given such a statistic, the algorithm aggregates data until convergence is seen with respect to that statistic.

It is important to note that ADAGE looks for structurally mature snapshots, rather than attempting to find a sparse snapshot that represents the entirety of the timeline. One would certainly expect that statistics change substantially in different parts of the timeline; indeed, when studying network evolution of some graph statistic, one would hope that the statistic changes.

The remainder of this chapter proceeds as follows. We first survey related work.

We then proceed to a description of ADAGE, followed by our extensive case study on malware detection using a propagation-based algorithm to demonstrate the usefulness of ADAGE in practice. Finally, we end by presenting our conclusions.

## 3.2 Prior Work and Our Differences

While the bulk of research on networks deals with a static representation, the recent past has witnessed increasing interest in studying the structure of time-evolving networks. Related work can be grouped into three main parts: (1) patterns and models for time-evolving networks, (2) mining time-evolving networks, and (3) analysis of aggregation intervals.

### 3.2.1 Models for Time-evolving Networks

A body of work is concerned with discovering laws and patterns in longitudinal networks [107, 9, 8, 105, 11]. In [107], the authors examine a set of time-evolving networks, and find that they obey two main power laws: *densification* or the growth of the average degree, and *shrinking* of the diameter of the network over time, contrasting with previous assumptions as those made in [105, 11]. In [8, 9], Akoglu et al. propose models for generating time-evolving networks, while satisfying additional power laws observed in real data, such as the eigenvalue power law. While all of these pieces of research split the evolving network into snapshots of arbitrary durations (depending on the type of the data), ADAGE seeks to identify aggregation intervals that are more meaningful and better structured relative to some metrics or tasks.

### 3.2.2 Mining Time-evolving Networks

Related literature has also addressed the problem of designing algorithms to mine different properties of time-evolving networks [17, 21, 161, 169, 20]. In [169, 20], the authors present a framework for analyzing group behavior and finding communities

over time, whereas [17] focuses on empirical evolution of groups in large social networks. The aim of [21] is to mine frequent patterns of interaction that appear more than expected in a series of snapshots of a network. Sun et al. [161] propose a method for mining patterns and anomalies in large evolving networks.

### 3.2.3 Aggregation Intervals

More closely related to ADAGE are [104, 84, 159]. In [104], a call network is analyzed using *fixed-length* aggregation intervals. The goal of [104] is to evaluate the impact of these intervals on call patterns, as opposed to that of ADAGE, which is finding good *variable-length* intervals, independently of the nature of the data. In [84], the authors are concerned with a different task: combining edges of a temporal contact network into a *single* snapshot. TWIN is aimed at finding underlying cyclical patterns or rhythms to streaming data [159]. To the best of our knowledge, no prior work considers malware detection by partitioning a data stream. In this work, we present an extensive case study of ADAGE on malware detection using a propagation-based algorithm.

## 3.3 Description of ADAGE

ADAGE is an online method for aggregating streaming edges into a sequence of structurally mature networks. Given a network statistic (e.g., exponent of the degree distribution), ADAGE aggregates the time-stamped edges into a network until the value of the statistic converges. Figure 14 depicts the length of time intervals automatically detected by ADAGE on Facebook wall-postings vs. the exponent of the degree distribution of the composed graphs. A graph represents the following relationship between users: *user i posted on user j's wall.* Each time step on the x-axis is an hour. 40 hours of data were needed to generate the first structurally mature graph. That is, it took 40 hours worth of edge streams to compose a graph with a stable degree distribution exponent. In the next interval, it took only 10 hours to

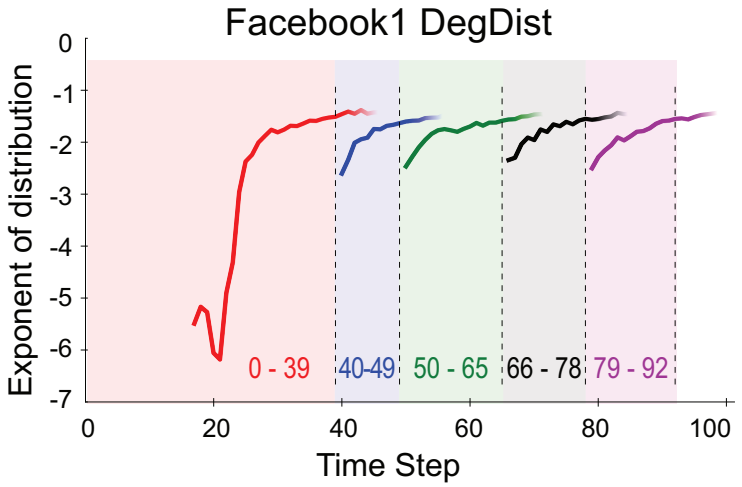



Figure 14: Intervals automatically detected by ADAGE with degree distribution exponent as the graph statistic on Facebook wall-postings. Dashed lines indicate intervals. Each time step is one hour. The curves do not start at the beginning of each interval because there is not enough data early in the interval to calculate the exponent of the degree distribution.

generate a structurally mature graph; and so on. Note that a fixed-length interval would not have worked well in this particular example.

ADAGE takes as input a (discretized) sequence of edge sets $E_1, E_2, \ldots$ arriving at times $T_1, T_2, \ldots$ and a function $f(G)$, which outputs the value of a specified statistic on the graph $G$. At each time $T_i$, $f$ is applied to the current aggregated graph $G_i$ to obtain a statistic value $r_i$. The $r_i$ values are then inspected for convergence.

ADAGE can take any network statistic such as exponent of degree distribution, exponent of triangle count distribution, clustering coefficient, number of nodes in the largest connected component, effective diameter, etc. The choice of statistic for ADAGE depends on the nature of phenomenon under study. For example, the exponent of the degree distribution is a good statistic to track if the phenomenon is expected to exhibit the Pareto principle.

Figure 15 provides an overview of ADAGE, which begins at time $T_1$ and aggregates data until convergence of the chosen graph statistic is detected (i.e., adding more edges will not greatly alter the value of the statistic). To determine whether convergence has

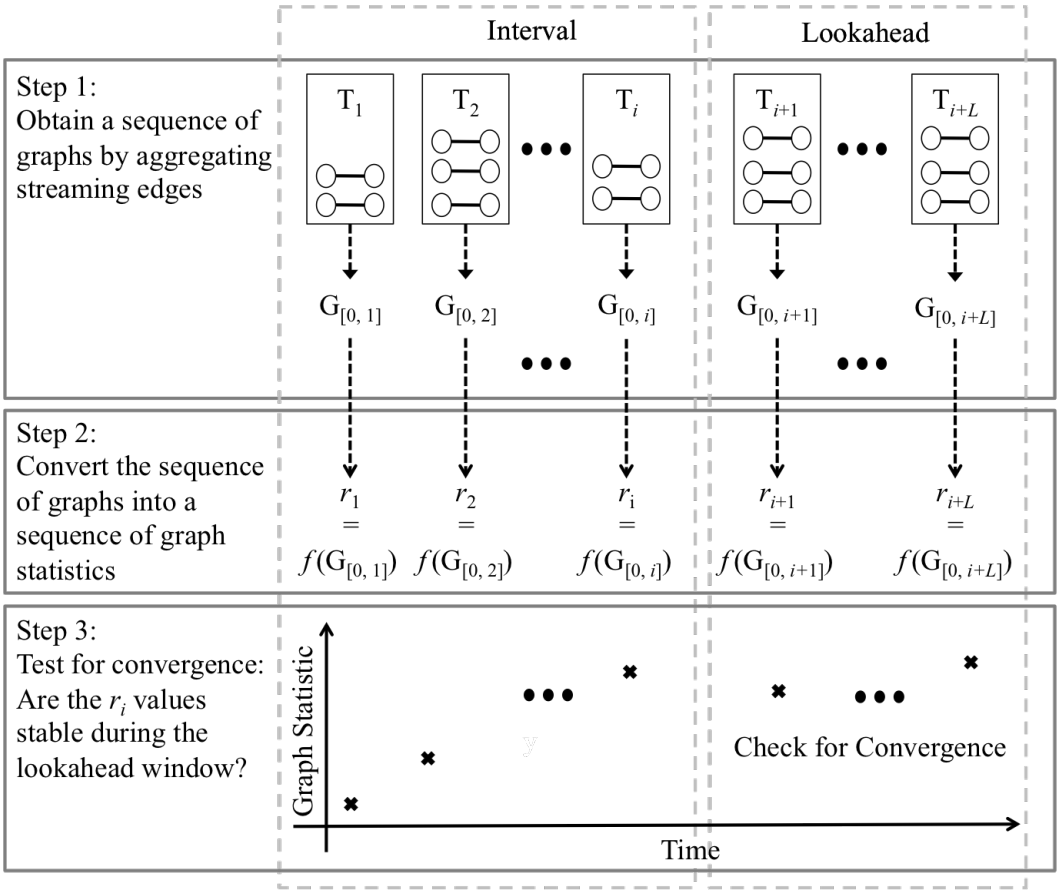



Figure 15: Overview of the ADAGE algorithm. A stream of time-stamped edges is aggregated until convergence is detected on the chosen graph statistic.

occurred at time $T_i$, ADAGE examines the value $r_i$ and the set of values $\{r_{i+1}, ..., r_k\}$ seen during the lookahead window $[T_{i+1}, T_{i+L}]$. The length $L$ of the lookahead window is calculated using a parameter $b$, where $L = b \times i$ ($L$ depends on the length of the interval so far). To avoid automatically detecting convergence after very short intervals, the window length is set to be at least 10. The allowed deviation in values $r_i, ..., r_k$ is controlled by a threshold parameter $c$: the difference between the largest and smallest values cannot exceed threshold $t$, which is equal to $c$ times the smallest value (assuming all values are positive). A parameter study suggested that $b = 0.1$ and $c = 0.1$ produce good results. Once convergence is detected, ADAGE outputs the graph and restarts the aggregation process.

ADAGE has several important strengths. It is *simple*, allowing for easy implementation and adoption to real problems. It is *flexible*; it can be tailored for any network statistic. It is *efficient*, easily accommodating statistic approximation through sampling, calculation of statistic values distributed over multiple processors, or modification of convergence parameters.

ADAGE makes two assumptions. First, it assumes that the stream of edges is discretized in time—e.g., the data might be discretized in seconds. ADAGE aggregates multiple seconds to produce intervals larger than a second. Clearly, if the initial discretization is too coarse (e.g., a year), then ADAGE might output each of these initial intervals as a structurally mature graph snapshot. Second, ADAGE (in its current form) assumes that once an edge has been added to a snapshot, it is not removed for the remainder of that interval. This assumption can be easily modified if the edge stream is labeled (i.e., add edge or delete edge).

## 3.4 Case Study on Malware Detection

In the context of malware detection, prior work [38] used a machine-file graph that enables to capture the relationships between machines and files, denoting which file appears on which machine. The prior work infers the nature of the unknown files by propagating goodness scores between the files and the machines in the graph. Assume a setting with a finite stream of time-stamped machine-file relationships. In this case, the prior work would consider the final, full graph that includes all the relationships. We made the novel observation that leveraging the smaller snapshots of the graph generated from the intervals determined by ADAGE can enable us to detect malware more accurately—by propagating goodness scores between the files and the machines as the prior work does—in comparison to using the final graph. This is because it is often the case that infected machines receive a short burst of malicious files over a time-span of minutes, therefore longer snapshots destroy the purity of the graph's

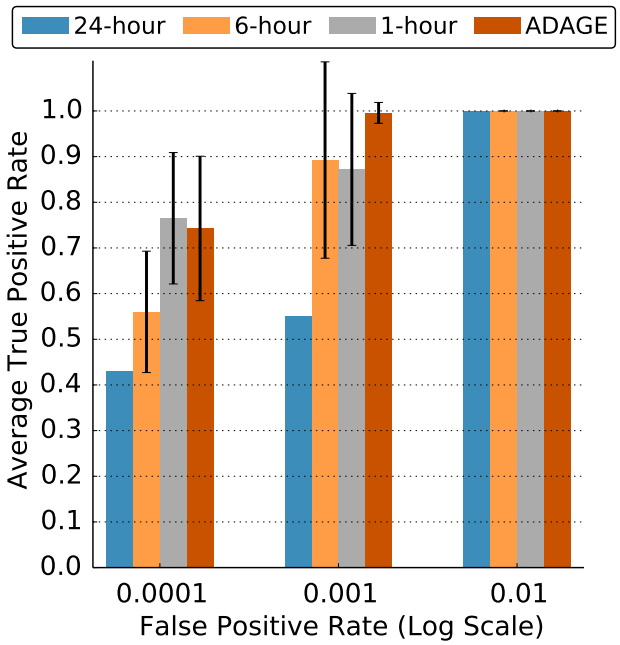



Figure 16: Results for malware detection on a machine-file graph. For a fixed false positive rate, a higher true positive rate indicates better performance. (1) Shorter intervals can sometimes produce better results than longer intervals. (2) ADAGE offers a principled method for identifying intervals that are competitive with ad-hoc fixed-length intervals.

connected components by polluting these bursty malware clusters with increasing numbers of benign files. Effectively, longer snapshots lose the finer granularity needed to detect short-lived trends in the data by increasing the graph's density.

To validate our observation, we performed an extensive case study as follows. We obtained a dataset from Symantec's Worldwide Intelligence Network Environment (WINE), which samples and aggregates datasets that Symantec uses in its day-to-day operations to share them with the research community [61]. Our dataset covers one day with a time granularity of 10 minutes, and it contains 3,392,983 machine-file relationships between 574,733 files and 53,174 machines. Some files in the dataset are known to be malicious or benign. As [38] reported a power-law degree distribution for their machine-file graph (with a few files residing on many machines and many files residing only on a few machines), we chose the exponent of the degree distribution as the network statistic that ADAGE should track. We randomly selected 30

starting points within the duration of our dataset and used ADAGE to determine the appropriate intervals to generate the graph snapshots. ADAGE took on average 4.19 seconds to detect structurally mature graphs, each with around 0.5 million edges. We also generated graph snapshots with the following fixed-length intervals: 24-hour (the final, full graph), 6-hour, and 1-hour. Afterwards, we reconstructed the approach in [38] to detect malware on these graphs by propagating goodness scores between the files and the machines. To measure performance, we computed the average true positive rate at different false positive rates. Here, the true positive rate is the fraction of malware instances correctly labeled as bad, and the false positive rate is the fraction of benign files incorrectly labeled as bad. Figure 16 shows the results for ADAGE and the fixed-length intervals. We achieved an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots determined by ADAGE in comparison to 43% true positive rate at 0.01% false positive rate with the final graph that [38] would consider. Also, we observe that ADAGE was able to *automatically* find aggregation lengths that match or outperform the other two shorter fixed-length intervals. This case study validated our observation, which is now patent-pending.

## 3.5 Conclusions

This chapter presents ADAGE, a flexible algorithm for partitioning a timeline of streaming edge data into variable-length intervals in order to generate a sequence of structurally mature graphs. ADAGE was developed in a joint effort led by our collaborators; we contributed mainly with an extensive case study on malware detection using a propagation-based algorithm to demonstrate the usefulness of ADAGE in practice. We made the novel observation that leveraging the smaller snapshots of a machine-file graph generated from the intervals determined by the ADAGE algorithm can enable us to detect malware more accurately in comparison to using the final, full graph that includes all the machine-file relationships. We validated our observation

with an extensive case study over 574 thousand files, achieving an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots in comparison to 43% true positive rate at 0.01% false positive rate with the final graph. This observation we made is patent-pending.

# CHAPTER IV

# EDOCS: EFFORT-BASED DETECTION OF COMMENT SPAMMERS

Detecting comment spammers that use comment threads on social media platforms to post spam content is an important problem because spam comment messages have become prevalent [3] and dangerous, with some containing links to malware sites [95]. The majority of the existing techniques consider each comment message independently and attempt to determine if it is spam or not by examining the properties of the comment and its sender [118, 3, 95, 151, 50].

We made the novel observation that comment spammers tend to be lazy and put limited *effort* towards preparing and disseminating their comments, therefore it might be possible to detect the comment spammers if we can quantify the effort scores of the social media users (i.e., the users with low effort scores are expected to be spammers). For instance, we observed that some spammers recycle the comment messages and share the same IP addresses with other spammers, as each message is time-consuming to craft and obtaining unique IP addresses is costly. Assuming that the comment messages and the IP addresses are the two effort-requiring resources, graphs enable us to capture the relationships between the users and these resources, denoting which user posted a particular comment message and had a specific IP address. By doing so, we differ from the existing techniques as we consider all the comment messages in relation to each other in the broader context. Our EDOCS algorithm leverages a graph that represents such effort-related relationships, on which it performs message propagation to quantify the effort scores of the users, and it then flags the users with

Material adapted from work appeared at IEEE S&P 2015 [166].

low effort scores as spammers.

EDOCS detected comment spammers across over 197 thousand users accurately with 95% true positive rate at 3% false positive rate as well as preemptively (i.e., it detected spammers early on), and it outperformed the existing technique used by Yahoo (exact performance details proprietary). EDOCS is patent-pending, has been integrated into Yahoo's anti-abuse technology for their social media platforms, and guards multiple online communities from comment spammers.

## 4.1 Introduction

In recent years, social media has become ubiquitous and important for content sharing. An example of how users contribute content to a social media platform is through comment threads in online articles (e.g., news), which allow users to share their insights and engage in discussions with each other. An important aspect of the comment space is its open nature; in most social media platforms one can post a comment anonymously or with an account that can be obtained in a matter of seconds. Also, comments posted on a popular social media platform can easily reach a significant number of users.

Unfortunately, this open nature of the comment space provides malicious users with various opportunities to abuse it. For instance, abusers often use comment threads to post content irrelevant to the article. Such content is typically referred to as spam, posted by the so-called comment spammers [118]. Comment spammers are posing a serious problem; a recent study showed that more that 75% of the one million blog comments collected were indeed spam [3]. Furthermore, some spam comment messages are extremely malicious; they contain text luring users to click links leading to malware sites [95].

However, detecting comment spam is challenging for the following reasons. Comment spam is different from other forms of spam in that a typical spam comment

message is usually short and carefully crafted by humans; even human experts have hard times differentiating some spam comments from legitimate ones [95].[1] In contrast, the majority of spam email messages, for instance, are generated by botnets using certain predefined templates [143]—an important property leveraged by many approaches tackling email spam (see [27] for a survey). Relying solely on human experts to detect comment spam is also not feasible; human experts simply do not have the bandwidth to deal with the enormous amounts of content generated by users in today's social media era [95]. In addition, recent research showed that human experts are not very effective in detecting spam messages [132, 131].

The existing approaches proposed for comment spam take a comment-level view to the problem in that they attempt to classify a comment message as spam or not spam by mainly considering the characteristics of the comment and its sender [118, 3, 95, 151, 50]. We take a different slant on the problem and propose *Effort-based Detection of Comment Spammers* (EDOCS), a graph-based user-level approach that quantifies how much *effort* a user exerted over his or her comments, to detect if the user is a comment spammer or not. As we will explain below, we expect that the effort scores of the comment spammers are lower than those of the legitimate users.

The remainder of this chapter proceeds as follows. We first survey related work. We then proceed to a description of EDOCS, followed by the experiments we conducted to demonstrate its effectiveness. Finally, we end by presenting our conclusions.

## 4.2 Prior Work and Our Differences

Comment spam detection is a relatively new area of research that has become important with the increasing popularity of the social media platforms. Below, we review work related to ours.

[1] In our context, human experts are editors whose job responsibility include labeling users' comments as spam or not spam in a social media platform.

Mishne et al. [118] present an approach that compares the language models built from the comment, the associated article or blog post, and pages linked by the comment. The authors expect these language models to be different as spammers usually create links between sites that have no semantic relation, and they exploit the difference in the models using the Kullback-Leibler (KL) divergence measure to classify the comments.

Kantchelian et al. [95] define spam as content that is uninformative in the information-theoretic sense and propose a metric called content complexity that measures the informativeness of the comments using the entropy rate. The authors leverage this metric to identify a set of features adjusted to comment spam detection, and they develop a latent logistic regression classifier based on these features, which can tolerate noisy and missing class labels.

Cormack et al. [50] focus on spam filtering for short messages such as comments and mobile (SMS) messages, and determine that they contain an insufficient number of words to properly support bag of words or word bigram-based spam classifiers. The authors show that the performance of these classifiers can be improved considerably by expanding the set of features to include orthogonal sparse word bigrams as well as character bigrams and trigrams. Among the various classifiers evaluated, the Dynamic Markov Compression (DMC) method is found to perform best on short messages and message fragments.

Abu-Nimeh and Chen [3] present a multi-stage approach that extracts the terms frequently appear in the comments using the term frequency-inverse document frequency (TF-IDF) method and runs them against a support vector machine (SVM) classifier. To improve the accuracy of the classifier, they combine it with several heuristics and decide whether to classify a comment as spam or not spam by weighing the classifier and heuristics results in a final score.

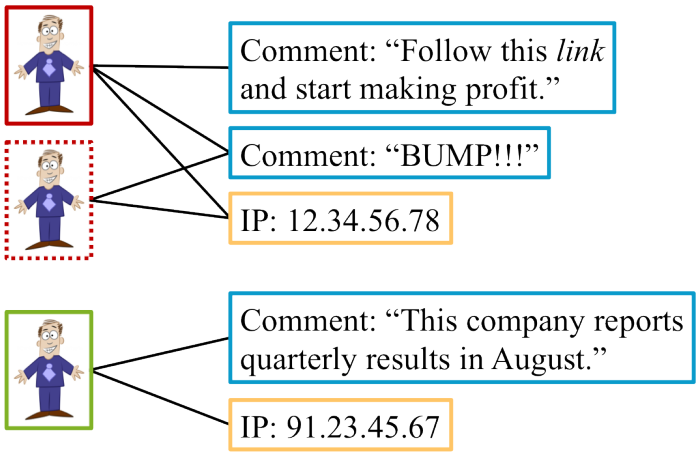



Figure 17: Our EDOCS algorithm leverages a graph that captures the relationships between the social media users and the effort-requiring resources of comment messages and IP addresses to detect comment spammers. In this toy graph, the users in the red and green rectangles are spammers and a legitimate user, respectively. (Cartoon image from `wikihow.com`)

Sculley and Wachman [151] consider the online setting where the SVM classifier makes a decision on a new comment, is told if its prediction is correct, updates its hypothesis accordingly, and then awaits a new example. The authors show that online SVMs give good classification performance on online comment spam filtering, and they propose a relaxed online SVM method that achieves nearly equivalent performance at reduced computational cost.

In summary, all of the above pieces of research consider each comment message independently and attempt to determine if it is spam or not by examining the properties of the comment and its sender. In contrast, we consider all the comment messages in relation to each other in the broader context by leveraging a graph that captures the relationships between the users and the effort-requiring resources of comment messages and IP addresses.

## 4.3 Our Approach: The EDOCS Algorithm

### 4.3.1 Why Quantifying Effort Can Help Detect Spammers?

We made the novel observation that comment spammers tend to be lazy and put limited *effort* towards preparing and disseminating their comments, therefore it might be possible to detect the comment spammers if we can quantify the effort scores of the social media users (i.e., the users with low effort scores are expected to be spammers). For instance, we observed that some spammers recycle the comment messages and share the same IP addresses with other spammers, as each message is time-consuming to craft and obtaining unique IP addresses is costly. We propose EDOCS to utilize this observation, by analyzing a bipartite graph of users and effort-requiring feature values (see Figure 17 for an example) to quantify how much effort a user exerted over his or her comments. EDOCS outputs an overall *effort score* for each user, taking into account all the comments that the user posted.

### 4.3.2 The EDOCS Algorithm

EDOCS operates on a bipartite graph of users and effort-requiring feature values. A user is connected to all the feature values that apply to him or her (e.g., an edge connecting the user with his or her IP address). EDOCS performs iterative message propagation on this graph. Specifically, messages are first propagated from users to feature values, where they are aggregated using feature-specific aggregation functions, and these aggregated messages are then propagated back to the users. The propagation ends when a maximum number of iterations is reached, after which an overall effort score is computed for each user using a general aggregation function.

In its current form, EDOCS performs the message propagation for two iterations given the scale of our dataset (see details below), and it utilizes the two important features present in our dataset: the body of the comment and the IP address of the comment poster. If a user posts the same comment body multiple times, possibly

Table 6: Characteristics of our comments dataset.

| | |
|---|---|
| Number of users | 197,464 (20.03% spammers) |
| Number of comments | 1,201,277 |
| Mean/median number of comments per user | 6.08/1 |
| Dataset duration | May 1–31, 2014 |
| Duration of follow-up period | June 1–August 5, 2014 |

with other users, and shares the same IP address with other users, this might be an indication of a spamming activity or campaign. To capture this, EDOCS executes with the following message values and aggregation functions.

- *Comment body effort:* Each user node sends to the neighboring comment body nodes a message containing as its value the total number of times the user posted the corresponding message. Each comment body node computes the sum of all the incoming messages' values and sends the reciprocal of the sum to the neighboring user nodes.

- *IP effort:* Each user node sends to the neighboring IP address nodes a message containing the value 1. Each IP address node computes the sum of all the incoming messages' values and sends the reciprocal of the sum to the neighboring user nodes.

- *Overall effort:* Each user node computes the sum of all the messages' values arriving from the comment body nodes and normalizes the sum by the total number of comments the user posted. Similarly, the user node computes the sum of all the messages' values arriving from the IP address nodes. Finally, the user node returns the sum of these two values as the overall effort score for the corresponding user.

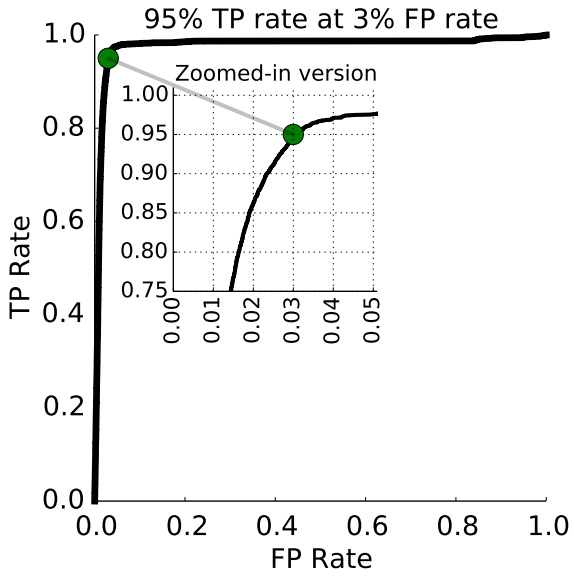



Figure 18: Receiver operating characteristic (ROC) curve for the spammer detection experiment. EDOCS achieved 95% true positive (TP) rate in detecting spammers at 3% false positive (FP) rate while labeling over 197k users.

## 4.4 Experiments

### 4.4.1 Dataset

We use a dataset containing user comments posted on the finance portal of a large internet company during May 2014. The characteristics of our dataset are shown in Table 6. A user is assumed to be a spammer if he or she posted at least one comment labeled as spam by human experts.

### 4.4.2 Detecting Spammers

Figure 18 shows EDOCS's effectiveness in detecting spammers with a receiver operating characteristic (ROC) curve; EDOCS achieved an impressive 95% true positive (TP) rate at 3% false positive (FP) rate, assuming that spammers belong to the positive class. We generated the ROC curve as follows: (i) we ran EDOCS to obtain an effort score for each user; (ii) we considered each effort score in ascending order (recall that low effort scores are indicative of spammers) and used the effort score

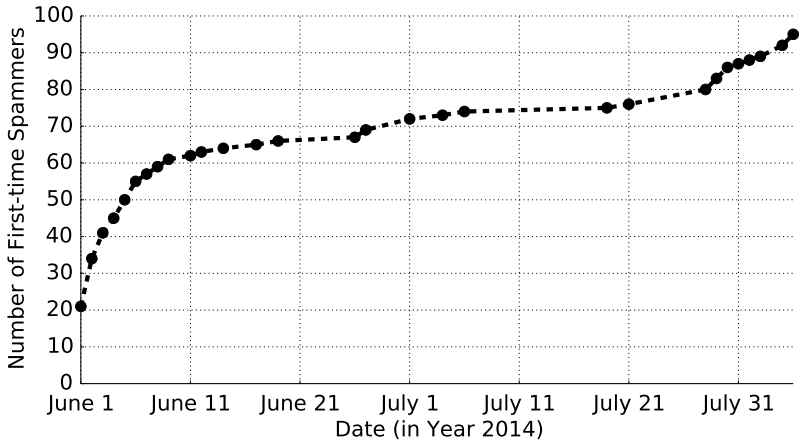



Figure 19: Conversion trend of users from "clean" to spammer based on the date of their first spam comment messages during the follow-up period (June 1–August 5, 2014). EDOCS preemptively detected these 95 users (top right corner) as spammers using data from May 2014.

as a cutoff value for classification—a user who had an effort score smaller than the cutoff value was labeled as spammer, or clean otherwise; (iii) using the classifications of users generated from each cutoff value, we finally computed a pair of TP rate and FP rate values; plotting and connecting these pairs of values gave us the smooth ROC curve in Figure 18.

### 4.4.3 Follow-up on False Alarms

We next focus on the users belonging to the FP set that we obtained from the cutoff value used in the 95% TP rate at 3% FP rate result in Section 4.4.2. Note that these are the users that EDOCS labeled as spammers, however they did not have any spam message within the duration of our dataset. To examine if these users were indeed "clean", we followed them for two more months (June 1–August 5, 2014) and we checked if they posted any spam comments. Out of 937 users who had a comment during this follow-up period, 95 of them posted at least one spam comment message, resulting in a 10.1% clean-to-spammer conversion rate. Figure 19 shows the conversion trend based on the date of the first spam comment messages. Note

that conversions occur consistently, showing the effectiveness of EDOCS in detecting spammers preemptively (i.e., it can detect spammers early on).

## 4.5 Conclusions

We tackled the crucial problem of comment spam and proposed EDOCS, a graph-based approach that quantifies how much effort a user exerted over his or her comments, to detect if the user is a comment spammer or not. Our experimental evaluation of EDOCS showed its effectiveness in detecting comment spammers accurately with 95% true positive rate at 3% false positive rate as well as preemptively, and it outperformed the existing technique used by Yahoo (exact performance details proprietary). EDOCS is patent-pending, has been integrated into Yahoo's anti-abuse technology for their social media platforms, and guards multiple online communities from comment spammers.

# CHAPTER V

# CHARACTERIZING SMOKING AND DRINKING ABSTINENCE FROM SOCIAL MEDIA

Alcohol and tobacco are among the top causes of preventable deaths in the United States [120]. Achieving long-term abstinence of tobacco or alcohol is difficult [175]—most abstainers are known to relapse within one to three months of cessation. Prior work examining addiction behavior manifested on social media investigates mainly the role of linguistic attributes in characterizing health challenges related to addiction [124, 110]. Also, these pieces of research use crowdsourcing to obtain information about the abstinence status of the individuals. However, simply looking at social media posts may not always allow third-party judges to reliably capture abstinence status.

In our work, which consists of two parts, we focused on two prominent smoking and drinking cessation communities on the social media site Reddit: StopSmoking and StopDrinking. These communities are identified as "self-improvement communities" on Reddit and are geared toward providing support and motivation to smoking and drinking addiction sufferers. A unique aspect of these communities is that they allow the users to acquire "badges". Badges are a mechanism by which the users can self-report the duration of their abstinence. We collected data on the users' badges, posts, comments, and associated metadata from these communities, and developed statistical models to analyze the role of social media language, interactions, and engagement in characterizing smoking/drinking abstinence and relapse. Addiction literature indicates social support to act as an important mediator of stress during smoking/drinking

Material adapted from work appeared at ACM Hypertext 2015 [164].

urges [153, 72]. In this context, graphs enable us to capture the interactions and engagement between the users, which reflect access to social support. Specifically, our models leverage a graph that represents which user provides social support to whom by writing comments on their posts in the communities. In summary, through our work, we extend the existing body of research by using self-reported abstinence information on smoking and drinking, and examining the additional role of interaction and engagement in characterizing these addiction-related health challenges.

The first part of our work, which we present in this chapter, focuses on characterizing abstinence from smoking and drinking. We used the badges of 1,168 users to construct ground truth information on short-term (<40 days) and long-term (>one year) abstainers, and we formulated and identified the key linguistic and interaction characteristics of these abstainers based on activity in the communities spanning eight years, from 2006 to 2014. We developed supervised learning-based statistical models based on these characteristics to distinguish long-term abstinence from short-term abstinence with over 85% accuracy. We found linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence.

The second part of our work, which we present in Chapter 6, focuses on characterizing relapse to smoking and drinking. Here, we used longitudinal data on the badges of 5,991 users to determine their abstinence or relapse status, and we formulated and identified the key engagement and linguistic characteristics of the abstainers and relapsers based on activity in the communities spanning almost nine years, from 2006 to 2015. We developed a robust statistical methodology based on survival analysis to examine how participation in the communities and the characteristics above relate to the risk of relapse. Our results show that although participation in the communities is not linked to high likelihood of smoking/drinking abstinence during the one/two months post-cessation, it shows a stable trend of heightened chance of abstinence

beyond three years, suggesting the efficacy of the communities in preventing relapse in the long term. Furthermore, we found positive affect and increased engagement to be predictors of abstinence.

The two parts of our work differ from each other in terms of the problem statement, the statistical method, and the dataset as follows. (1) The first part focuses on characterizing attributes of short-term and long-term abstinence from smoking/drinking. The second part focuses on modeling relapse events self-reported by individuals, and how they, collectively, might indicate the effectiveness of the communities in preventing relapse. (2) The first part uses a supervised learning-based statistical technique. The second part identifies the limitations of such supervised learning techniques in analyzing relapse events, and employs techniques from the survival analysis literature. (3) The first part considers a dataset with one badge per user. The second part expands this dataset with a unique method to obtain daily badges, and considers a dataset with multiple badges per user to determine the relapse events of the users.

## 5.1 Introduction

Health and well-being challenges such as smoking, alcoholism, and impulsive eating are known to be influenced by individuals' social environment [72], which are moving online, as social media sites become more popular. Indeed, the use of social media for health-related discourse have increased sharply in recent years [71]. Such use acts as a constantly available and conducive source of information, advice, and support, as well as known to foster positive behavior change [89]. Meanwhile, this new social interaction paradigm has begun to provide us with an opportunity to observe individuals' psychological states and social milieu, often in a real-time, longitudinal fashion.

We focus on the health challenge of addiction, specifically addiction to tobacco or alcohol. Alcohol and tobacco are among the top causes of preventable deaths in

the United States [120]. In addition to contributing to traumatic death and injury, alcohol is associated with chronic liver disease, cancers, acute alcohol poisoning, and fetal alcohol syndrome. Similarly, smoking is associated with lung disease, cancers, and cardiovascular disease [82]. Achieving long-term abstinence of tobacco or alcohol is difficult [175]—most abstainers are known to relapse within one to three months of cessation. In fact, many individuals who want to quit have been observed to go through short phases of relapse and cessation [73]. While there is a rich body of research on identifying factors associated with such short-term relapse or cessation [153, 178, 124], limited research examines the cues associated with long-term abstinence. This is largely due to the difficulty in compiling high quality self-reported data on abstinence from suitable populations, spanning over long periods of time.

In this work, we examine how social media language and interactions may be leveraged to *characterize long-term abstinence from tobacco or alcohol*. As of May 2013, 72% of online adults use social networking sites; the number is more than 80% for individuals under the age of 50.[1] Based on reports from the Centers for Disease Control and Prevention (CDC), this demographic aligns well with the age group in which heavy smoking and/or drinking are prevalent [150]. This suggests that social media may be a viable platform for mining cues associated with abstinence.

To this end, we focus on two prominent smoking and drinking abstinence communities on the social media site Reddit: StopSmoking[2] and StopDrinking[3]. These two communities together consist of more than 68 thousand subscribed Reddit users as of December 2015, and as described on their public pages, serve as "a place for Reddit users to motivate each other to control or stop smoking/drinking". A participating user may request to have a "badge" (see Figure 20) that indicates self-reported information about the duration of their smoking/alcohol abstinence. The badges are

[1] `www.pewinternet.org/data-trend/social-media/social-media-use-all-users/`
[2] `www.reddit.com/r/StopSmoking`
[3] `www.reddit.com/r/StopDrinking`

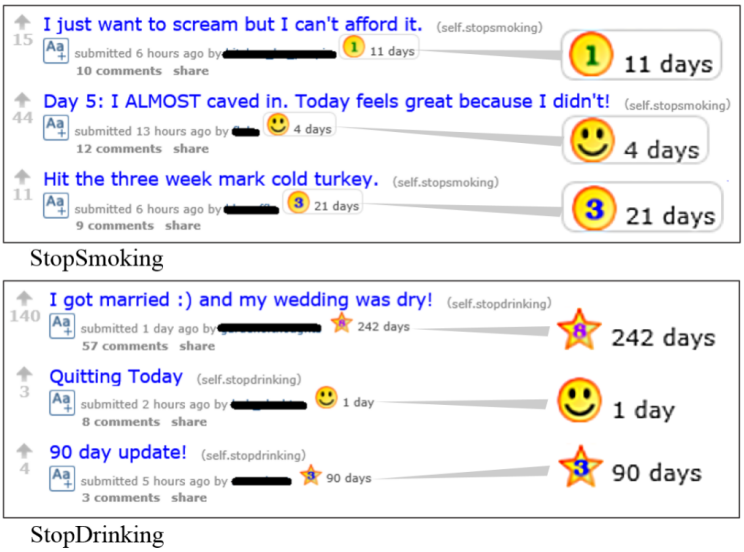



Figure 20: Examples of the users' abstinence badges on the StopSmoking and StopDrinking subreddits. The abstinence stage is displayed inside the badge icon (e.g., circle-shaped smiley face for "under one week") and the actual number of days of abstinence is reported next to it (e.g., 4 days).

dynamically updated in the system on a daily basis, unless the users request a change to their badges. The main contributions of this work include:

- We collect and study a novel dataset from Reddit that describes 1,168 users' self-reported information on their duration of smoking or drinking abstinence via the badges. We use the badge information to identify short-term and long-term abstainers.

- We formulate and identify the key linguistic and interaction characteristics of short-term and long-term abstainers based on activity spanning eight years, from 2006 to 2014.

- We build a supervised learning framework based on the characteristics above to distinguish long-term abstinence from short-term abstinence with over 85% accuracy, 88% precision, and 82% recall.

- Our findings present a number of significant discoveries that may help researchers better understand the role of social media language and interactions in assessing and determining tobacco or alcohol use. We find that:
  - the nature of affect manifested in Reddit posts and comments as well as the tenure of participation in Reddit communities are indicative of short-term or long-term abstinence;
  - the network properties of the users (e.g., indegree) based on their interaction patterns also bear significant explanatory power towards characterizing these addiction-related health outcomes.

We note here that our goal in this work is not to predict future success or failure in abstaining from tobacco or alcohol use. That is, we do not attempt to predict which individual will transition from being a short-term abstainer to long-term abstainer or will relapse while being a short-term or long-term abstainer. Rather, we study a set of successful abstainers and attempt to characterize the attributes of long-term smoking or drinking abstinence from social media. Through such characterization, we evoke the potential use of social media towards addressing public health challenges, in particular addiction to tobacco or alcohol.

## 5.2 Prior Work and Our Differences

### 5.2.1 Behavioral Science and Addiction

Clinical research on addiction shows that decreased psychosocial stress is associated with transitions from smoking to abstinence [128]. Smokers who fail to quit or relapse after a short period report high levels of stress prior to initial abstinence or at one, three, and six months after cessation [175]. Additionally, recent work analyzing the size and structure of individuals' social networks has found that their connections and interactions therein are related to health-related behaviors and goals [41]. Availability

of a strong, trusting network of friends can provide practical and emotional support, which can reduce their smoking or drinking urges [97, 48].

The findings of this extensive body of research provide evidence on the relationship between behavior and addiction. However, they rely heavily on small, often homogeneous samples of individuals, not necessarily representative of the larger population. Furthermore, these studies are typically based on surveys, relying on retrospective self-reports about mood and observations regarding addiction episodes. This method limits temporal granularity as it involves recollection of historical facts. Some of these limitations are circumvented through the use of wearable sensors and other electronic equipment that capture behavioral and affective data in real time without explicit intervention [153]. However, these methods are often expensive and intrusive because they need participants to use the equipment over a period of time.

As such, most behavioral science research on substance abuse has focused on relapse [134, 114, 178]. In fact, few population-based cohort studies have examined long-term abstinence (a year or more) among former smokers or alcoholics. It is important to quantify the relationship between the duration of abstinence and the likelihood of continued abstinence for the evaluation of ongoing public health interventions and the design of smoking or drinking cessation programs. Additionally, understanding factors associated with long-term abstinence is critical due to the high rate of relapse—most individuals attempting to quit tobacco or alcohol abuse go through multiple short-term phases of abstinence and relapse [72].

Our research specifically tries to address this problem. We develop computational approaches that can characterize the attributes of long-term smoking or drinking abstinence from social media. We derive a promising non-intrusive way to examine psychosocial attributes associated with long-term health outcomes by analyzing longitudinal and fine-grained activity in online communities.

### 5.2.2 Social Media, Health, and Addiction

Social media research has indicated that individuals' psychological states and social support status relating to health and well-being may be gleaned via analysis of language and conversational patterns. These include utilizing social media, largely Twitter, to understand conditions and symptoms related to diseases [139], cyberbullying and teenage distress [59], postpartum depression [56], mental health [57, 137, 85, 49], obesity and public health [1], exercise and mental health [144]. Broadly, this body of work investigated the role of linguistic attributes in describing or predicting health challenges.

We extend this body of research by examining the role of both language and social interactions gleaned from social media. Specifically, we build statistical language models that go beyond dictionary approaches. Additionally, we explore how network measures (e.g., indegree, neighborhood density, centrality, etc.) derived out of social interactions may bear explanatory power in the context of tobacco or alcohol addiction. Furthermore, we focus on Reddit, which remains underexplored in comparison to other social media platforms like Twitter.

There has been some research examining addiction behavior manifested on social media, however this body of work is limited. Relationship between displayed alcohol use on Facebook and self-reported information on alcohol abuse was examined in [122, 23, 123]. The authors in [35] explored sentiment manifested by individuals in Twitter by following a pro-marijuana profile. The structure of social circles of prescription drug abusers was investigated in [79]. Using Twitter, the authors in [125] examined perceptions of tobacco products. Another work conducted a study examining characteristics of individuals who express a desire to quit smoking on Twitter [124]. More recently, researchers have studied the prescription drug abuse recovery community Forum77 [110]. In a method similar to [124], they identified dictionary-based linguistic attributes of individuals in various phases of recovery, and were able

to characterize recovery trajectory of these individuals.

With the exception of [123] and [110], none of the above pieces of research focuses on predicting health challenges related to addiction. Furthermore, it is important to note that, the ground truth labels on recovery in [124] and [110] were obtained via crowdsourcing. Simply looking at social media posts may not always allow third-party judges to reliably capture abstinence status. Additionally, reasons such as idiosyncratic or personal usage patterns of social media as well as differential social norms and stigma may motivate or preclude some individuals from explicitly reporting abstinence information in social media content. Hence, self-reported abstinence information is extremely valuable. In this work, we leverage self-reported abstinence information on smoking and drinking.

## 5.3 Data

We begin with a short overview of Reddit. Reddit is a highly popular social media platform, where the users are often referred to as "redditors". They can submit content in the form of link posts or text posts. Posts are organized by areas of interest or sub-communities called "subreddits". For instance, some popular subreddits are r/Politics, r/programming, and r/science.[4] Redditors can engage on a post via "upvotes" or "downvotes"; the post's *score* is the difference between these two quantities. They can also post comments on a post and respond in a comment thread. Over time, redditors accrue reputation in two forms: *link karma* and *comment karma*. Link karma is proportional to the difference between the upvotes and downvotes in all the link posts users made. Comment karma refers to the same difference for all their comments. In 2014, Reddit had 71 billion page views, over 8,000 active communities, 55 million posts, and 535 million comments.[5]

[4]Subreddits are typically referred to with the prefix "r/". We omit the prefix when no ambiguity arises.

[5]`www.redditblog.com/2014/12/reddit-in-2014.html`

In this work, we focus on the following two self-improvement subreddits: StopSmoking and StopDrinking. We refer to them as SS and SD, respectively. Both subreddits host *public* content that can be viewed without a Reddit account. At the time of the writing of this chapter, SS had over 37 thousand subscribed users, while SD had 31 thousand subscribed users.

As we described before, both subreddits allow users to acquire "badges" to help track their abstinence progress (see Figure 20). Such badges are subreddit-specific, and are displayed next to the username whenever the user posts or comments on the subreddit (ref. Figure 20). Both SS and SD identify different stages of abstinence inside the badge icon (e.g., circle-shaped smiley face for "under one week"), although the actual number of days of abstinence is reported next to it as well.

Typically, a user makes a badge request to the moderators of the subreddit they are interested in, through the subreddit's interface or by privately messaging the moderators. Badges are then awarded by the subreddit moderators either manually (SD) or automatically through an application known as "badgebot" (SS). Both subreddits are heavily moderated and follow a set of guidelines. For instance, SD cautions against providing medical advice on the forum, conducting surveys, or advertising links to recovery centers.

### 5.3.1 Data Collection

We used Reddit's official API[6] to collect posts, comments, and associated metadata from the subreddits. Our data collection proceeded in three phases.

**Phase 1.** We collected a sample of users in SS and SD. The Reddit API limits crawling historical posts on a subreddit to the past 1,000 posts, so we obtained the most recent 1,000 posts from each of the two subreddits. The crawl took place in November 2014. For each post, we collected the title of the post, body or textual

[6] `www.reddit.com/dev/api`

Table 7: Summary statistics of the crawled dataset. The post and comment lengths are measured in words.

| | StopSmoking (SS) | | StopDrinking (SD) | |
|---|---|---|---|---|
| | All data | Ground truth data | All data | Ground truth data |
| Users | 1,859 | 635 | 1,383 | 533 |
| Total posts from users | 86,835 | 36,713 | 59,201 | 30,178 |
| Total comments from users | 766,574 | 306,560 | 492,573 | 229,656 |
| Date of earliest post | Dec. 09, 2006 | Dec. 09, 2006 | Feb. 18, 2006 | Feb. 18, 2006 |
| Date of earliest comment | Aug. 29, 2006 | Aug. 29, 2006 | Aug. 02, 2007 | Aug. 02, 2007 |
| Date of latest post | Nov. 23, 2014 | Nov. 23, 2014 | Nov. 23, 2014 | Nov. 23, 2014 |
| Date of latest comment | Nov. 23, 2014 | Nov. 23, 2014 | Nov. 23, 2014 | Nov. 23, 2014 |
| Mean / Median comment karma | 4,390.2 / 846 | 5,065.4 / 1,391 | 3,808.6 / 406 | 4,610.2 / 745 |
| Mean / Median link karma | 1,312.7 / 88 | 1,626.2 / 201 | 1,184.7 / 7 | 1,794.9 / 38 |
| Mean / Median comments per post | 6.8 / 5 | 7.1 / 5 | 12.6 / 9 | 13.2 / 9 |
| Mean / Median post score | 37.5 / 4 | 36.9 / 4 | 34.3 / 5 | 34.1 / 5 |
| Mean / Median comment score | 5.5 / 1 | 5.5 / 2 | 5.2 / 2 | 5.0 / 2 |
| Mean / Median post length | 55.2 / 15 | 55.3 / 14 | 67.5 / 17 | 62.7 / 15 |
| Mean / Median comment length | 31.9 / 16 | 32.6 / 17 | 36.7 / 18 | 39.2 / 19 |

content, ID, timestamp, author ID, author's comment and link karmas, and score of the post. We collected the same information for each comment on the post as well. We then used the API to obtain the badge value of the post author and each of the comment authors, if available.

**Phase 2.** We extracted the list of unique authors of the posts and comments who had a badge. This gave us 1,859 users for SS and 1,383 for SD (ref. Table 7). The distributions of the SS and SD users across the various abstinence stages displayed in the badges are shown in Figure 21. The badge values of these users were eventually used to construct ground truth data on smoking and drinking abstinence, which we will discuss below. We purposefully excluded the users for whom the API did not return any badge value. No badge information meant that we did not know about their smoking or drinking abstinence status at the time of the crawl.

**Phase 3.** For users with badges, we collected their posts, comments, and associated metadata, this time across Reddit. Note that these posts and comments could have been shared on any subreddit, outside of SS/SD. Like before, for every user, the Reddit API limits crawling to the most recent 1,000 posts or comments shared by

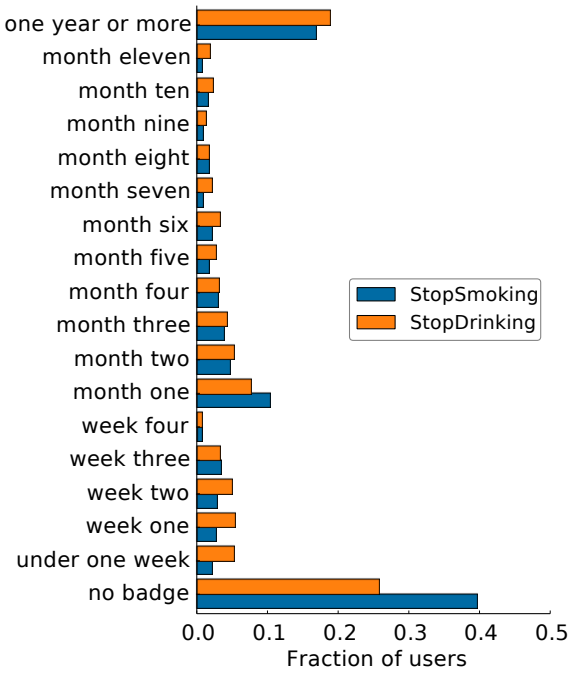



Figure 21: Distributions of the users in StopSmoking (SS) and StopDrinking (SD) across the various smoking and drinking abstinence stages, displayed in the subreddit-specific badges.

the user. Using this method, we obtained 86,835 posts and 766,574 comments for the 1,859 SS users, and 59,201 posts and 492,573 comments for the 1,383 SD users.

We report the summary statistics of the crawled data in the "All data" columns for SS and SD in Table 7. Also important to note here that, per our crawl, each user in the dataset had a recent post or comment in SS/SD, therefore our dataset is likely to be free of any users who stopped being active in SS/SD and do not pay attention to their badges therein.

### 5.3.2 Ground Truth Creation

We constructed ground truth information on smoking and alcoholism abstinence from the crawled badges of the users. Since the badge information is self-reported, we consider it as a reliable, high-quality signal of a user's abstinence status. While characterizing the different abstinence statuses would be insightful, the skewness in

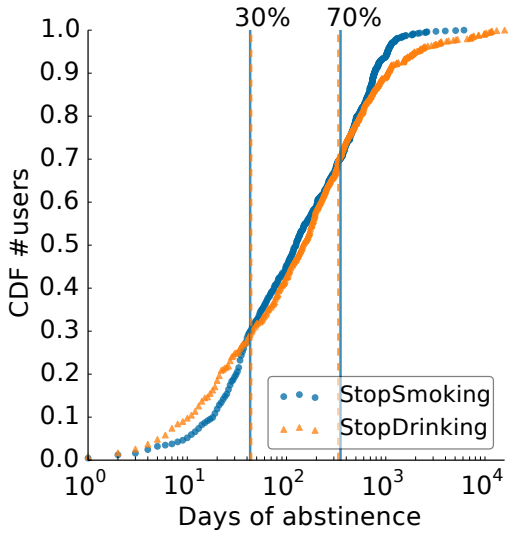



Figure 22: Cumulative distribution functions (CDFs) of the number of users over the abstinence duration (in days) in StopSmoking (SS) and StopDrinking (SD).

the number of users among the different abstinence stages and the sparsity of users per stage (see Figure 21) debarred us from pursuing this direction. Instead, we examined whether we could utilize Reddit activity and interaction of users towards a binary classification task—determining whether a user is likely to belong to the short-term abstinence category or to the long-term abstinence category, given his or her historical data.

To identify the suitable durations to qualify for short-term or long-term abstinence, we leverage the cumulative distribution functions (CDFs) of abstinence duration obtained from the badges in SS and SD (Figure 22). The CDFs show stable patterns before the 30 percentile and after the 70 percentile. The 30 percentile mark for SS is 43 days while it is 44 days for SD; the 70 percentile mark is 350 days and 333 days for SS and SD, respectively. Prior research in addiction [175] indicates frequent relapse to happen at 1-2 months after quitting, which aligns with our 30 percentile mark. Furthermore, individuals who successfully abstain from smoking/alcohol for a year or more have been found to be less likely to relapse in the future [178]. Therefore, we consider the users within the 30 percentile mark to be the short-time abstainers

Table 8: List of the explanatory variables used in the statistical models for StopSmoking (SS) and StopDrinking (SD). SCC and WCC refer to strongly and weakly connected components, respectively.

| Explanatory variables |
|---|
| *Language variables:* |
| counts for the 300 uni/bi/trigrams |
| mean, median PA, NA SS/SD |
| *Addiction variables:* |
| addiction words count |
| mean, median PA, NA OSR |
| *Interaction variables:* |
| #posts, #comments SS/SD |
| #posts, #comments OSR |
| mean, median $\Delta$ between contents SS/SD |
| mean, median $\Delta$ between contents OSR |
| mean, median content scores SS/SD |
| mean, median content scores OSR |
| mean, median content lengths SS/SD |
| mean, median content lengths OSR |
| link, comment karma |
| tenure, recency SS/SD |
| tenure, recency OSR |
| #contents in each of the 15 related subreddits |
| indegree, outdegree, degree |
| reciprocity, #triangles, clustering coefficient |
| betweenness, closeness, eigenvector centralities |
| SCC size, WCC size |

and those beyond the 70 percentile mark to be the long-term abstainers.

This categorization gave us 635 users in SS (318 users/50.07% long-term abstainers) and 533 users in SD (268 users/50.28% long-term abstainers). In the rest of this chapter, we use this user set for the task of characterizing long-term abstinence from tobacco or alcohol. Summary statistics on these users can be found in the "Ground truth data" columns for SS and SD in Table 7.

## 5.4 Statistical Method

We now present the statistical method we employ to characterize long-term abstinence from tobacco or alcohol. For this goal, we introduce the variables outlined below and summarized in Table 8.

### 5.4.1 Response Variable

Our binary response variable represents if a user is a *short-term* or a *long-term* abstainer of smoking/drinking.

### 5.4.2 Explanatory Variables: Language

Our first set of explanatory variables focuses on extracting linguistic attributes from a user's posts and comments in SS/SD. Here, we converted the textual content of all the posts and comments in SS/SD to lowercase and extracted the top-100 most frequent unigrams, bigrams, and trigrams (three sets of 100 items each) following the conventional bag-of-words model.[7] These 300 n-grams do not include any phrase that is solely comprised of stopwords. We introduce a count variable for each n-gram, representing the total number of times that the corresponding n-gram appears in the user's posts or comments.

As another dimension of language, we also consider the sentiment of the posts and comments with VADER [91]. VADER is a lexicon and rule-based sentiment analysis tool that is tailored to specifically detect sentiment expressed in social media. Using VADER, we introduce four variables that correspond to the mean and median of the positive sentiment (PA) and negative sentiment (NA) scores of a user's posts and comments in SS/SD. Together, this set of explanatory variables contains 304 variables and we refer to them as the *language variables*.

### 5.4.3 Explanatory Variables: Addiction

Our second set of explanatory variables focuses on the content (posts or comments) shared by a user in subreddits other than SS/SD (we henceforth refer to this set of subreddits as OSR).[8] To examine if smoking or drinking related content in OSR could

[7]Our statistical models suffered from high dimensionality when we considered more than 300 n-grams.

[8]SD (SS) becomes an OSR when we focus on smoking (drinking).

Table 9: Addiction-related lexicons for smoking and drinking.

| | |
|---|---|
| **Smoking**: | acid, alcohol, baked, blaze, blazed, blunt, blunts, bong, bongs, bowl, bowling, bowls, bud, cannabis, chew, chronic, cig, cigar, cigarette, cigarettes, cocaine, coke, crack, dank, dip, doobie, dope, drug, drugs, drunk, ecstasy, fag, ganja, grass, grizzly, herb, heroin, high, hit, hookah, joint, joints, lsd, marijuana, meth, nicotine, party, piece, pills, pipe, pipes, pot, reefer, ripped, roach, school, sex, shit, skoal, smoke, smokes, smoking, snuff, spliff, stone, stoned, stoner, stones, tobacco, toilet, toke, toking, wasted, weed, fucked up, mary jane |
| **Drinking**: | acid, alcohol, alcoholic, alcoholism, awesome, bar, beer, beers, beverage, booze, boozing, brew, cocaine, cocktail, coke, college, crack, crazy, crunk, dance, dope, drink, drinking, drinks, drug, drugs, drunk, ecstasy, friends, fucked, fun, girls, hammered, hangover, heroin, high, intoxicated, liquor, lsd, marijuana, meth, parties, party, partying, pills, pissed, pong, pot, rave, rum, sex, shitfaced, shot, shots, smashed, smoke, sober, stoned, trashed, up, vodka, wasted, weed, whiskey, wine |

potentially help characterize long-term abstinence, we complied two addiction-related lexicons for smoking and drinking based on words in Urban Dictionary[9]. Urban Dictionary is a suitable choice due to the informal nature of online language. Specifically, we utilized a snowball approach in which we seeded the dictionary searches with "smok*" and "alcohol*". We followed the "related words" returned by the dictionary results on these two seed words. We recursively adopted this approach over three more iterations. The final two lexicons are shown in Table 9. Since a user is unlikely to use every word in the lexicon, we consider a single count variable that represents the total number of times that any of the words in the lexicon appears in the user's posts or comments. We also introduce four variables that correspond to the mean and median of the PA and NA scores of the users' posts and comments in OSR—we again use VADER for this purpose. This set of explanatory variables contains 5 variables and we refer to them as the *addiction variables*.

### 5.4.4 Explanatory Variables: Interaction

Our third set of explanatory variables focuses on the various aspects of interaction.

[9] `www.urbandictionary.com`

1. *Activity measures.* We introduce variables for the number of posts and comments in SS/SD and OSR, the mean and median differences in hours ($\Delta$) between consecutive contents in SS/SD and OSR, the mean and median content scores in SS/SD and OSR, the mean and median content lengths (in characters) in SS/SD and OSR, and the user's link and comment karmas. Also, we include variables that represent the number of days since the earliest and latest contents (tenure and recency, respectively) in SS/SD and OSR.

2. *Participation in related subreddits.* Since abstainers might seek support from or contribute to other subreddits as well, we also extracted the list of the 100 most widely used subreddits, other than SS and SD themselves, based on the posts and comments of the users. Two researchers familiar with Reddit thereafter individually scanned the list to rate their relevance to our task. Researchers referred to prior addiction literature during this task to identify behavioral attributes associated with smoking/alcohol addiction [48]. Subreddits with the following characteristics were deemed relevant—emotional discourse subreddits (e.g., r/depression), religious discourse subreddits (e.g, r/Buddhism and r/atheist), fitness subreddits (e.g., r/Fitness), and subreddits on other types of addiction and recovery (e.g., r/cripplingalcoholism). Abstainers are known to engage to greater emotional expression, including personal and subjective topics like religion [128]. Fitness and exercise are also known to be a helpful characteristic of abstinence [48].

   The final set of related subreddits considered here are shown in Table 10. For each of these subreddits, we introduce a count variable that represents the total number of posts and comments that the user made in the corresponding subreddit.

3. *Graph measures.* To further quantify the interaction between the users in

Table 10: Related subreddits—subreddits other than StopSmoking (SS) and StopDrinking(SD) where users post/comment.

| | |
|---|---|
| **Smoking**: | StopDrinking, electronic_cigarette, BabyBumps, Fitness, relationships, Christianity, personalfinance, atheism, IAmA, MakeupAddiction, SkincareAddiction, loseit, Frugal, Showerthoughts, Buddhism |
| **Drinking**: | REDDITORSINRECOVERY, alcoholism, StopSmoking, relationships, cripplingalcoholism, depression, Christianity, Drugs, CasualConversation, IAmA, atheism, Fitness, MakeupAddiction, electronic_cigarette, DebateReligion |

SS/SD, we leverage a network we construct based on the users' posting and commenting patterns in SS/SD. Specifically, if user A comments on user B's post or comment, we establish a directed edge with a weight of 1 from user A to user B in the network. The total weight of an edge denotes the number of "directed" interactions between the corresponding users. We introduce several graph-centric variables, representing a user's local and global relations with other users in SS/SD: the indegree, outdegree, and degree; reciprocity, the number of triangles to which the user participates (#triangles), and clustering coefficient; the betweenness, closeness and eigenvector centralities; and the number of users in the strongly (SCC) and weakly connected components (WCC) to which the user belongs. Note that for #triangles, clustering coefficient and the centrality measures, we consider an undirected network in which an edge exists only if it appears in both directions in the original network. We refer the reader to [5] for the details of these measures. This set of explanatory variables contains 48 variables and we refer to them as the *interaction variables*.

### 5.4.5 Statistical Models

We employ Ridge regression [83] to classify our binary response variable (short-term or long-term smoking/drinking abstinence). Most of our explanatory variables correspond to English phrases, which posit the collinearity (i.e., excessive correlation

Table 11: Summary of different model fits. Null is the intercept-only model. Deviance measures the goodness of fit. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.01}{3}$).

| | StopSmoking (SS) | | | | StopDrinking (SD) | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Deviance | df | $\chi^2$ | p-value | Deviance | df | $\chi^2$ | p-value |
| Null | 880.3 | 0 | | | 738.9 | 0 | | |
| Language | 438.9 | 304 | 441.4 | $< 10^{-6}$ | 353.5 | 304 | 385.4 | $10^{-3}$ |
| Language + Addiction | 418.5 | 309 | 461.8 | $< 10^{-7}$ | 340.8 | 309 | 398.1 | $< 10^{-3}$ |
| Language + Addiction + Interaction | 326.9 | 357 | 553.4 | $< 10^{-9}$ | 273.2 | 357 | 465.7 | $< 10^{-4}$ |

between phrases) and sparsity (i.e., some phrases occurring infrequently) properties. Ridge regression guards against problems related to collinearity and sparsity by shifting the weights of the correlated and sparse variables to the more explanatory ones. We use 10-fold cross-validation to determine the best tuning constant that controls the strength of the ridge penalty and also to prevent overfitting to the dataset.

To understand the explanatory powers of our independent variables, we consider three statistical models: (i) the Language model, (ii) the Language + Addiction model, and (iii) the Language + Addiction + Interaction model, which consist of (i) the language, (ii) the language and addiction, and (iii) the language, addiction, and interaction variables, respectively. The first two models are motivated from prior work [124, 110], and through the third, we examine the additional role of interaction in characterizing abstinence. In these models, we represent each user as feature vectors that are standardized to zero mean and unit variance.

## 5.5 Results

In this section, we present the results of our two tasks: characterizing long-term abstinence from tobacco and from alcohol.

### 5.5.1 Deviance Results

To evaluate the goodness of fits of our three models, namely Language, Language + Addiction, and Language + Addiction + Interaction, we use *deviance*. Briefly

put, deviance is a measure of the lack of fit to data, hence lower values are better. It is calculated by comparing a model with the saturated model—a model with a theoretically perfect fit, which we consider to be the intercept-only model and refer to as *Null*. Table 11 provides a summary of the different model fits. Due to the randomness introduced by cross-validation, we ran our models 10 times and here we report the results corresponding to the lowest deviances that we obtained in any of the runs.

Compared to the Null models, we observe that all three of our models provide considerable explanatory power with significant improvements in deviances in both SS and SD. The difference between the deviance of a Null model and the deviances of the other models approximately follows a $\chi^2$ distribution, with degrees of freedom equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of Language with that of Null in SS, we see that the information provided by the language variables has significant explanatory power: $\chi^2(304, N = 635) = 880.3 - 438.9 = 441.4, p < 10^{-6}$. This comparison with the Null model is statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.01}{3}$ since we consider three models). We observe similar deviance results for the Language + Addiction and Language + Addiction + Interaction models in both SS and SD, with the latter models possessing the best fits and highest explanatory powers.

From the fits of the Language + Addiction + Interaction models, Table 12 presents the top-30 positive and top-30 negative $\beta$ values for the variables corresponding to the n-grams and the top-7 positive and top-7 negative $\beta$ values for the other variables. The variables with negative and positive $\beta$ values classify a user as short-term and long-term abstainer, respectively. Note that we standardize the feature vectors before regression, hence the $\beta$ values correspond to standardized features. We do not report the statistical significance of the $\beta$ values in the form of $p$-values because they are hard

Table 12: $\beta$ values corresponding to the 74 features with the highest explanatory power for StopSmoking (SS) and StopDrinking (SD). "OSR" stands for subreddits other than SS/SD. The prefix "r/" indicates a related subreddit. "aa" stands for Alcoholics Anonymous.

| StopSmoking (SS) | | | | StopDrinking (SD) | | | |
|---|---|---|---|---|---|---|---|
| feature | $\beta$ | feature | $\beta$ | feature | $\beta$ | feature | $\beta$ |
| indegree | -0.28 | tenure SS | 0.75 | indegree | -0.26 | tenure SD | 0.83 |
| median content length SS | -0.24 | #comments OSR | 0.35 | closeness centrality | -0.20 | #comments OSR | 0.25 |
| degree | -0.23 | tenure OSR | 0.24 | median NA SD | -0.16 | r/REDDITORSINRECOVERY | 0.24 |
| r/Buddhism | -0.18 | mean content score SS | 0.20 | mean NA OSR | -0.16 | mean PA SD | 0.18 |
| recency SS | -0.17 | comment karma | 0.18 | r/Fitness | -0.15 | tenure OSR | 0.16 |
| median NA SS | -0.16 | addiction words count | 0.18 | link karma | -0.15 | #posts OSR | 0.14 |
| outdegree | -0.16 | r/electronic_cigarette | 0.14 | SCC size | -0.15 | r/relationships | 0.13 |
| feature (n-gram) | $\beta$ | feature (n-gram) | $\beta$ | feature (n-gram) | $\beta$ | feature (n-gram) | $\beta$ |
| i started | -0.31 | year | 0.32 | in the past | -0.33 | year | 0.33 |
| i need to | -0.26 | keep it up | 0.27 | i'm going to | -0.31 | i got sober | 0.27 |
| this time | -0.23 | think about it | 0.21 | week | -0.24 | months | 0.25 |
| i'm going to | -0.23 | pack a day | 0.20 | i know i | -0.18 | i quit drinking | 0.23 |
| i want to | -0.22 | i still | 0.19 | i need to | -0.17 | i don't drink | 0.23 |
| as much as | -0.19 | keep it | 0.18 | day | -0.17 | a drink | 0.19 |
| trying to quit | -0.19 | never | 0.18 | i need | -0.17 | meetings | 0.19 |
| thanks for the | -0.19 | since i quit | 0.18 | i feel | -0.16 | find | 0.19 |
| if you don't | -0.18 | if you want | 0.18 | i don't know | -0.14 | was able to | 0.17 |
| in the morning | -0.18 | a year | 0.17 | to quit | -0.14 | years | 0.16 |
| feel like | -0.17 | worked for me | 0.16 | and i don't | -0.14 | as much as | 0.16 |
| i don't want | -0.16 | you want | 0.16 | last | -0.13 | keep up the | 0.15 |
| started | -0.16 | going to be | 0.16 | want to be | -0.13 | stay | 0.14 |
| the last | -0.13 | i would | 0.15 | the first time | -0.12 | stay sober | 0.14 |
| try to | -0.13 | i smoked | 0.15 | have a problem | -0.12 | in the first | 0.13 |
| feeling | -0.13 | hang in there | 0.15 | so much | -0.12 | sobriety | 0.13 |
| last | -0.13 | a non smoker | 0.14 | back to | -0.12 | at a time | 0.13 |
| i want | -0.13 | you'll | 0.14 | don't know | -0.11 | still | 0.13 |
| thanks for | -0.13 | get a | 0.14 | i'm | -0.11 | part of | 0.13 |
| you don't have | -0.13 | you're | 0.14 | i can't | -0.11 | one day at | 0.13 |
| i've | -0.13 | so much | 0.13 | i think i | -0.11 | people | 0.12 |
| right now | -0.12 | keep | 0.12 | i'm not | -0.11 | a time | 0.12 |
| 2 | -0.12 | you don't need | 0.12 | i know that | -0.11 | i was drinking | 0.12 |
| in the past | -0.12 | helped me | 0.12 | i don't want | -0.11 | congrats on | 0.12 |
| in my life | -0.11 | you quit | 0.12 | not drinking | -0.11 | i really | 0.12 |
| to quit smoking | -0.11 | it gets | 0.12 | drinking i | -0.10 | i got | 0.12 |
| i quit smoking | -0.11 | like a | 0.12 | i've been | -0.09 | aa | 0.12 |
| able to | -0.11 | years | 0.12 | thank you | -0.09 | life | 0.12 |
| i got | -0.11 | you want to | 0.12 | i feel like | -0.09 | if you don't | 0.12 |
| as well | -0.10 | a pack a | 0.12 | i want to | -0.09 | you don't want | 0.11 |

to interpret for strongly biased estimates such as those arise from Ridge regression [75].

The contribution of the different explanatory variables in the two characterization tasks is notable. In both, phrases are notable variables that distinguish short-term and long-term abstinence. In fact, the variables that have the highest explanatory

power for short-term abstinence in SS/SD are the phrases *"i started"* and *"in the past"*, respectively. We conjecture that the short-term abstainers use these phrases to indicate new intentions: *"i started an attempt on monday..."* and *"it feels great to be sober and have my dark drinking days in the past"*, respectively. Furthermore, the phrases associated with short-term abstinence are related to current sensation, urge, or confession ( *"i need to", "i feel"*), and appreciation and acknowledgement of support, perhaps because they are newcomers in the community ( *"thanks for the", "thank you"*). E.g., notice the post excerpt below:

> *i need to find more friends that don't drink so much*

In contrast, the phrases associated with long-term abstinence are mostly about encouragement and boosting morale ( *"keep it up", "hang in there"*) and advisory ( *"worked for me', "was able to"*):

> *for those of you behind me, keep it up! i believe in you!*

Examining some of the non-phrase variables with negative $\beta$ values, we observe that indegree is a strong indicator of short-term abstinence. This is likely because the short-term abstainers' contents are typically support-seeking in nature, which attract responses from a variety of users in the SS/SD communities. The negative sentiment of contents is also a significant indicator of short-term abstinence. We conjecture that this is likely due to the tendency of the short-term abstainers' disclosures about recent failures, challenges, and struggles related to quitting. Addiction literature also indicates that increased negative affect and stress are associated with early abstainers of smoking/drinking [153]:

> *i [...] struggle with depression and used alcohol to escape from my often difficult reality*

Table 13: Performance metrics corresponding to the three statistical models for StopSmoking (SS) and StopDrinking (SD).

| | Language | | Language + Addiction | | Language + Addiction + Interaction | |
|---|---|---|---|---|---|---|
| Measure | StopSmoking | StopDrinking | StopSmoking | StopDrinking | StopSmoking | StopDrinking |
| F1 score | 0.70 ± 0.06 | 0.78 ± 0.04 | 0.78 ± 0.05 | 0.80 ± 0.05 | 0.86 ± 0.03 | 0.85 ± 0.05 |
| Accuracy | 0.74 ± 0.05 | 0.81 ± 0.04 | 0.80 ± 0.04 | 0.81 ± 0.04 | 0.86 ± 0.03 | 0.85 ± 0.05 |
| Precision | 0.81 ± 0.06 | 0.91 ± 0.07 | 0.83 ± 0.05 | 0.91 ± 0.06 | 0.90 ± 0.04 | 0.88 ± 0.06 |
| Recall | 0.62 ± 0.09 | 0.69 ± 0.06 | 0.74 ± 0.06 | 0.71 ± 0.07 | 0.82 ± 0.04 | 0.83 ± 0.06 |
| Specificity | 0.86 ± 0.04 | 0.93 ± 0.04 | 0.86 ± 0.05 | 0.93 ± 0.05 | 0.91 ± 0.04 | 0.88 ± 0.05 |

Focusing on some of the non-phrase variables with positive $\beta$ values, we observe that tenure in SS/SD and OSR are strong indicators of long-term abstinence. Prior work has indicated that long-term social engagement has a positive impact on the psychological states of individuals [56]. Hence, we conjecture that longer tenure on Reddit helps keep individuals intending to abstain from smoking/drinking more motivated and focused towards their respective self-improvement goals. Furthermore, users' comment karma characterizes long-term abstinence in SS, suggesting that social endorsement obtained from the greater Reddit community in the form of upvotes possibly motivated individuals to succeed in their abstinence goals.

We also see that the mean content score in SS and the mean positive sentiment of contents in SD are strong indicators of long-term abstinence from smoking and drinking, respectively, which are likely related to the supportive tone expressed in such content. Addiction literature indicates social support to act as a mediator of stress during smoking/drinking urges [153]. E.g., the following excerpt expresses positive sentiment:

> *every time when i remember i quit smoking it makes me happy and a little proud*

### 5.5.2 Classification Results

To evaluate how well our three statistical models distinguish the long-term and short-term abstinence categories, we randomly split the dataset into 90% training and 10%

testing partitions. We trained our models only on the training partitions and measured their classification performance on the testing partitions. Due to the randomness introduced by cross-validation, we performed the aforementioned procedure 10 times to obtain accurate performance estimates. Assuming that long-term abstinence is our positive class, Table 13 presents the classification results with respect to the F1 score, accuracy, precision, recall, and specificity metrics. We report for each metric the mean and standard deviation of the 10 values that we obtained from the 10 iterations on the testing sets.

In general, we observe that the best performing model in both SS and SD is Language + Addiction + Interaction, which achieves the mean F1 scores of 0.86 and 0.85 in SS and SD, respectively. Considering the minimum of the values for SS and SD, this model also achieves a mean accuracy of 0.85, a mean precision of 0.88, a mean recall of 0.82, and a mean specificity of 0.88. This model is followed by Language + Addiction and then Language in terms of performance. Not only the mean values of the performance metrics for Language + Addiction + Interaction are higher than those for the other two models, the ranges of the values are also narrower in Language + Addiction + Interaction (lower standard deviations).

The good performance of Language + Addiction + Interaction is also evident from the receiver operator characteristic (ROC) curves in Figure 23. To obtain the ROC curves, we first sorted the probabilities that the users are long-term abstainers as output by the models in ascending order. We then generated 250 threshold points equidistant in the range [0, 1] and applied them on the probabilities of the users in the testing partitions; for each threshold value, all users with probabilities above that value are labeled as long-term abstainers, or short-term abstainers otherwise. This process generated 250 pairs of true positive (TP) rate and false positive (FP) rate values for each testing partition, plotting the average of the 10 TP rate and FP rate values computed using the same threshold value across the 10 experiments on the

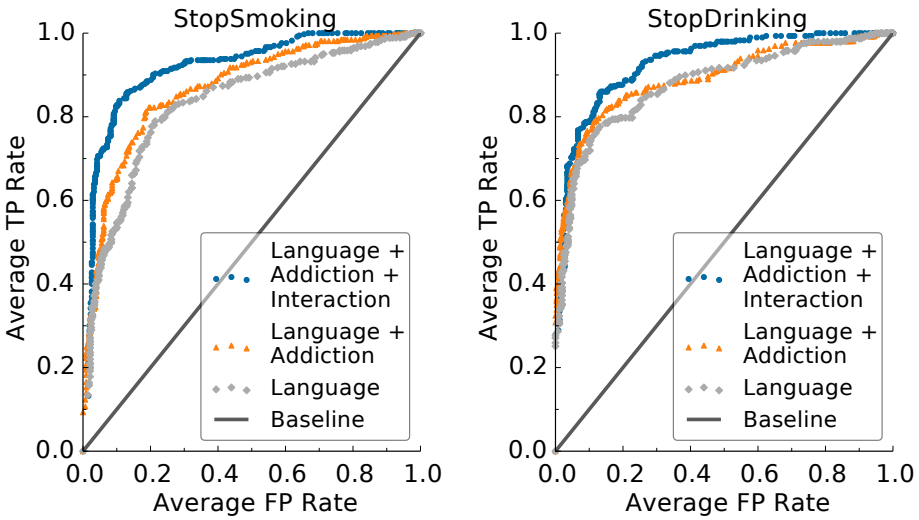



Figure 23: Receiver operating characteristic (ROC) curves showing average true positive (TP) and false positive (FP) rates corresponding to the three statistical models for StopSmoking (SS) and StopDrinking (SD). Long-term abstinence is the positive class.

testing partitions gave us the ROC curves in Figure 23. We observe from the figure that the performance of Language + Addiction + Interaction is superior to the other two models in both SS and SD in the whole spectrum of the average TP rate and FP rate values.

## 5.6 Discussion

### 5.6.1 Clinical Relevance

Our findings indicate that linguistic and interaction cues gleaned from activity in SS and SD forums may be used to understand short-term or long-term abstinence tendencies among users. Such ability to proactively identify one's abstinence status may be used to create early warning systems or interventions that are integrated in social platforms. These early warning systems could analyze one's activity on the platform and engage appropriately if the probability of long-term abstinence drops below a certain level. Certainly, such systems could raise ethical and privacy concerns, and must therefore be carefully designed and developed. However, if successful, these systems

may be used in clinically meaningful ways that provide great benefits. For instance, an individual may more easily keep track of his or her activities and interactions on a social media platform and share them with a therapist, which may subsequently lead to more effective treatment.

Broadly, tracking the patterns of changes in the explanatory variables we identified could help clinicians, medical professionals, and policy makers better understand people's experiences around long-term abstinence from tobacco or alcohol, and the strategies that may have worked for them. Since, traditionally, it has been challenging to understand and identify factors associated with long-term smoking or drinking abstinence [175], our research can also help identify previously underexplored variables that may contribute towards the success or failure of abstinence.

Finally, and importantly, through our statistical models that identify short-term and long-term abstainers, we can begin to determine the abstinence status of those individuals for whom badge or other self-reported information on abstinence is not available. This can be particularly valuable in bringing in-time help and support to individuals who intend to quit smoking or drinking and use a social media platform, however have not adopted the practices of accruing badges, imbibed in the two online communities we study.

### 5.6.2 Implications for Social Media Research

**Design Considerations.** We believe our findings have strong design-related implications for social media research. Below, we describe several design ideas inspired by our research, which may help tailor social media platforms to cater to individuals aiming to abstain from smoking or drinking. Literature indicates that individuals desirous of quitting smoking or drinking often go through repetitive phases of cessation and relapse [73]. Hence, new users joining these abstinence communities, or those who have been short-term abstainers may benefit from content on the forum

that discusses the challenges and struggles in this early phase. Mechanisms could be created to engage in a conversation with other long-term members on what to expect during this phase, how to combat desires of smoking or drinking urges, or for general positive reinforcement of their abstinence goal.

Post excerpts containing phrases and other linguistic constructs associated with long-term abstinence may also be promoted to users intending to quit smoking or drinking. They may also be directed to connect with other users in the community who have had success in tobacco or alcohol abstinence over a period of time—social support and higher levels of social capital have been known to help individuals fight addiction urges [72]. Moderators of these recovery communities may also direct requests for advice or help to appropriate users in the community who are actively engaged and have had experiences of long-term abstinence. Since we also found that posting activity or commentary in certain other subreddits were associated with long-term abstinence, users may also be recommended to participate in those other communities or forums where they might additionally obtain support for beating addiction urges or gather general positive reinforcement of their desire to abstain from smoking or drinking.

In addition, our work showed that network features derived out of the social interaction offered considerable explanatory power. That is, the presence of a strong support network on the forum is likely to play an important role in encouraging long-term abstinence. As a design idea, newcomers' posts could be promoted to prominent positions in the forums' timelines to attract more attention, increasing their likelihood of receiving responses. In turn, this would broaden engagement of the whole community, decrease user churn, and thereby increase member retention. This could lead to a self-reinforcing positive cycle that attracts and helps increasingly more people.

Furthermore, in these Reddit communities, reputation is associated with "badges"

that indicate the duration of abstinence of a user from smoking/drinking. In a way, making such badge information accessible to visitors and users of the forum not only is likely to boost self-esteem because of improved reputation in the community, but also in general, is likely to induce positive feelings towards abstinence, and encourage and inspire others to do so as well.

**Uniqueness of Reddit.** We also discuss the effectiveness of addiction recovery communities like SS or SD in general. Although many online communities exist to help individuals in addiction recovery, SS and SD are unique because they encourage long-term abstinence. This is indicated by the fact that almost 50% of the users in our dataset were abstainers for three or more months. We thus believe that participation in these Reddit forums are likely to help individuals adopt a positive attitude and approach towards addiction recovery. Moreover, the ability to be anonymous or pseudonymous can be an additional facilitating element of abstinence—Reddit accounts do not need any personally identifiable information. Users can thus engage in candid and honest discourse, without worrying about the social stigma that often comes with being a victim of addiction. In fact, a considerable fraction (10%) of users in our dataset explicitly only posted on these two subreddits, perhaps indicating that either they are on Reddit simply to participate in these abstinence forums, or have alternate account(s) on Reddit for non-addiction recovery related discourse. Also, even though some of the explanatory variables that we consider in our statistical models are Reddit-specific, our statistical models can be generalized to other social media platforms, especially to those that possess similar attributes implicitly or explicitly (e.g., link karma on Reddit vs. number of retweets on Twitter as a manifestation of a user's reputation on the online platform).

### 5.6.3 Limitations

Our work is of course not free from limitations. We acknowledge that generalizations of our work might not be easily applied across large populations or on arbitrary addiction contexts. As we pointed out, SS and SD are specialized self-improvement communities; most likely, individuals who choose to join them are already motivated to quit addiction. Moreover, since these are largely communities of abstainers, it is possible that individuals new to quitting may feel uncomfortable joining the communities or can feel uncomfortable to be participating. Further biases inherent to Reddit exist as well—the average redditor is a 20-something male[10], perhaps more "tech-savvy", and therefore more likely to resort to online platforms to obtain abstinence support compared to the general population. Additionally, since we did not have information on whether the long-term abstainers sought support through offline means, we are limited in the way we evaluate the effectiveness of the particular forums for addiction recovery. We also note that we focused on smoking and drinking addiction recovery, obviously extending our findings to other kinds of addiction (e.g., prescription or recreational drugs) would need additional investigation.

As we also pointed out earlier, an important point to note about this work is that we do not *predict* abstinence of individuals in SS/SD. That is, based on our findings, we are not able to make (causal) claims as to whether someone will continue to abstain smoking or drinking in the future, or will relapse. This requires tracking an individual's activity and their abstinence reports, i.e., the badge values, over time. In prior literature on clinical studies of addiction behavior, use of survival analysis methods have been found to be particularly helpful in forecasting the likelihood of experiencing a relapse. We leverage these statistical approaches in Chapter 6 to predict smoking or drinking relapse based on social media activities.

[10] `www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_reddit_usage_2013.pdf`

We also note that a known concern with many recovery communities is member retention—failure to recover often demotivates individuals and leads them to leave the platform. While it is challenging to measure the overall retention rate for SS and SD based on our data, the focus on both self-reported abstinence information through badges and the users who had a recent post or comment in SS/SD ensures that we consider a population of individuals who are attempting to abstain from smoking/drinking and continuing to use Reddit. Also, as mentioned earlier, in our ground truth dataset, we had nearly 50% users who are short-term abstainers. However, per our current data, we cannot be sure of the nature of such short-term abstinence—i.e., whether individuals were attempting to quit smoking/drinking for the first time, or it followed a recent relapse experience. This is because Reddit's API allows our program to access only the *current* badge of a user. Hence, we were not able to determine the nature of short-term abstinence of users in our dataset. For instance, we do not know if they had relapsed shortly before, or if they are attempting to quit for the first time. Finally, as Reddit also imposes that only the most recent thousand posts and comments of every user may be retrieved, we were limited in how far back we could go to examine redditors' historical activity.

## 5.7 Conclusions

We presented a computational framework to understand smoking and drinking abstinence of individuals from social media. We compiled and studied a previously unexplored source of data—activity on the Reddit communities StopSmoking and StopDrinking. We leveraged the badge feature in these forums to construct self-reported ground truth information on the abstinence status of users to characterize long-term abstinence. Our statistical models incorporated a variety of language and interaction attributes to distinguish long-term abstinence from smoking or drinking from short-term abstinence with 85% accuracy. We found that linguistic cues like

affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence. Through our findings, we provided insights into how social media may be leveraged to tackle addiction-related health challenges.

# CHAPTER VI

# CHARACTERIZING SMOKING AND DRINKING RELAPSE FROM SOCIAL MEDIA

Alcohol and tobacco are among the top causes of preventable deaths in the United States [120]. Achieving long-term abstinence of tobacco or alcohol is difficult [175]—most abstainers are known to relapse within one to three months of cessation. Prior work examining addiction behavior manifested on social media investigates mainly the role of linguistic attributes in characterizing health challenges related to addiction [124, 110]. Also, these pieces of research use crowdsourcing to obtain information about the abstinence status of the individuals. However, simply looking at social media posts may not always allow third-party judges to reliably capture abstinence status.

In our work, which consists of two parts, we focused on two prominent smoking and drinking cessation communities on the social media site Reddit: StopSmoking and StopDrinking. These communities are identified as "self-improvement communities" on Reddit and are geared toward providing support and motivation to smoking and drinking addiction sufferers. A unique aspect of these communities is that they allow the users to acquire "badges". Badges are a mechanism by which the users can self-report the duration of their abstinence. We collected data on the users' badges, posts, comments, and associated metadata from these communities, and developed statistical models to analyze the role of social media language, interactions, and engagement in characterizing smoking/drinking abstinence and relapse. Addiction literature indicates social support to act as an important mediator of stress during smoking/drinking

urges [153, 72]. In this context, graphs enable us to capture the interactions and engagement between the users, which reflect access to social support. Specifically, our models leverage a graph that represents which user provides social support to whom by writing comments on their posts in the communities. In summary, through our work, we extend the existing body of research by using self-reported abstinence information on smoking and drinking, and examining the additional role of interaction and engagement in characterizing these addiction-related health challenges.

The first part of our work, which we present in Chapter 5, focuses on characterizing abstinence from smoking and drinking. We used the badges of 1,168 users to construct ground truth information on short-term (<40 days) and long-term (>one year) abstainers, and we formulated and identified the key linguistic and interaction characteristics of these abstainers based on activity in the communities spanning eight years, from 2006 to 2014. We developed supervised learning-based statistical models based on these characteristics to distinguish long-term abstinence from short-term abstinence with over 85% accuracy. We found linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence.

The second part of our work, which we present in this chapter, focuses on characterizing relapse to smoking and drinking. Here, we used longitudinal data on the badges of 5,991 users to determine their abstinence or relapse status, and we formulated and identified the key engagement and linguistic characteristics of the abstainers and relapsers based on activity in the communities spanning almost nine years, from 2006 to 2015. We developed a robust statistical methodology based on survival analysis to examine how participation in the communities and the characteristics above relate to the risk of relapse. Our results show that although participation in the communities is not linked to high likelihood of smoking/drinking abstinence during the one/two months post-cessation, it shows a stable trend of heightened chance of

abstinence beyond three years, suggesting the efficacy of the communities in preventing relapse in the long term. Furthermore, we found positive affect and increased engagement to be predictors of abstinence.

The two parts of our work differ from each other in terms of the problem statement, the statistical method, and the dataset as follows. (1) The first part focuses on characterizing attributes of short-term and long-term abstinence from smoking/drinking. The second part focuses on modeling relapse events self-reported by individuals, and how they, collectively, might indicate the effectiveness of the communities in preventing relapse. (2) The first part uses a supervised learning-based statistical technique. The second part identifies the limitations of such supervised learning techniques in analyzing relapse events, and employs techniques from the survival analysis literature. (3) The first part considers a dataset with one badge per user. The second part expands this dataset with a unique method to obtain daily badges, and considers a dataset with multiple badges per user to determine the relapse events of the users.

## 6.1 Introduction

Addiction challenges, especially to legal substances like tobacco and alcohol, constitute the third leading cause of preventable death and disability in the United States [153]. Tobacco and alcohol use are critical substance abuse problems and kill far more people than all other substance use, homicides, suicides, motor vehicle accidents, and risky sexual behaviors combined [82]. However, maintaining abstinence from tobacco or alcohol is difficult [175]. Research indicates that 80-90% of those who attempt to quit smoking or drinking relapse within a year of their quit dates [73]. In fact, a study found that those who relapse following an attempt to quit had a 95% probability of resuming their regular pattern of smoking/drinking [153]. Hence, there is a rich body of research on identifying precipitants of short-term smoking or drinking cessation [153, 178, 124]. However, limited research provides robust statistical and

empirical insights into cues that may be associated with abstinence or relapse in the longer term. This is largely because of the difficulty in recruiting individuals identified with this stigmatized health behavior as well as the practical, ethical, and monetary challenges of long-term tracking of abstinence and relapse experiences [155, 58, 19].

Use of social media platforms and online communities has been found to be linked to improved self-efficacy and well-being, including facilitating recovery from health challenges [71, 127]. Research has indicated that these platforms provide a constantly available and conducive source of information, advice, and psychosocial support, as well as foster positive behavior change [111, 89]. In the context of substance abuse and addiction, recent research has been able to identify cues of social media behavior and affect associated with abstinence and relapse [122, 124]. Empirical investigations and quantitative evidence on *how* participation in social media communities may relate to tobacco/alcohol addiction cessation are, however, limited.

In this work, we address gaps in prior work by examining how activity in an addiction cessation social media community may be used to analyze smoking and drinking relapse events. Thereby, we explore the efficacy of the community in preventing relapse in the long term. Our motivation lies in the observation that the social environment and other psychological influences have particularly been found to play critical roles in tobacco and alcohol cessation [72]. Therefore, analysis of participation, engagement, and linguistic constructs of content shared in social media support communities are likely to provide insights relating to one's health outcomes and well-being status. We focus on two specific research questions:

**RQ 1:** How is participation in social media communities that provide support toward smoking and drinking cessation associated with the risk of relapse? Additionally, based on participation in these communities, can we infer the likelihood of relapse over time?


Figure 24: Screenshots from the StopSmoking and StopDrinking subreddits, showing example post topics and abstinence badges. The badge icon contains the abstinence stage (e.g., star-shaped smiley face for "one year and beyond"), while the actual number of days of abstinence is reported next to it (e.g., 365 days). The usernames are blurred for anonymity.

**RQ 2:** Are engagement (e.g., receiving extensive feedback from others) and linguistic constructs of content shared (e.g., expressing greater positive sentiment) within these communities predictors of likelihood of relapse to smoking/drinking?

We focus on two prominent smoking and drinking cessation communities on the social media site Reddit: StopSmoking[1] and StopDrinking[2]. These two communities are identified as "self-improvement communities" on Reddit and are geared toward providing support and motivation to smoking/drinking addiction sufferers. A unique aspect of these communities that makes them suitable for our investigation is that they allow individuals seeking help and support on smoking/drinking cessation to acquire "badges" (see Figure 24). Badges are a mechanism by which individuals can self-report the duration of their smoking/alcohol abstinence. The badges are set up to be updated automatically everyday unless a user reports a relapse or a change to

[1] `www.reddit.com/r/StopSmoking`
[2] `www.reddit.com/r/StopDrinking`

their abstinence status.

The main contribution of this work revolves around the study and analysis of relapse and abstinence experiences of over 14 thousand individuals from these two Reddit communities, based on their self-reported badge information. Specifically:

- We devise a methodology to collect longitudinal data on a user's badges in these communities, and thereafter use the badges to identify addiction abstinence or relapse status.
- We employ a robust statistical methodology based on survival analysis [80] to estimate the likelihood of experiencing a relapse event—this method is suitable for analyzing data like ours where the outcome variable is the time until the occurrence of an event of interest (i.e., relapse).
- We formulate and identify key engagement and linguistic characteristics of abstainers and relapsers based on participation in the communities spanning almost nine years, from 2006 to 2015.

Our results present a number of significant insights that may help researchers better understand the role of social media participation in tobacco or alcohol relapse and abstinence. We find that the likelihood of experiencing a relapse to smoking/drinking within a day of abstinence is very high; 45%/33% of individuals in the communities we study are estimated to undergo this event. The median survival time is 25/56 days for smoking/drinking, i.e., half of the population is projected to relapse within about one/two months from start of our study. However, the rate of survival improves significantly beyond three years, suggesting the potential of the communities we study for sustaining cessation among those who do not relapse for a considerable amount of time. Finally, we observe that the linguistic constructs used by the Reddit users in their posts and comments as well as their interaction patterns that capture access to social support are important predictors in preventing relapse.

We discuss the role of social media communities in acting as mediators supporting addiction cessation and the implications for designing timely, adaptive interventions towards promoting sustained health recovery.

## 6.2 Prior Work and Our Differences

### 6.2.1 Addiction Cessation and Relapse

What factors and precipitants lead to addiction relapse (e.g., smoking or drinking) have invited the interest of behavioral scientists and addiction researchers for decades [108, 175]. Typically, such risk factors are categorized into affective, behavioral, cognitive, and social antecedents. The prevailing theory is that stress and cognitive impairment increase the likelihood of relapse, while social and emotional support tend to act as buffers toward mitigating urges to relapse [128, 97].

However, since there is a direct clinical implication around issuing just-in-time interventions to prevent relapse [103, 154], the vast majority of existing efforts have focused on identifying the near real-time antecedents of a relapse [153]. Limited research exists in understanding factors that may be associated with promoting abstinence (and preventing relapse) in the long term. Quantification of these factors is equally important, as they can help evaluating ongoing public health interventions and the design of smoking or drinking cessation programs. An exception is the work of Christakis et al. [41] where the size and structure of individuals' social networks were analyzed to find that their connections and interactions relate to reduced smoking tendencies in the long term (also [45]). Similarly, other work has found that access to a strong, trusting network of friends can provide practical and emotional support toward maintaining abstinence [12, 48].

Most of the above studies are, however, retrospective [134]. They identify risk factors in a post-hoc manner based on survey data and retrospective self-reports about mood and observations about relapse episodes. This method limits temporal

granularity as it involves recollection of historical facts. Prospective or predictive studies analyzing abstinence and experiences of relapse to smoking/drinking, especially over long periods of time are limited [130]. This is because most of them rely on individuals to actively volunteer and provide self-reported information about their addiction status, making compliance over time not only difficult, but also expensive. Furthermore, since tobacco addiction and alcoholism are stigmatized [58, 19], subject recruitment from the general population is a challenging task. For instance, most prior studies have focused on the 4-5% of smokers who attended smoking cessation clinics or reached out to a counseling hotline [115].

In this work, we leverage participation of individuals in a support community on the social media site Reddit to address some of the above challenges. Longitudinal large-scale data obtained from social media allows us to assess the likelihood of relapse or abstinence over a long period of time. By identifying how participation, engagement, and the nature of content shared relate to relapse, we are further able to explore the role played by an online support community in improving self-efficacy toward long-term abstinence.

### 6.2.2 Online Health Communities, Recovery and Coping

People afflicted by medical conditions often find support via online health communities [64, 147, 127]. One study suggests that 30% of the U.S. Internet users have participated in medical or health-related groups [94]. Besides support, these communities serve a range of purposes, including seeking advice [94], connecting with experts and individuals with similar experiences [64, 156, 81, 89], sharing questions and concerns around treatment options [64], sensemaking [112] and understanding professional diagnoses [141], enabling better management of chronic health conditions [113, 89, 90], and fueling discussions with healthcare providers [64]. In this light, approaches to community building have been proposed, e.g., [77, 176], and the

role of participation in such communities toward promoting ailment recovery and coping has been examined in a number of different domains, such as cancer and diabetes [156, 86, 111]. Taken together, this rich body of work supports the notion that people struggling with smoking or drinking cessation may benefit from participation in support communities online, which we examine in this work.

### 6.2.3 Social Media and Inference of Health Status

Recent research in social computing has been able to utilize the abundant and growing repository of social media data to provide a new type of "lens" into inferring health and well-being status of individuals and populations, such as influenza and depression [139, 55, 57, 56, 85, 171]. A common observation in these works has been that social interactions and linguistic constructs of content shared by individuals could be utilized toward building robust computational inference frameworks of health risk. Our work builds on this direction by examining to what extent participation, engagement, and attributes of linguistic expression in a social media support community could signal relapse to smoking or drinking.

Although limited, there has been some recent work examining social media cues associated with addictive behaviors, including tobacco use and prescription drug use [122, 23, 125]. Murnane and Counts [124], for instance, found that among individuals who announced an intent to quit smoking on Twitter, relapsers expressed more negative sentiment compared to those who ceased their smoking behavior during the time of the study. The predictive ability of these cues toward relapse or abstinence was, however, not explored. MacLean et al. [110] adopted a method similar to [124] to study a prescription drug abuse recovery community. They were able to identify linguistic attributes of individuals in various phases of recovery, where recovery stages were identified through crowdsourcing techniques. Finally, in our work [164], which we present in Chapter 5, we examined Reddit support communities to characterize

attributes of short and long-term abstinence from smoking and drinking but did not examine factors that can be predictive of risk to relapse in the long term.

While the latter two pieces of work did demonstrate some predictive capability of the identified cues in inferring relapse or abstinence, their methodology is inadequate to estimate long-term trajectories of likelihood of relapse or abstinence. We extend this body of work by (1) utilizing self-reported information on abstinence or relapse status of individuals in a support community, and (2) developing a robust statistical methodology, motivated from the survival analysis literature, to explore how participation in the communities we study is related to relapse events over time.

## 6.3 Data

Towards our research goals, we focus on obtaining data from two communities in the popular social media Reddit: StopSmoking and StopDrinking, both of which are considered self-improvement subreddits. We refer to them as SS and SD, respectively, through the rest of the chapter. Both subreddits host *public* content that can be viewed without a Reddit account. As mentioned above, they are support communities for individuals intending to quit tobacco or alcohol abuse, garner thousands of subscribers, and have been examined in our prior work [164], which we present in Chapter 5, to study patterns of tobacco and alcoholism cessation.

**"Badges" as Proxies of Abstinence Progress.** A key aspect of these subreddits is that they allow users to acquire "badges" to help track their abstinence progress (see Figure 24). Such badges are subreddit-specific and are displayed next to the username whenever the user posts or comments on the subreddit (ref. Figure 24). Typically, a user makes a badge request to the moderators of the subreddit he or she is interested in through the subreddit's interface or by privately messaging the moderators. Badges are then awarded by the subreddit moderators either manually (SD) or automatically through an application known as "badgebot" (SS). In the

absence of direct user interaction, we utilize the information displayed via the badges as a proxy for self-reported ground truth data on abstinence.

### 6.3.1 Data Collection

Our data collection proceeded as follows. In our prior work [164], which we present in Chapter 5, we used Reddit's official API[3] to obtain a dataset containing users' posts, comments, and associated metadata from SS and SD. In total, we had data for 1,859 SS users (86,835 posts and 766,574 comments) and 1,383 SD users (59,201 posts and 492,573 comments). This crawl also gave us the most recent badge value of each user, i.e., the badge value as displayed on the day of crawl, which was dated November 23, 2014. The drawback of Reddit's API is that it does not provide information about the historical badge values of a user. As we are interested in characterizing and analyzing the temporal patterns of relapse events in SS and SD in this work, we devised a method to obtain longitudinal (daily) data on the badge values for each user in the dataset, going forward from November 2014.

**Longitudinal Data on Badges.** Specifically, we created two "user dictionaries" containing the author IDs of the existing SS and SD users in the dataset, and built a badge value dataset by performing daily crawls on each user for the next five months, from November 24, 2014 to April 23, 2015. The Reddit API limits crawling historical posts on a subreddit to the past thousand posts, so to capture new SS/SD content, each day we obtained the most recent thousand posts and their associated comments in SS and SD, and we stored the new posts or comments in a data batch. For each post, we collected its title, body or textual content, ID, timestamp, and author ID. We collected the same information for each comment on the post as well. We included any new user (author of a new post or comment) that we observed during the daily crawls to the corresponding user dictionary. If the API did not return a badge value

[3] `www.reddit.com/dev/api`

Table 14: Summary statistics of the crawled dataset ("All data" columns) and the dataset used in the statistical models ("Survival data" columns). $\mu$ and $\sigma$ correspond to the mean and standard deviation, respectively. The post and comment lengths are reported in words.

| | StopSmoking (SS) | | StopDrinking (SD) | |
|---|---|---|---|---|
| | All data | Survival data | All data | Survival data |
| Users | 7,221 | 2,917 | 7,224 | 3,074 |
| Total posts from users | 372,414 | 163,480 | 285,055 | 133,887 |
| Total comments from users | 3,424,350 | 1,496,799 | 2,907,379 | 1,333,245 |
| Date of earliest post | Aug. 02, 2006 | Aug. 02, 2006 | Feb. 18, 2006 | Feb. 18, 2006 |
| Date of earliest comment | Aug. 02, 2006 | Aug. 18, 2006 | Jul. 09, 2007 | Jul. 09, 2007 |
| Date of latest post | Apr. 23, 2015 | Apr. 23, 2015 | Apr. 23, 2015 | Apr. 23, 2015 |
| Date of latest comment | Apr. 23, 2015 | Apr. 23, 2015 | Apr. 23, 2015 | Apr. 23, 2015 |
| $\mu \pm \sigma$ \| median comments per post | $7.33 \pm 7.21$ \| 5 | $7.59 \pm 7.41$ \| 6 | $11.70 \pm 12.85$ \| 8 | $11.85 \pm 13.35$ \| 8 |
| $\mu \pm \sigma$ \| median post length | $50.18 \pm 109.83$ \| 14 | $49.91 \pm 108.25$ \| 14 | $71.66 \pm 162.98$ \| 16 | $65.10 \pm 133.80$ \| 16 |
| $\mu \pm \sigma$ \| median comment length | $31.31 \pm 51.62$ \| 16 | $32.02 \pm 52.56$ \| 16 | $33.71 \pm 57.34$ \| 16 | $34.41 \pm 55.43$ \| 17 |

for a user, we assigned a special badge value of "NA" to the user.

**Historical Activity on Reddit.** Additionally, we collected each user's historical activity on the platform, i.e., posts, comments, and associated metadata, shared in subreddits beyond SS and SD, and we stored the new posts or comments in a separate data batch. We henceforth refer to this set of subreddits as OSR (Other SubReddits).

**Summary Statistics.** We report the summary statistics of the final crawled dataset in the "All data" columns for SS and SD in Table 14. Figure 25 shows the cumulative distribution functions (CDFs) of the abstinence duration obtained from the badges in SS and SD at the end of the data collection period. These CDFs exclude 3,838 users/53.15% in SS and 3,548 users/49.11% in SD whose observed, final badge values were NA (i.e., they did not have a badge value on the last day of the crawl). We observe from the figure that the majority of the users abstained for either a short period of time (less than a week) or a long period of time (more than a year), in essence they are bimodal distributions. It is important to note that, per our crawl, each user had at least one recent post or comment in SS/SD, therefore our dataset is

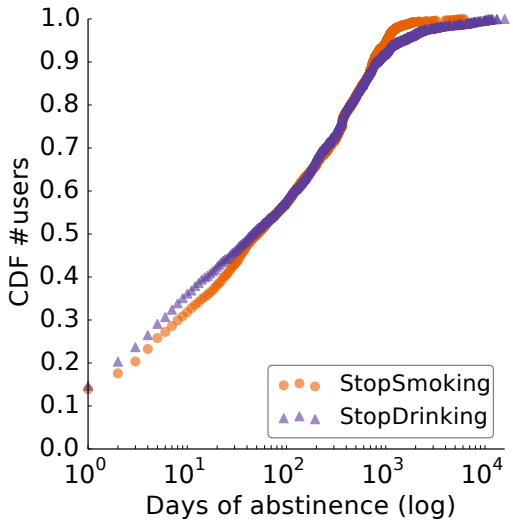



Figure 25: Cumulative distribution functions (CDFs) of the number of users over the abstinence duration (in days) in StopSmoking (SS) and StopDrinking (SD), leveraging the badge values at the end of the data collection period.

likely to be free of users who are no longer active in SS/SD.

### 6.3.2 Capturing Abstinence Success and Failure from Badges

Now, we discuss how we measure smoking/drinking abstinence success and failure from the longitudinal (daily) badge values of the users. We first used the collection of the daily badge values of a user to establish a *badge sequence* for the user. Figure 26 shows several example badge sequences. We defined the abstinence and relapse events from smoking/drinking based on the badge sequences of the users as follows:

- ***Abstinence.*** We assumed that the users with strictly increasing badge sequences have successfully abstained from smoking/drinking during our time period of analysis.

- ***Relapse.*** We assumed that the badge sequences of users who experienced a relapse will be characterized by either (a) an increasing badge sequence with a sudden drop, or (b) a badge sequence with a repeating badge values of 1 (this case captures the users who relapsed on their first day of abstinence).

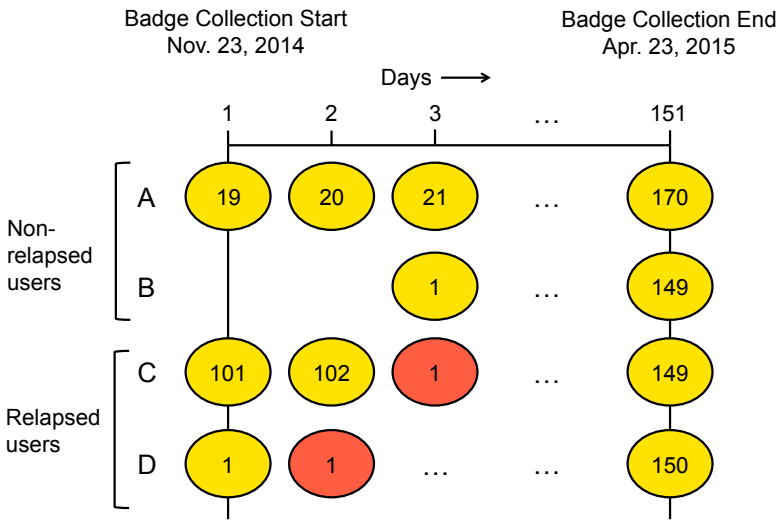



Figure 26: Example badge sequences (rows) obtained from the collection of the daily badge values (values inside the circles) of the users. Users $A$ and $B$ have strictly increasing badge sequences, indicating successful abstinence, whereas the badge sequences of users $C$ and $D$ have a drop (102→1) and a repeating badge values of 1, respectively, which indicate a relapse (highlighted in red).

However, our preliminary analysis of the badge sequences revealed a few points to consider for our subsequent statistical analysis. Specifically, these considerations were important to handle gaps in self-reporting of the badge values by the users.

(1) *Missing badge values.* There were users with missing badge values in our dataset. The badge sequences of 3,342 users/46.28% in SS and 2,994 users/41.45% in SD consisted of only NA values. No badge information means that we do not know about the smoking/drinking abstinence statuses of these users and, hence, they were disregarded.

(2) *Sparse badge values.* A related point comprises the users with few badge values. As we continued to include new users in our dataset during the daily crawls, for those users admitted shortly before the data collection period ended, we were able to collect only a small number of badge values. To ensure that we have

a precise and comprehensive picture of the users' abstinence or relapse history, we omitted the users with an NA badge value and those who had less than 10 badge values.[4]

(3) *Irregularities in values of badge sequences.* Finally, we observed irregularities in the badge sequences of some users. Two prevalent examples were sudden jumps between consecutive badge values (e.g., from the badge value of 30 to 150) and falloffs to large badge values (e.g., from the badge value of 200 to 100). To ensure the integrity of the badge sequences, we omitted the users with badge sequences violating any of the following heuristic rules: for any two consecutive badge values $b_t$ and $b_{t+1}$, (i) the difference $b_{t+1}-b_t$ should be either negative, 0, 1, or 2, and (ii) if $b_{t+1}-b_t < 0$, then $b_{t+1}$ should be less than or equal to 10. These rules allowed capturing the expected behavior (increasing badge sequences with possible drops to small badge values) and presumably minor system glitches (two consecutive badges with identical values or increasing values that differ by 2), while disallowing the majority of the irregularities that we observed.

We report the summary statistics of the filtered dataset in the "Survival data" columns for SS and SD in Table 14. We refer to it as survival data since we leveraged this dataset for our subsequent survival analysis-based statistical method. Figure 27 shows the daily volumes of relapses and the cumulative distribution functions (CDFs) of the number of users over the total number of relapses experienced by the users. We observe that 2,566 users/87.97% in SS and 2,479 users/80.64% in SD did not relapse during the period of our study. Of those who relapsed, the majority relapsed once (213 users/7.3% in SS and 291 users/9.47% in SD). Some users relapsed many times; our inspection of their badge sequences revealed that they contain consecutive badge

[4]We opted for a conservative approach and omitted the small number of users (492 users/6.81% in SS and 559 users/7.74% in SD) who initially did not have a badge but later obtained one (as reflected in our crawl) to ensure the accuracy of our statistical analysis.

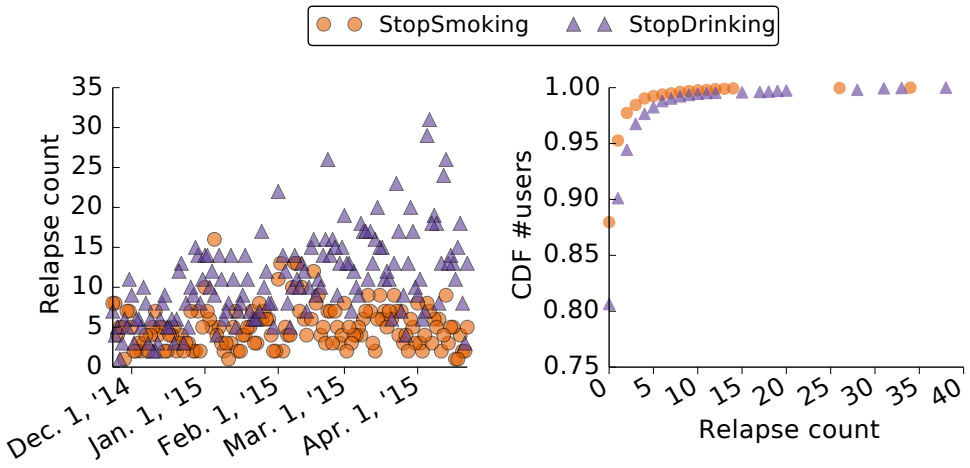



Figure 27: *Left*: Daily volumes of relapses observed in StopSmoking (SS) and StopDrinking (SD). *Right*: Cumulative distribution functions (CDFs) of the number of users over the total number of relapses experienced by the users.

values of 1, indicating that these users suffered from consecutive failed attempts to abstain from smoking/drinking.

## 6.4 Statistical Method

### 6.4.1 Explanatory Variables

We first introduce the variables utilized to analyze smoking and drinking relapse events; they are outlined below and summarized in Table 15. The choice of these variables were framed in the light of prior literature on health recovery and addiction cessation [153, 178] and align with the goals of RQ 1 and RQ 2.

**Engagement.** Our first set of explanatory variables focus on various aspects of engagement within the SS and SD communities. We consider three dimensions of engagement: self-disclosure, the support received from other users (in-support), and the support provided to other users (out-support).

Literature indicates that self-disclosure can be an important therapeutic ingredient and is linked to improved physical and psychological well-being [43]. In the context of

Table 15: List of explanatory variables used in the statistical models for StopSmoking (SS) and StopDrinking (SD).

| |
|---|
| **Engagement variables:** |
| self-disclosure SS/SD |
| in-support SS/SD |
| out-support SS/SD |
| **Language variables** (grouped for brevity, those for SS/SD are LIWC related): |
| first person singular, first person plural, second person, third person pronoun words counts SS/SD |
| "body", "health" words counts SS/SD |
| past, present, future tense words counts SS/SD |
| positive affect, negative affect, "swear" words counts SS/SD |
| addiction words count OSR |

health conditions that are typically considered socially stigmatized, such as addiction, self-disclosure has been noted to be a basic element in the attainment of improved health [140]. This is because self-disclosure results in disinhibition [158], which is known to play a positive role in psychological counseling. In SS/SD, the majority of the posts have a self-disclosing nature, including reflections of feelings, thoughts, and experiences related to quitting (see Figure 24 for a sample of post topics), whereas through the comments the users provide feedback or encouragement to the author of the original post. As such, we capture self-disclosure by considering the users' tendency to submit posts (relative to comments) and define the corresponding variable as the ratio of the number of posts to the total number of posts and comments the user has in SS/SD.

Addiction literature also indicates social support to act as an important mediator of stress during smoking/drinking urges [153]. We consider two forms of social support: in-support and out-support. For both, we consider the users' commentary activities in SS/SD (as a response to a post or another comment) as the primary mechanism of providing feedback and support in these communities. Specifically, we define in-support to be the average number of comments received per post submitted by the user. As the initiator of the discussion in the post, we assume that all the

comments on the post contribute towards the in-support of its author (even if some of the comments are directed to other comments). We capture out-support by considering the users' tendency to respond to other users' posts and comments (relative to the number of users who responded to them). To this end, we leverage a network we construct as follows: if user A comments on user B's post or comment, we establish a directed unweighted edge from user A to user B in the network (if it does not already exist). Then, based on the network, we define the out-support of the user to be the ratio of his or her outdegree to the sum of his or her outdegree and indegree.

This set of explanatory variables therefore contains three variables and we refer to them as *engagement variables*.

**Language.** Our second set of explanatory variables focus on extracting linguistic attributes from a user's posts and comments in SS/SD and OSR. The Linguistic Inquiry and Word Count (LIWC: `www.liwc.net`) is a proprietary database containing 74 psychologically meaningful linguistic categories and the word patterns associated with each category (which includes exact matches as well as prefixes like addict*). Prior work has used LIWC to characterize and distinguish women suffering from postpartum depression [56], individuals at risk for depression [57], and smokers on Twitter who are at risk for relapse [124]. We introduce a count variable for each of the 12 LIWC categories we deemed the most relevant (see Table 15), representing the number of times that any of the words in the corresponding category appear in the user's content.

To examine if smoking or drinking-related content in OSR can potentially help characterize smoking and drinking relapse events, we adapt the addiction-related smoking and drinking lexicons that we utilize in our prior work [164], which we present in Chapter 5 (see Table 9 therein). Since the user is unlikely to use every word in the lexicon, we consider a single count variable (referred to as addiction words

count), representing the total number of times that any of the words in the lexicon appear in the user's posts or comments.

Together, this set of explanatory variables contains 13 variables and we refer to them as *language variables*.

### 6.4.2 Survival Analysis

*Why survival analysis?* As achieving long-term abstinence of tobacco or alcohol is challenging [175], relapse to smoking or drinking is a behavior change that can happen anytime, even after years of cessation. However, in studies of human subjects, it is often the case that the study period is not long enough to observe whether the event of interest (relapse in our case) has happened or not. Consequently, the analysis of the probability of "survival" (e.g., prevention of relapse) during the study period as a dichotomous variable (relapsed vs. not relapsed) using conventional statistical techniques (e.g., a linear regression technique or a chi-squared test) fails to account for non-comparability between subjects whose relapse is observed during the study period versus not [80]. Also, simply ignoring subjects who do not experience the event of interest has been noted to produce biased underestimates of survival [148]. Therefore, we borrow techniques from the survival analysis literature for the purposes of our study.

In the survival analysis literature, if the event does not happen before the study ends, the subjects are considered to be *right-censored* at the last assessment time [80]. Another important concept is that of the *survival function* $S(t)$, which denotes the probability that an individual survives at least to time $t$. The Kaplan-Meier method is a widely used nonparametric technique to graphically construct the unconditional survival function without covariates [80]. It is important to note that this method provides an estimation of the survival function if the underlying data is censored (as in our case), but the estimated function is still useful for forecasting purposes [32].

We leverage the Kaplan-Meier method to examine how participation in SS/SD is associated with the risk of relapse (RQ 1).

### 6.4.3 Cox Regression

We also employ Cox regression [52] to examine associations between time to first relapse (as reflected in our dataset) and our explanatory variables (RQ 2). The analysis on the users' subsequent relapses is left as future work. The Cox regression is a statistical technique to analyze survival data where time to event is formulated as a function of possible prognostic factors [67]. The response variable in Cox regression is typically represented as a pair of values: time to event and a status indicator denoting whether the event of interest has happened or not. We leverage the users' badge values to determine their response variable values. E.g., consider the following response variable values:

(a) If user $A$ had the badge value of 30 when they experienced the first relapse, then their values for the response variable would be the pair (time to event = 30, relapsed = "yes").

(b) In contrast, if user $B$ did not experience a relapse and had the badge value of 150 on the last day of our observation period, then their values for the response variable would be the pair (time of event = 150, relapsed = "no"), denoting that user $B$'s relapse time is right-censored.

(c) A key point to consider in our case is that users may join SS/SD at any time during their cessation period and thereby specify any value for their initial badge in SS/SD. E.g., if user $C$ has been abstaining from smoking/drinking for 200 days and decides to join SS/SD, they would pick 200 as their initial badge value. In this case, we consider user $C$ as a *delayed entry* [80] to our study. The Cox regression supports such delayed entries as the user $C$; the response variable is

then represented as a triplet of values: starting time of the observation, ending time of the observation, and a status indicator as before. Thus, if user $C$ had the badge value of 300 when they experienced the first relapse, then their values for the response variable would now be the triplet (observation start = 200, observation end = 300, relapsed = "yes").

**Statistical Models.** To understand the explanatory powers of our independent variables, we consider three statistical models: the ENGAGEMENT model, the LANGUAGE model, and the ENGAGEMENT + LANGUAGE model, which consist of the engagement, language, and engagement and language variables, respectively. The LANGUAGE model is motivated from prior work investigating the role of linguistic attributes in describing or predicting health challenges from social media [124, 110], and through the other two models, we examine the additional role of engagement in characterizing smoking and drinking relapse events. In these models, we log-transform the language variables (which denote counts) to correct for outliers and skewness.

## 6.5 Results

### 6.5.1 RQ 1: Participation and Likelihood of Relapse

Per our RQ 1, we begin by examining how the extent of participation in the SS and SD communities relates to estimates of smoking/drinking relapse and abstinence. To that end, Figure 28 shows the survival functions obtained for SS/SD using the Kaplan-Meier method. Both SS and SD have an initial drop-off with 55% and 67% of the users estimated to be at risk of relapse beyond the first day of abstinence.

We also obtain the median survival time from our Kaplan-Meier estimator, which is the time at which 50% of the users are estimated to have relapsed. The median survival time for SS is 25 days (95% confidence interval (CI) = [1, 127]), whereas for SD it is considerably longer with 56 days (95% CI = [35, 102]). These short median

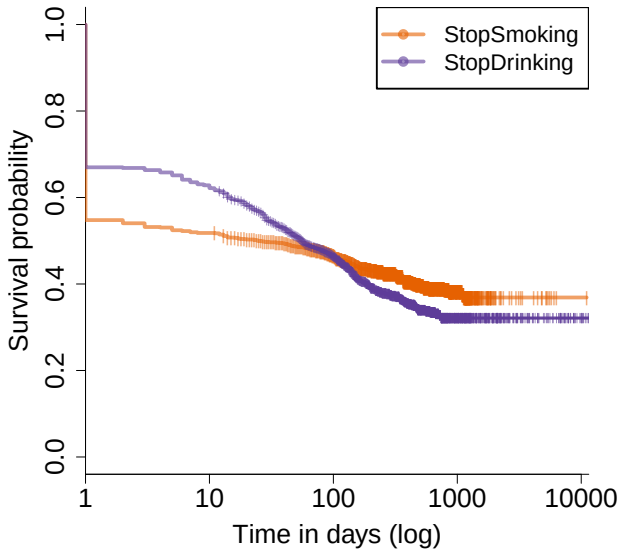



Figure 28: Survival functions obtained for StopSmoking (SS) and StopDrinking (SD) using the Kaplan-Meier method.

survival times of SD and SS align with established studies in the addiction literature [130]. In a way, we find social media-based empirical evidence that bolsters the known fact that smoking or drinking cessation is difficult, and the experiences of individuals who participate in the Reddit support communities align with observations about the same made in clinical populations [153].

However, we find that the probability of survival (not experiencing a relapse event) 500 days after being on the SS community is 40%, while the same for the SD community is 34%. Therefore, although a significant fraction of the populations on both communities are expected to relapse in the short term, survival trend shows a pretty stable pattern in the longer term. In other words, beyond 1000 days, the likelihood of experiencing a relapse event is low in both communities.

Survival curves can also be used to estimate the likelihood that a user who has not experienced a relapse event at a specific time point will continue to abstain from smoking/drinking for an additional length of time (calculated by dividing the

probability of survival at time $t_j$ by the probability of the same at time $t_i$, where $j > i$). For example, the probability that a user in SD who did not relapse by 50 days would continue to do so for another 50 days is $0.46/0.51 = 90.2\%$. If the user does not relapse in 500 days, the probability of continuing the same for another 500 days is $0.32/0.34 = 94.1\%$. So, as the time of abstinence increases, the likelihood of ever experiencing a relapse event decreases. This analysis provides an alternative explanation of the observation in the paragraph above.

What is interesting, however, is the noticeable difference in the survival probabilities for SS and SD. We observe that the SD users are more likely to maintain abstinence beyond any number of days up to about 100 days, after which the SS users become more likely to maintain abstinence in the long run. This finding may be explained by the fact that while there is considerably high concomitance between the health behaviors of smoking and drinking [154], smokers tend to relapse at a faster rate than alcoholics; however, those smokers who have maintained abstinence for a while have a greater likelihood than alcoholics to continue to quit post cessation [24].

Overall, we conclude that in the context of RQ 1, participation in the SS and SD communities can lend us valuable insights into patterns and estimates of the likelihood of relapse over time, both in the short and long terms.

### 6.5.2 RQ 2: Role of Engagement and Linguistic Variables

Recall that the goal of RQ 2 is to examine how attributes of engagement as well as linguistic constructs derived from content shared on SS/SD are associated with and predictive of the likelihood of relapse in the future.

**Assessing Goodness of Fit.** First, we evaluate the goodness of fits of our models using *deviance*. Table 16 provides a summary of the different model fits. Compared to the Null models, we observe that all three of our models provide considerable explanatory power with significant improvements in deviances in both SS and SD.

Table 16: Summary of different model fits. Null is the intercept-only model. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{3}$).

| | StopSmoking (SS) | | | | StopDrinking (SD) | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Deviance | df | $\chi^2$ | p-value | Deviance | df | $\chi^2$ | p-value |
| Null | 4,235.95 | 0 | | | 7,619.08 | 0 | | |
| Engagement | 4,184.84 | 3 | 51.11 | $< 10^{-10}$ | 7,529.27 | 3 | 89.81 | $< 10^{-18}$ |
| Language | 4,123.96 | 13 | 111.99 | $< 10^{-17}$ | 7,484.28 | 13 | 134.80 | $< 10^{-21}$ |
| Engagement + Language | 4,104.15 | 16 | 131.80 | $< 10^{-19}$ | 7,424.20 | 16 | 194.88 | $< 10^{-32}$ |

The difference between the deviance of a Null model and the deviances of the other models approximately follows a $\chi^2$ distribution, with degrees of freedom (df) equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of the Engagement model with that of the Null model in SS, we see that the information provided by the engagement variables has significant explanatory power: $\chi^2(3, N = 2,917) = 4,235.95 - 4,184.84 = 51.11, p < 10^{-10}$. This comparison with the Null model is statistically significant after the Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{3}$ as we consider three models). We observe similar deviance results for the Language and Engagement + Language models in both SS and SD, with the latter model possessing the best fit and highest explanatory power.

**Assessing Predictive Power of the Cox Regression Models.** Next, we report the 10-fold cross-validated concordance scores of our Cox regression models to evaluate their predictive power. Briefly put, concordance is a generalization of the area under the receiver operating characteristic (ROC) curve and it measures how well a model discriminates between different responses. Specifically, it is the fraction of the pairs of observations in the data where the observation with the higher survival time has the higher probability of survival predicted by the model [80]. Generally speaking, a concordance of greater than 0.5 indicates a good prediction ability (the value of 0.5 denotes no predictive ability). Here, we first randomly split our dataset into

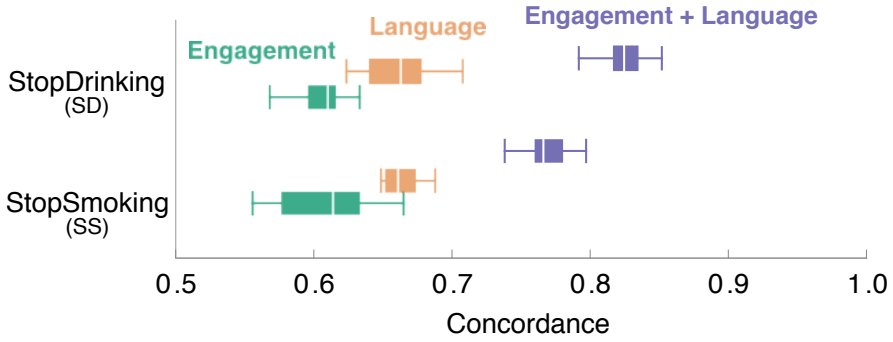



Figure 29: Boxplots for the 10-fold cross-validated concordance scores of the statistical models. The ENGAGEMENT + LANGUAGE model possesses a significant predictive power with a mean concordance of 0.77 in SS and 0.82 in SD. The boxplots are spread out vertically to avoid spatial overlap.

10 folds and then considered each fold one by one: we trained our models on the remaining 9 folds and computed the concordance scores of the models on the fold under consideration. This led to 10 concordance scores for each model, generated from the same set of folds. Figure 29 shows the boxplots for these concordance scores. We observe that the best performing model in both SS and SD is ENGAGEMENT + LANGUAGE, which possesses a significant predictive power with a mean concordance of 0.77 and 0.82 in SS and SD, respectively.

Summarily, we conclude that both engagement and language variables include valuable signal relating to the likelihood of relapse or abstinence in the SS/SD communities, compared to either of the categories alone. How do and by how much do these engagement and language variables relate to the risk of relapse? To address this, we present a discussion of the different notable predictors in the next subsection.

**Predictors of Relapse and Abstinence.** In Table 17, we present expanded results of our best-performing Cox regression model (ENGAGEMENT + LANGUAGE), reporting hazard ratios (HRs) and 95% confidence intervals (CIs) of different explanatory variables in this model. Note that the hazard ratio for a variable denotes the risk

Table 17: Results of Cox regression examining the associations between time to first smoking/drinking relapse and the explanatory variables. "OSR" corresponds to subreddits other than StopSmoking (SS)/StopDrinking (SD).

| | StopSmoking (SS) | | StopDrinking (SD) | |
|---|---|---|---|---|
| Explanatory variable | Hazard ratio | [95% CI] | Hazard ratio | [95% CI] |
| self-disclosure SS/SD | 0.87 | [0.34, 2.23] | 0.22 ** | [0.10, 0.48] |
| in-support SS/SD | 1.03 | [0.98, 1.08] | 1.02 * | [1.01, 1.04] |
| out-support SS/SD | 0.30 ** | [0.15, 0.62] | 0.17 ** | [0.10, 0.29] |
| first person singular pronoun words count SS/SD | 1.55 * | [1.07, 2.23] | 1.27 | [0.97, 1.66] |
| first person plural pronoun words count SS/SD | 1.10 | [0.84, 1.42] | 0.95 | [0.81, 1.11] |
| second person pronoun words count SS/SD | 0.89 | [0.72, 1.11] | 0.98 | [0.84, 1.14] |
| third person pronoun words count SS/SD | 0.90 | [0.70, 1.14] | 0.93 | [0.81, 1.06] |
| "body" words count SS/SD | 0.99 | [0.76, 1.31] | 1.04 | [0.87, 1.23] |
| "health" words count SS/SD | 1.02 | [0.81, 1.27] | 0.80 ** | [0.68, 0.93] |
| past tense words count SS/SD | 0.68 ** | [0.53, 0.88] | 0.80 * | [0.65, 0.98] |
| present tense words count SS/SD | 1.28 | [0.90, 1.83] | 1.41 ** | [1.09, 1.83] |
| future tense words count SS/SD | 1.01 | [0.79, 1.31] | 0.95 | [0.80, 1.13] |
| positive affect words count SS/SD | 0.69 ** | [0.52, 0.91] | 0.83 | [0.66, 1.05] |
| negative affect words count SS/SD | 0.99 | [0.75, 1.32] | 1.12 | [0.92, 1.37] |
| "swear" words count SS/SD | 0.99 | [0.75, 1.33] | 0.90 | [0.75, 1.09] |
| addiction words count OSR | 0.70 ** | [0.63, 0.78] | 0.80 ** | [0.75, 0.85] |

** $p < 0.01$, * $p < 0.05$

of a user relapsing with one unit increase in the value of the corresponding variable. A hazard ratio smaller than 1 indicates a decreased daily risk of relapse (increased survival rate), while a hazard ratio larger than 1 indicates an increased daily risk of relapse (decreased survival rate).

We observe from Table 17 that the language variables are particularly important variables that characterize smoking and drinking relapse events. Below, we highlight the results for some of the prominent language variables, including examples of the most common phrases to provide missing context.

*First person singular pronouns* are associated with high risk of smoking relapse (HR=1.55, meaning that the risk of relapse to smoking increases by 55% with one unit increase in the value of the log of the first person singular words count SS variable). This category contains words such as "i" and "me"; e.g., a post excerpt from an SS user who eventually relapsed: "*i*'m [...] craving a smoke all day, and now that [...], *i* don't have anything to distract *me* anymore". We presume that since use of

first person singular pronouns indicates high self-attentional focus and psychological distress [175], risk of relapse may be heightened due to experience of stress or depressive episodes as indicated in the addiction literature [154]. Additionally, *lower* use of *second person pronouns* (flipping the ratio to denote the decrease in value, HR=1/0.89=1.12 for SS) and *third person pronouns* (HR=1/0.90=1.11 for SS) are indicative of lowered social interaction with the greater community and linked to increased risk of relapse [43] (though, these interactions are not statistically significant).

*Past tense words* are associated with low risk of smoking/drinking relapse (HR=0.68 for SS; HR=0.80 for SD). This category contains words such as "had" and "felt"; e.g., a comment excerpt from an SS user who maintained abstinence: "i *had* a dream where i smoked one cig, i *felt* incredible sad that my progress *was gone*". This observation is supported by the literature that reflecting on past experiences is known to improve decision-making abilities among addiction quitters, including improving self-control and reducing impulsivity to relapse urges [44]. Additionally, *present tense words* are associated with high risk of drinking relapse (HR=1.41). This category contains words such as "know" and "seem"; e.g., a comment excerpt from an SD user who eventually relapsed: "i *don't know* about withdrawals but many cups of tea and lots of candy *seem* to help the cravings". Literature has indicated that focus on the here and now, as captured by the use of present tense words, tend to be linked to lowered cognitive functioning and increased mental health challenges—both of which show comorbidity with addiction [115].

*Positive affect words* are associated with low risk of smoking relapse (HR=0.69). This category contains words such as "fun" and "yay"; e.g., a comment excerpt from an SS user who maintained abstinence: "*great* man! *thanks* for dropping in and [...]! you *inspire* me". Our finding is supported by the literature that has found that experience of positive emotions, including regulatory efforts to alleviate negative mood states is strongly linked to smoking cessation and relapse prevention [34, 124].

In contrast, use of *negative affect words* increases the likelihood of drinking relapse (HR=1.12, though this interaction is not statistically significant). Literature indicates increased negative affect to be associated with symptoms such as mental instability, helplessness, loneliness: factors known to trigger addiction urges [108].

Next, *"health" words* are associated with low risk of drinking relapse (HR=0.80). This category contains words such as "medic*" and "alcohol*"; e.g., a comment excerpt from an SD user who maintained abstinence: "i [...] and got *medicine* designed to help *alcoholics detox* from *alcohol* safely". Recognizing the needs of one's health and well-being is known to lead to better lifestyle choices and improvement in self-regulation and self-efficacy [110].

*Addiction words* are also associated with low risk of smoking/drinking relapse (HR=0.70 for SS; HR=0.80 for SD). One explanation behind this observation could be that some users tend to use other subreddits (OSRs) to receive feedback about the various challenges related to quitting; e.g., a post excerpt submitted to the subreddit *Anxiety* by an SS user: "i had a couple of panic attacks, and decided to quit *smoking* since i figured they were from [...]". Moreover, as with the discussion of health and well-being topics, awareness of one's addiction challenges and risk has been known to increase one's cognitive control and therefore reduce risk of relapse [97].

Finally, examining the *engagement variables*, we observe that self-disclosure significantly reduces the risk of drinking relapse (HR=0.22). Also, in-support is associated with high risk of smoking/drinking relapse (HR=1.03 for SS, though this interaction is not statistically significant; HR=1.02 for SD). We conjecture this might be because the users who received greater support from the SS/SD communities are those who are more vulnerable to relapse. Alternatively, it could also be the support-seeking nature of the content shared by users struggling to maintain abstinence, which attracts responses from the greater community. Finally, we observe that out-support is associated with low risk of smoking/drinking relapse (HR=0.30 for SS; HR=0.17

for SD). Prior work has indicated that social engagement has a positive impact on the psychological states of individuals [56]. Hence, we conjecture that greater feedback to other users on the support communities we study helps keep individuals more motivated and focused towards their respective self-improvement goals.

## 6.6 Discussion

Our results show that participation in the smoking and drinking support communities we study may not be linked to abstinence in the short term—half of the population is estimated to relapse to smoking/drinking within 25/56 days post-cessation. However, the relatively smaller proportion of individuals who *do* survive past the initial few months are estimated to experience sustained abstinence over a long period of time (beyond three years). In essence, while for short-term abstinence our findings call into question the effectiveness of the communities, we found that in the course of time these platforms do provide individuals a place where they can improve their regulation and efficacy toward preventing risks of relapse. Direct comparison between our study sample from Reddit and clinical populations would be inappropriate. However, our observations align with the literature on addiction where it has been observed that although smoking in particular is highly relapse-prone, individuals who have abstained sufficiently long tend to have a considerably lowered probability of resuming their pre-cessation smoking choices [24].

We also discovered several characteristics of engagement and language that indicate increased or decreased chance of relapse. Higher self-attentional focus and detachment from the social realm (first, second, and third person pronoun use), and focus on the present increase the risk of relapse. On the other hand, reflection on one's health and addictive behaviors, expression of positive emotions, self-disclosure, and increased desire to provision support to others (engagement variables) heighten

the likelihood of abstinence. We also demonstrated the satisfactory predictive capability of these variables in estimating the communities' recovery behaviors (abstinence/relapse) over time. We believe these findings can have notable impact on several points of scientific and practical consideration. We discuss them below.

### 6.6.1 Scientific and Practical Relevance

**Clinical Research.** Given the predictive capability of our survival analysis-based method, early warning systems could be developed that analyze patterns of participation on the platform. These systems could engage appropriately if the likelihood of relapse in the broader community increases beyond a certain level. Provisions like this, however, could raise ethical and privacy concerns and must therefore be carefully designed and developed. If successful, such early warning systems could further provide scientific and clinical insights into understanding and identifying prospective factors associated with abstinence and relapse over time. They can also help discover previously underexplored variables that may contribute towards the success or failure of cessation in a community. Moreover, we found that the likelihood of abstinence and relapse can be projected and tracked over time. This could help clinicians, medical professionals, and policy makers better understand people's experiences around maintaining long-term abstinence from tobacco or alcohol, and the strategies that may have worked for them.

**Designing Health Interventions.** The different engagement variables and other linguistic constructs indicated by our results to be associated with increased likelihood of abstinence may also be utilized to design interventions. These can bring timely and personalized help to individuals in the community intending to abstain from smoking/drinking.

By identifying a link between variables that increase risk of relapse and an individual's Reddit activity, moderators could pair them up with peers in the community

for support. Social support and higher levels of social capital have been known to help individuals fight addiction urges [72]. In fact, finding "people like me" is a primary stated reason for user participation in online communities [71]. Encouraging or actionable content from others may also be promoted in their activity timelines; positive feedback may mediate urges to relapse and improve self-regulation toward abstinence, whereas content with instrumental information may help individuals identify and cope with the challenges and struggles that characterize cessation attempts. Moreover, since we also found that addiction-related posting activity or commentary in other subreddits is associated with increased likelihood of abstinence, provisions may be made to encourage relapse-prone individuals participate more in the broader social platform.

**Understanding and Tracking Community Efficacy.** Our computational approach also demonstrated the ability to proactively identify a community's efficacy toward promoting addiction cessation, including factors linked to such efficacy. Therefore, we believe our methods and the insights we gleaned may be used to create enabling reflective interfaces for community moderators or involved volunteers, so as to not only understand how participation in these platforms supports their goals of self-improvement, but also to make provisions to quantify and improve their effectiveness. These provisions could include a variety of mechanisms to alter community dynamics. Based on our survival analysis-based methodology, moderators could recognize time of vulnerability in the community, for instance, or when to direct requests for advice or help to appropriate, actively engaged users. Alternatively, platforms like AA (Alcoholics Anonymous) have benefited from their sponsorship program that claims to promote cessation in the long-term [170]. In a similar manner, moderators could pair up individuals in early stages of their cessation attempt with long-term members who would act as formal mentors.

### 6.6.2 Limitations

**Gaps in Self-Reports.** First, while we adopted several cautionary steps to ensure our inferences of abstinence and relapse events from badges accurately reflect a user's experience, we acknowledge that it may still suffer from some limitations. Just like self-reporting in survey approaches, our dataset also suffers from the challenges of falsified reporting (e.g., user not reporting that they had a relapse due to self-representation or social comparison concerns [58]), temporal gaps between actual relapse and when it is reflected in a user's badge, or failing to report the relapse event altogether. However, since we analyzed abstinence and relapse at the macro (or community) level, we expect these gaps in self-reporting to impact our findings to a lesser extent.

**Generalizability and Causality.** Finally, focusing on a large and prominent support community like SS or SD allowed us to analyze abstinence and relapse events over a diverse population; however, we caution against broad generalizations. The communities we study recognize themselves as "self-improvement communities", which implies that they likely tend to attract those individuals who are already considering quitting smoking/drinking actively. Furthermore, we cannot causally attribute abstinence or recovery to the different explanatory variables we investigate (participation, engagement, and language), especially because of the lack of information on whether the users we study sought support, counseling, or interventions through offline means.

## 6.7 Conclusions

We presented a survival analysis-based computational methodology to analyze and understand smoking and drinking relapse events of individuals in two support communities on Reddit. We leveraged the self-reported badge information of 14K users as a way to infer their abstinence status. We found that although participation in the

community is not linked to high likelihood of smoking/drinking abstinence during the one/two months post-cessation, it shows a stable trend of heightened chance of abstinence beyond three years. We also found that the linguistic constructs of the content shared by the users as well as the extent of their engagement in these communities are indicative of high or low risk of smoking/drinking relapse. Our work provides one of the first quantitative insights into evaluating the effectiveness of social media support communities in promoting cessation from smoking and drinking, and how social media may be leveraged to tackle addiction-related health challenges.

# CHAPTER VII

# INSIDER TRADING ANALYSIS: PATTERNS AND DISCOVERIES

The insiders of a company are corporate officers, directors, or beneficial owners who own more than 10% of the company's stock. While the insiders can legally trade their companies' stock in financial markets, some insiders exploit their roles and use *nonpublic* information about their companies as a basis for trade. This is called illegal insider trading and it is actively prosecuted by the Securities and Exchange Commission (SEC). To monitor trades by the insiders, SEC requires these trades to be disclosed via a form called *Form 4*. To the best of our knowledge, very little published research is available that uses computational techniques to help financial regulators and policymakers better understand the dynamics behind how the insiders trade.

We performed the first academic, large-scale exploratory study of the complete Form 4 filings from SEC, and made surprising and counterintuitive discoveries. We analyzed over 12 million transactions by around 370 thousand insiders spanning years 1986 to 2012, the largest reported in academia. Our analysis consists of two major components. The first explores the trading behaviors of the insiders from a temporal perspective. By analyzing the time series of the transactions, we discovered distinctive temporal patterns in the insiders' trades that may be explained by government regulations, corporate policies, and macroeconomic factors. For instance, we determined that a significant portion of the insiders makes short-swing profits (i.e., profit

Material adapted from work appeared at IEEE/ACM ASONAM 2013 [168] and in Springer SNAM Journal [165].

resulting from a combined purchase and sale, or sale and purchase, of the company's stock within a 6-month period) despite the existence of a rule designed to prevent short-swing trading.

The other main component of our analysis explores the trading behaviors of the insiders from a graph-based perspective. Specifically, it focuses on the insiders who consistently trade on similar dates, and therefore, might be sharing nonpublic inside information with each other. Graphs enable us to capture such relationships between all the insiders in the broader context. By constructing insider networks that represent these relationships and studying the characteristics of the networks, we found strong evidence that insiders form small clusters in which trade-related information might propagate both vertically (between higher-level and lower-level insiders) and horizontally (among lower-level insiders).

We believe this work could form the basis of novel tools for financial regulators and policymakers to detect suspicious trades based on our characterization of how the insiders trade. The results of this work were presented to SEC.

## 7.1 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 [22]). 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, 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.

### 7.1.1 Opportunities for Data Mining

Government regulators are increasingly interested in applying data mining techniques to detect fraud and illegal insider trading [78]. 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' relationships 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 mining as well as graph mining and social network analysis. 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, graph-based analysis is important for detecting illegal insider trading since insiders often share information with each other through their social networks. Graphs enable us to capture such relationships between all the insiders in the broader context. With graph-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 little 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.

### 7.1.2 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 non-parametric 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.

### 7.1.3 Contributions

We conduct an extensive large-scale analysis of insiders' trades using the Form 4 filings. Our analysis consists of three components. The first is based on time series 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 the prices of the insiders' transactions and the 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 graph mining and social network analysis; in this component we construct networks of insiders based on the similarity of the 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 to 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 towards 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 chapter.

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

## 7.2 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

Table 18: Summary statistics for our dataset. We focus on open-market sale and purchase transactions.

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

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 United States. Table 18 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 Sections 7.4.1 and 7.4.2) use the mapping from [63], 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

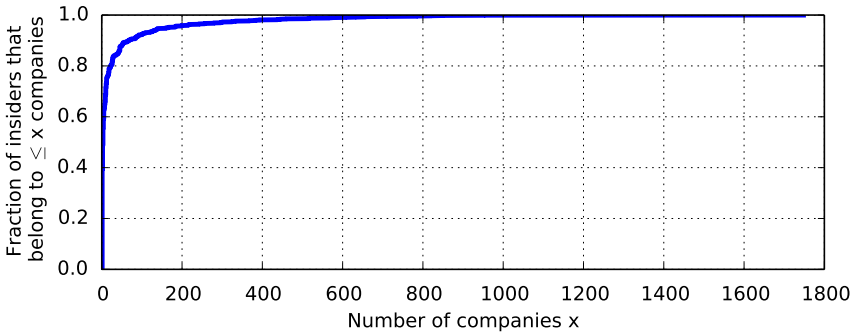



Figure 30: Empirical cumulative distribution function for the number of companies that insiders belong to in our dataset. A majority of insiders belong to a small number of companies.

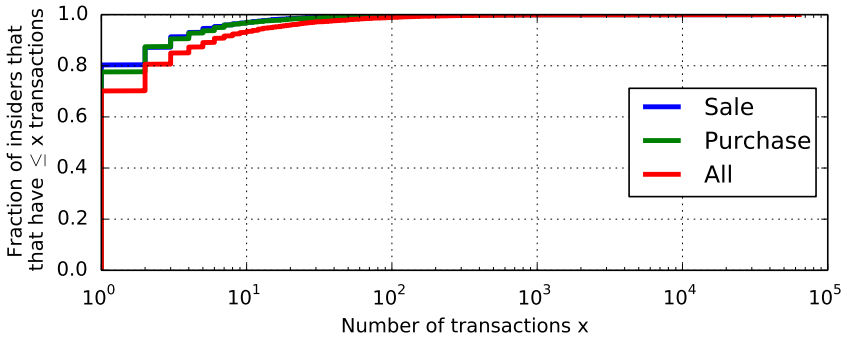



Figure 31: Empirical cumulative distribution function for the number of transactions that insiders have in our dataset. Note that the x-axis is in log-scale. A majority of insiders have a small number of transactions.

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 [173].

Figures 30 and 31 show the empirical cumulative distribution functions for the number of companies that insiders belong to and the number of transactions that

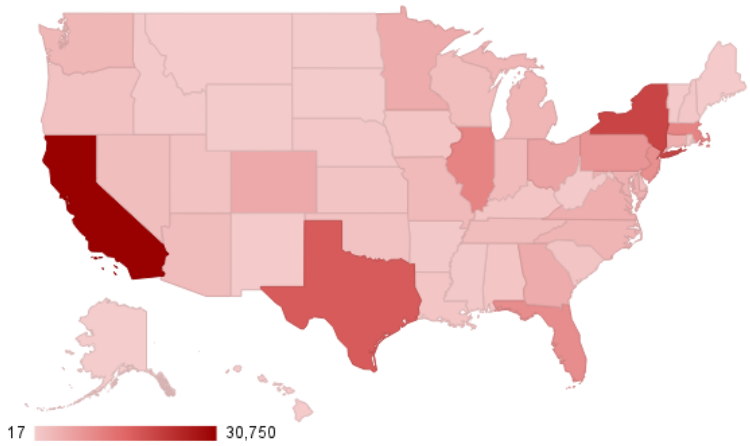



Figure 32: Geographical distribution of the number of transactions based on the zip codes of the insiders' companies. Darker color indicates higher number. The highest number of transactions initiate from the state of California.

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 32 shows the geographical distribution of the transactions based on the zip codes of the corporate headquarters. The highest number of transactions occur for companies headquartered in the state of California, followed by New York and Texas.

## 7.3 Prior Work and Our Differences

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.

### 7.3.1 Profiling Insiders

In the finance domain, Cohen et al. [47] 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 graph-based 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. [39] 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. [30] find that managers release abnormally positive news before stock option exercises to obtain relatively high sales prices, and Aboody et al. [2] show that managers tend to release bad news before stock option grants to fix lower strike prices. Brockman et al. [29] 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 et al. [109] 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 et al. [106] 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 to these works, we explore a significantly larger dataset both in terms of the number of companies and time span.

### 7.3.2 Detecting Potential Fraud and Illegal Trades

Goldberg et al. [76] 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 [60] 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. [100] 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 graph-based analysis is insider-centric as opposed to firm-centric, and we do not question the credibility of the SEC filings. In [160], 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 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 [116], which uses a graph-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 [135]. A more general system, Sherlock [18] 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.

### 7.3.3 Mining Financial Data

Fan et al. [65] present a data mining-based automatic trading surveillance system for large data with skewed distribution using multiple classifiers. Bizjak et al. [26] document the network structure in the interlocking board of directors to explain how inappropriately backdating compensation spreads. Adamic et al. [4] 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 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.

## 7.4 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

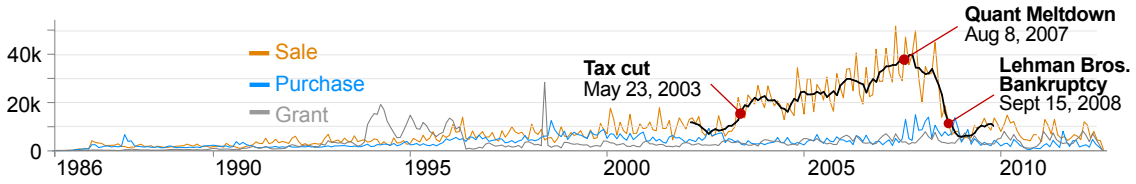



Figure 33: The daily count of *Purchase*, *Sale*, and *Grant* transactions (the most common types) over 1986-2012. 180-day centered moving average for Sale transactions shown in black. The change in the U.S. tax law in 2003 (reduced capital gains taxes) boosted Sale transactions for following years. Financial crises like the "Quant Meltdown" in 2007 and the burst of "housing bubble" in 2008 suppressed them.

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 [47]. The second factor is the relationships between insiders. If a group 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.

### 7.4.1 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 Figure 33. 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 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

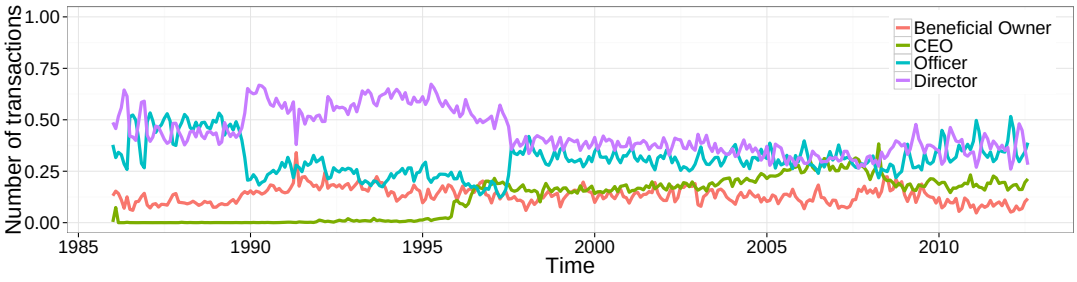



Figure 34: Transactions break down by role codes. Only the most frequent four codes are shown. Beneficial owners behave differently than the other insiders.

with the 2003 change in the United States tax law[1] that reduced capital gains taxes. The sharp drop in sales occurs after the "Quant Meltdown" of August 2007[2] [98] 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 [129]). 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.

Figure 34 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 towards 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 35 depicts trading activity in various sectors. In terms of the number

[1] Enacted May 23, 2003.
[2] A point identified, with hindsight, as the start of the financial crisis.

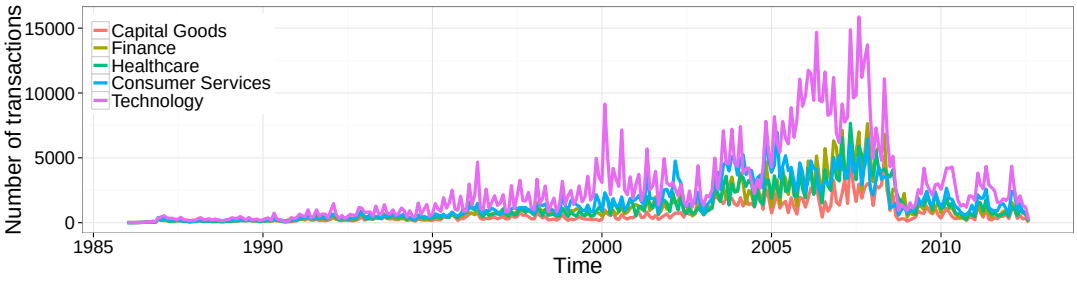



Figure 35: Transactions break down by sectors. Only the most frequent five sectors are shown. Most activity comes from the technology sector.

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 Figure 33. Inspired by [149], 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.

### 7.4.2 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 36 and 37 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→P. The other three transaction pairs are *purchase-then-sale* (P→S), *sale-then-sale* (S→S), and *purchase-then-purchase* (P→P). From Figures 36 and 37, we see that, in general, S→P and P→S pairs are less common

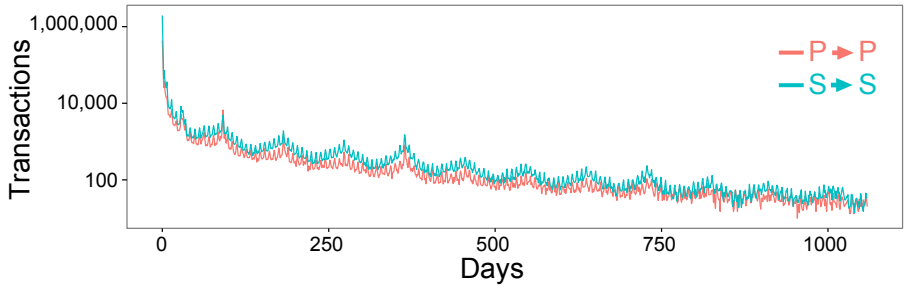



Figure 36: Time between consecutive transactions of the same type: purchase-then-purchase (P→P) and sale-then-sale (S→S). The pattern is oscillatory, with a cycle of about 90 days.

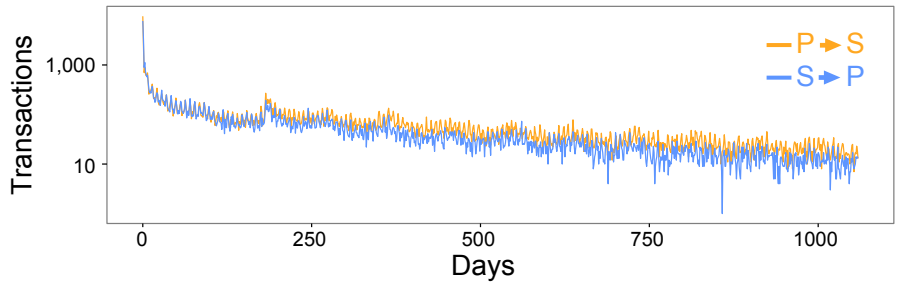



Figure 37: Time between consecutive transactions of different types: purchase-then-sale (P→S) and sale-then-purchase (S→P). The highest peak for both distributions is around the point corresponding to 180 days.

than P→P and S→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→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→P pairs. Another notable observation in Figure 36 is that the pattern is strongly oscillatory, with a cycle of about 90 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 Figure 37 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 statue 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 six-month period. As a result of the rule, one might expect that round-trip transactions completed within a six-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 $t_\ell$ ($t_k \leq t_\ell$), the profit earned is

$$\big(\log(P^C_{t_\ell}) - \log(P^C_{t_k})\big) \times P^C_{t_k} \times \min(ST^C_{t_k}, ST^C_{t_\ell}), \tag{3}$$

where $P^C_{t_i}$ is the market closing price of company $C$'s stock at date $t_i$ and $ST^C_{t_i}$ 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]

[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 less than six 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 six 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[.]

[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 match to reach the maximum possible profit. [40] claims that a transportation algorithm should be used to compute the maximum possible profit when multiple transactions occur within rolling six-month windows. Due to the sheer number of transactions, we only consider the consecutive transactions for simplicity.

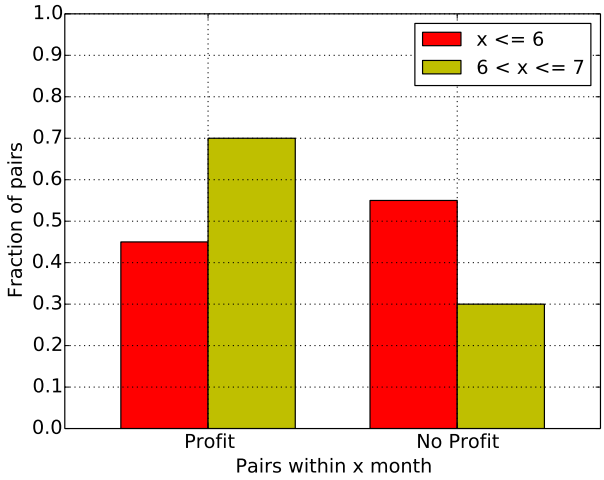



Figure 38: Fraction of consecutive opposite transaction pairs (P→S and S→P) that are profitable versus unprofitable. 45% of the pairs that occur within a 6-month period are profitable despite the short-swing profit rule, which requires insiders to forfeit profit from trades that occur within six months of each other.

Figure 38 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 greater than 6 but less than or equal to 7 months apart (the rule no longer applies).[5] Interestingly, approximately 45% of the pairs containing transactions that occur within six 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 $H_a : \mu_{profit} \neq 0$ 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 six months. While the data indicates that the short-swing profit rule may not completely deter

[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 product-moment correlation coefficient value of 0.12 indicates positive correlation between profit and number of shares traded ($p < 0.01$).

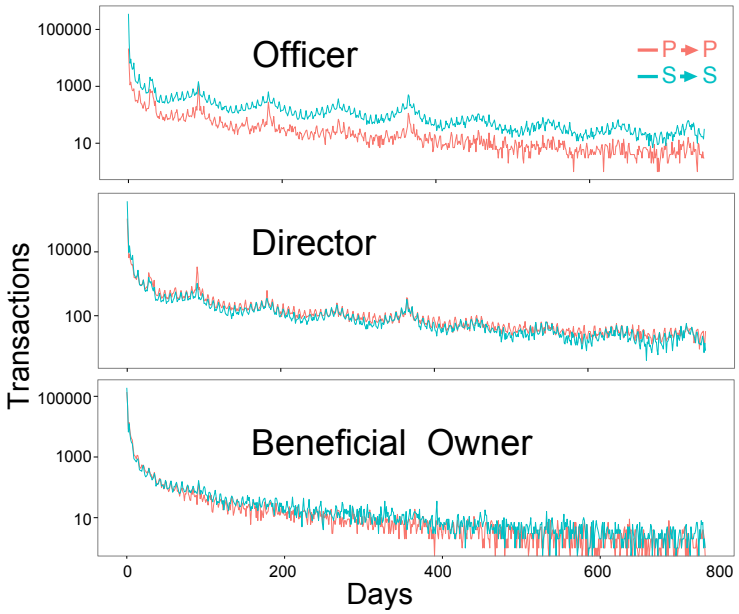



Figure 39: Transaction intervals for different role codes. Insiders in different roles trade differently.

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 Figure 39. 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 subject to corporate bylaws—though some beneficial owners are other companies rather than individuals. We further observe that the patterns for the other types insiders differ amongst themselves. For example, officers have significantly more S→S sequences than P→P sequences. This, again, is likely related to the stock options and grants given to the officers as part of their compensation package. Directors are generally

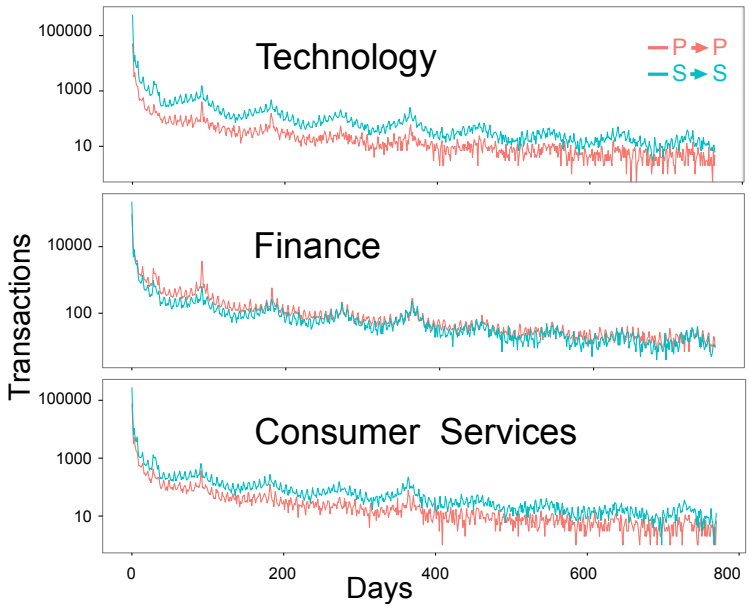



Figure 40: Transaction intervals for different sectors. Insiders in different sectors trade differently.

fewer in number and typically do not receive as much stock compensation.

Figure 40 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 are 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.

### 7.4.3 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 [7] and accounting scandals [6], and attempted to determine if 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 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.

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 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

[7]The dollar volume of a stock is a measure of its liquidity on a given day and it is computed by

**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** $TT =$ 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** $TT =$ 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** $I$

6-7).

Note that our goal here is to aggregate the "signals" from all the transactions of the insider, possibly for different 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 (line 8). Note that this ratio is greater than 0 and almost always upper-bounded by 1,[8] and it denotes the "magnitude" of the transaction in dollars relative to the other

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.

[8]The scenarios leading to a ratio greater than 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.

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 to 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 problem[9] 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 [145]. 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

[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 [145].

Table 19: 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 |

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

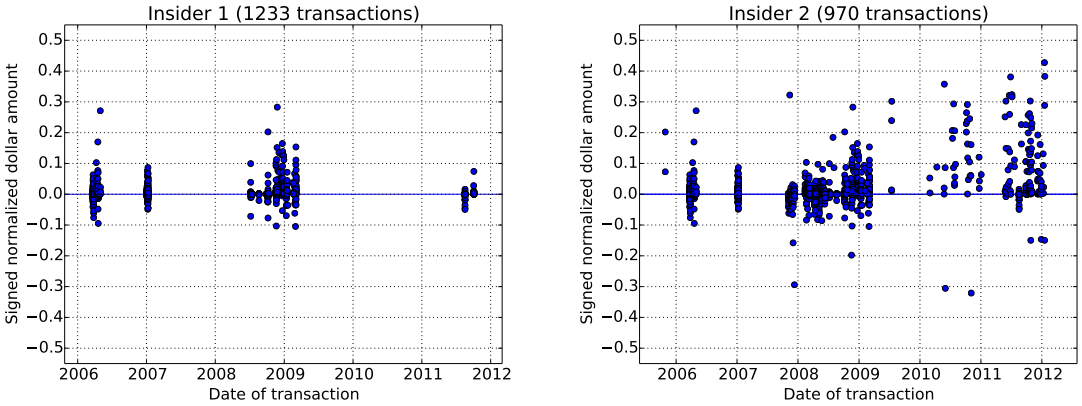
Table 19 lists the 29 insiders returned from the procedure with significant statistical results. The list is ranked in descending order according to the number of

[10] `www.crsp.uchicago.edu`

transactions. We also report if 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), 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 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, Figure 41 zooms in and shows the time series of the signed normalized dollar amounts for the transactions of the top-2 insiders in Table 19. Notice that the bulk of the transactions in both time series have positive normalized dollar amounts. This is particularly obvious for Insider 2, who almost consistently times his or 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.




(a) Insider with highest number of transactions (b) Insider with second-highest number of transactions

Figure 41: Time series of the signed normalized dollar amounts for the transactions of the top-2 insiders in Table 19; 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 to the market closing price. The bulk of the transactions are located above the straight line in both figures, illustrating that our approach can capture this trading behavior.

### 7.4.4 Constructing Networks of Insiders

We now study insider behavior from a graph-based perspective. We conjecture that insiders within and across companies may share nonpublic 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 work, 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. Note that trader

**Algorithm 2** 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** $G$

$T$ 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:

$$S(X_C, Y_C) = \frac{\left( \sum_{i=1}^{|X_C|} \sum_{j=1}^{|Y_C|} I(x_i, y_j) \right)^2}{|X_C| \times |Y_C|}, \tag{4}$$

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.

We now analyze two networks generated using the aforementioned process: the

Table 20: Simple network parameters for our Sale and Purchase networks.

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

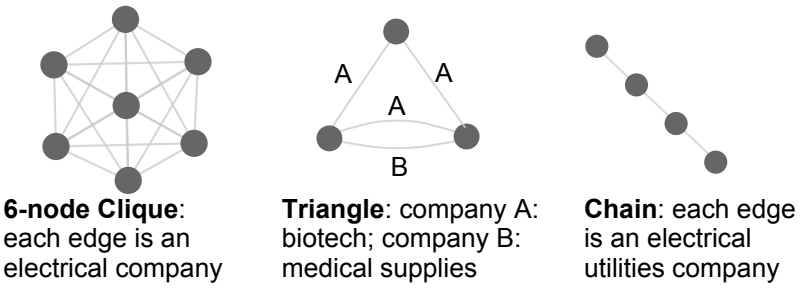



Figure 42: Examples of connected components from the Sale network. The insiders form different clusters in terms of shape.

*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 20 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

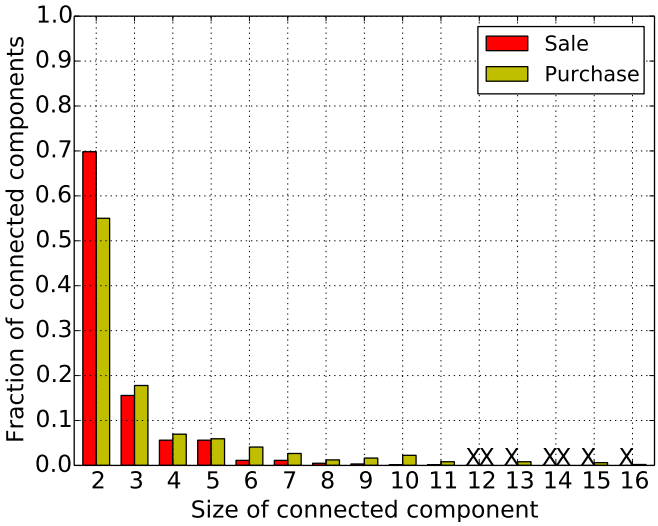



Figure 43: Distributions of the fraction of connected components with size of a particular value. "X" is used for values that are not applicable. Some insiders form large clusters in which trade-related information might propagate.

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 Figure 42. 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 Figure 43, 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 Figure 44, which is the largest connected component in the Purchase network.

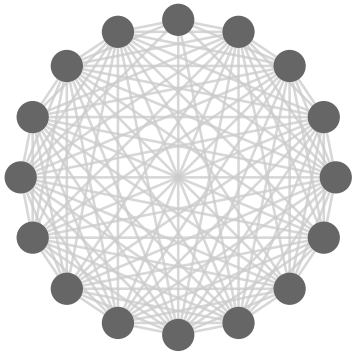



Figure 44: Largest connected component in the Purchase network: 16 insiders form a "trading clique".

Table 21: 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 Figure 42. Table 21 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? Figure 45 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 he

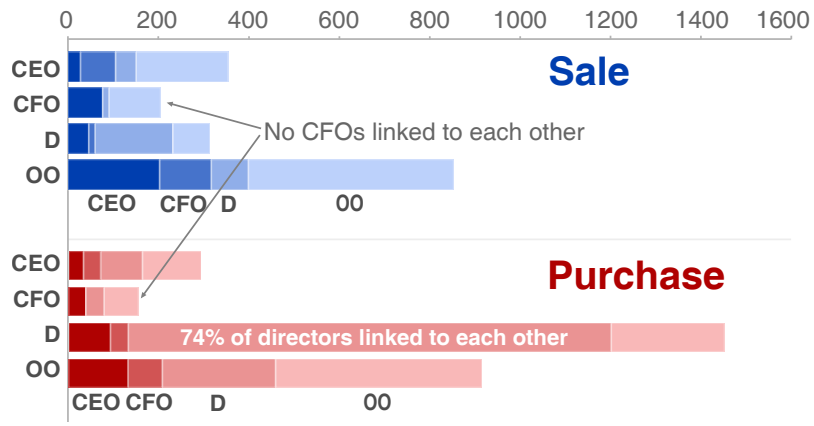



Figure 45: Counts for all combinations of *role pairs* (e.g., CEO-CFO, D-D), where D is *Director*, OO is *Other Officer*. High-level insiders (e.g., CEO, CFO) more likely to be linked to low-level insiders (e.g., Director).

or 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 Figure 46. 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 to 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

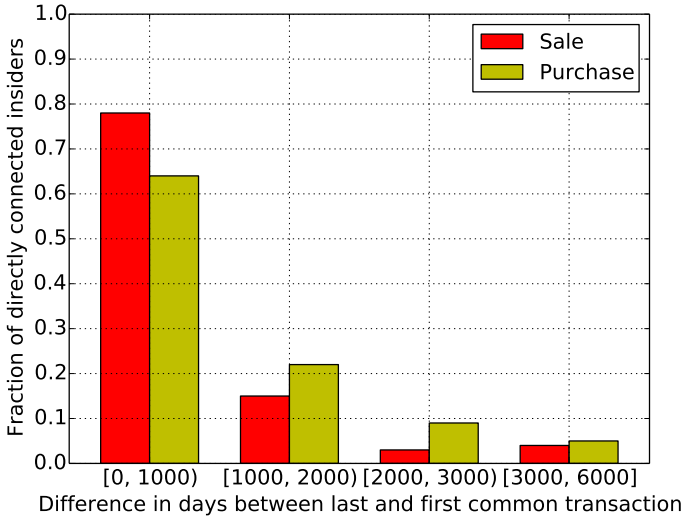



Figure 46: A comparison of the persistence of the similar trading behaviors of the insiders. The persistence is greater for purchase transactions.

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 $Y_C^1, Y_C^2, ..., Y_C^n$. We then take intersection of these $n$ sets, $I = Y_C^1 \cap Y_C^2 \cap ... \cap Y_C^n$, 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 his or 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 47 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 his or her neighbors also trade on that date. Note that our networks contain only the insiders with

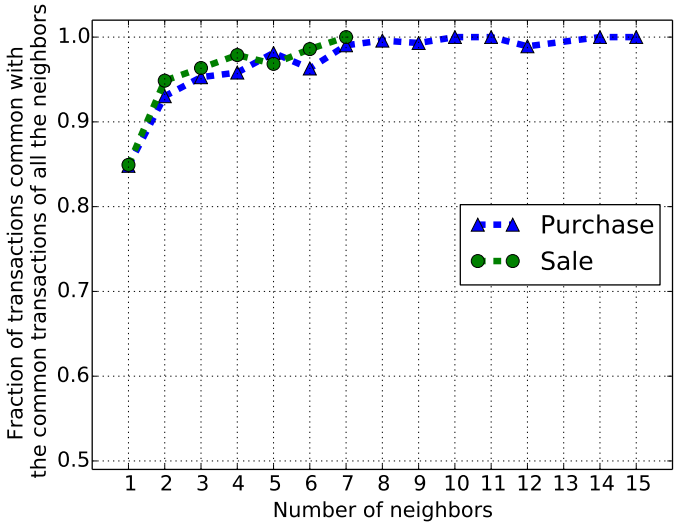



Figure 47: Collective trading behavior between the insiders and their neighbors: given that all the neighbors of an insider trade on a date, the insider is likely to trade on the same date.

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, therefore does not ensure collective trading behaviors between the insider and his or her neighbors. Similarly, the high clustering coefficients we observe for the connected components do not ensure collective trading behaviors across the whole spectrum of neighbor counts. 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.

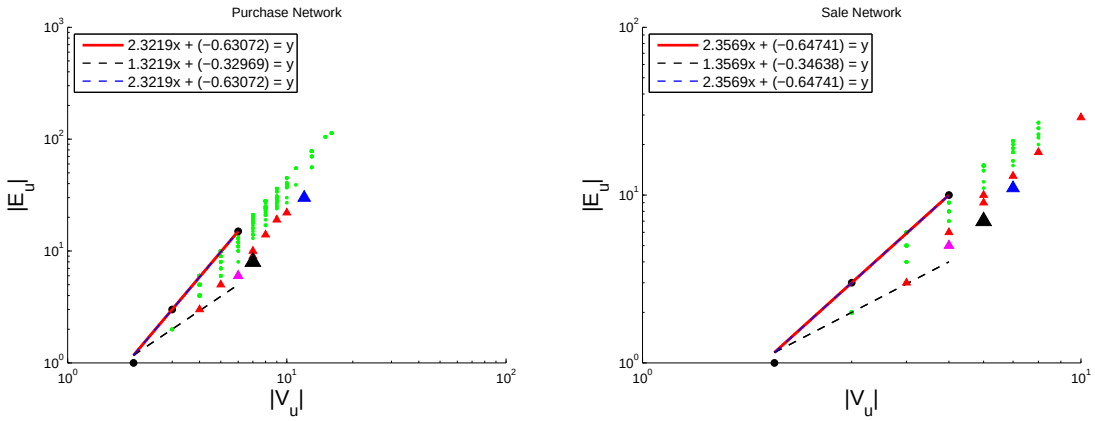
Some possible reasons for the collective trading behavior are the following. First, there might be information flow from the neighbors to the insiders. In other words, as the number of signals the insider receives increases, he or she is more willing to trade on a particular date. Second, the insider and his or her neighbors might have the same internal source of information. For instance, if both the insider and his or 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 his or 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.

### 7.4.5 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 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 [62] or edges [36]. Alternatively, random walk based methods have been suggested for identifying outliers in object similarity graphs [121], or bipartite graphs [162]. However, these methods exhibit some limitations: while we are interested in detecting anomalous nodes, i.e., insiders, [36] focuses on edges; the algorithm of [162] is designed for bipartite graphs, which does not apply to our networks; [62] 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 [121] is difficult to evaluate,




(a) Purchase Network (b) Sale Network

Figure 48: 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.

given that it requires parameter tuning, which can highly affect the results.

Akoglu et al. [10] 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 (i) it detects anomalous nodes in *general weighted* graphs, (ii) it does not assume any labels on the nodes, (iii) it yields results that are easy to interpret, and (iv) 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 [10] 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 Figure 48. 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


Figure 49: Insiders from several companies in different sectors/industries form a long chain in the Sale network.

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:

$$out - distance(u) = \frac{max(y_u, f(x_u))}{min(y_u, f(x_u))} \cdot log(|y_u - f(x_u)| + 1) \tag{5}$$

The value of *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 [10], 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, and look at some of the egos with the highest outlierness scores. In Figure 48, the ten most anomalous egos in each network are designated with larger triangles indicating higher outlierness scores. We discuss the interesting findings from this analysis in Section 7.5.

## 7.5 Notable Observations

In this section, we discuss interesting findings from our graph-based analysis and point out directions for future work. The graph-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 Figure 49 from the Sale network, which was found by our technique. At first glance, one may think that these insiders

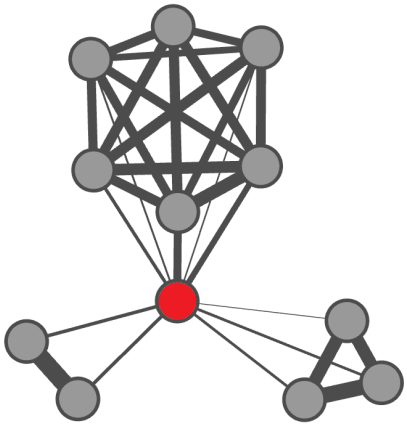


Figure 50: A visualization of the egonet of the middle node, flagged as anomalous by the method described in Section 7.4: the ego is connected to three cliques, which deviates from the pattern of the power-law fit for the Purchase network in Figure 48.

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 Section 7.4.5. Recall that this method flags nodes (or egos) whose neighborhoods' (or egonets) structures deviate from the general pattern across all nodes. In Figure 50, 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 directly connected to all the other nodes. The thickness of the edges is proportional to the value of the similarity function defined in Equation 4, 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 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.

## 7.6 Conclusions

This chapter presents the first academic, large-scale exploratory study of the complete insider trading data from SEC. We study the trades by insiders from temporal and grahp-based 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 who consistently trade on similar dates 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. We believe our work raises exciting research questions, opens up many opportunities for future studies, and has taken a major step towards helping financial regulators and policymakers understand the dynamics behind insider trading. The results of this work were presented to SEC.

# CHAPTER VIII

# CONCLUSIONS AND FUTURE DIRECTIONS

Graphs are now omnipresent, infusing into many aspects of our society. This thesis leverages graphs from the security, healthcare, and finance domains to benefit societies at large, by helping solve real-world problems affecting millions of individuals' daily lives, from cyber-attacks involving malware to tobacco and alcohol addiction. Our overarching goal is to help solve large-scale societal problems; in doing so, we take a graph-based perspective such that we represent the relationships between the entities central to the problems as well as information about the entities in the form of graphs, based on which we design and develop algorithms and models that contribute towards solving these problems. Our research groups into two interrelated topics, which form the main thrusts of the thesis.

In the first part of the thesis, entitled "Propagation-based Graph Mining Algorithms", we design and develop graph mining algorithms to propagate the information we possess about the entities between the nodes of our graphs based on the graph structure. In this part, we describe several propagation-based graph mining algorithms, which we briefly mention below.

In Chapter 2, we describe our Aesop algorithm for malware detection, which leverages the co-occurrence relationships between the files. Aesop detected malware across over 43 million files both more accurately (achieving 99.61% true positive rate at 0.01% false positive rate vs. 76.74% true positive rate at 0.01% false positive rate) and sooner (flagging them at least one week sooner) than the state-of-the-art technique [38]. Aesop is patented, has been integrated into Symantec's antivirus technology, and protects over 120 million people worldwide from malware.

In Chapter 3, we describe the ADAGE algorithm, which systematically determines the appropriate intervals to construct a sequence of graph snapshots from streaming edges. ADAGE was developed in a joint effort led by our collaborators; we contributed mainly with an extensive case study on malware detection using a propagation-based algorithm. In this chapter, we discuss how leveraging the smaller snapshots of a machine-file graph generated from the intervals determined by ADAGE can enable us to detect malware more accurately—by propagating goodness scores between the files and the machines as prior work [38] does—in comparison to using the final, full graph that includes all the machine-file relationships. We validated our observation with an extensive case study over 574 thousand files, achieving an average of 74% true positive rate at 0.01% false positive rate with the smaller snapshots in comparison to 43% true positive rate at 0.01% false positive rate with the final graph. This observation we made is patent-pending.

In Chapter 4, we describe our EDOCS algorithm for comment spammer detection, which quantifies the effort scores of the social media users. EDOCS detected comment spammers across over 197 thousand users accurately with 95% true positive rate at 3% false positive rate as well as preemptively (i.e., it detected spammers early on), and it outperformed the existing technique used by Yahoo (exact performance details proprietary). EDOCS is patent-pending, has been integrated into Yahoo's anti-abuse technology for their social media platforms, and guards multiple online communities from comment spammers.

In the second part the thesis, entitled "Graph-induced Behavior Characterization", we derive new insights and knowledge that characterize certain behavior of the entities using statistical and algorithmic techniques that incorporate information from our graphs as well as other useful information about the entities that might be captured externally. In this part, we describe several graph-induced behavior characterizations, which we briefly mention below.

In Chapters 5 and 6, we provide one of the first attempts at understanding the smoking/drinking abstinence and relapse experiences of individuals from social media, and present quantitative insights into evaluating the effectiveness of social media support communities in promoting cessation. By leveraging self-reported abstinence information, we developed statistical models to analyze the role of social media language, interactions, and engagement in characterizing smoking/drinking abstinence and relapse. As an example, we found linguistic cues like affect, activity cues like tenure, and network features like indegree to be indicative of short-term or long-term abstinence. Based on participation to the communities we study, we determined that individuals who continue to abstain beyond three years tend to maintain high likelihood of sustained abstinence, suggesting the efficacy of the communities in preventing relapse in the long term. We also found positive affect and increased engagement to be predictors of abstinence.

In Chapter 7, we performed the first academic, large-scale exploratory study of the complete insider filings from SEC, and made surprising and counterintuitive discoveries. As an example, by analyzing the time series of the transactions, we determined that a significant portion of the insiders makes short-swing profits (i.e., profit resulting from a combined purchase and sale, or sale and purchase, of the company's stock within a 6-month period) despite the existence of a rule designed to prevent short-swing trading. Also, in our graph-based analysis, we found strong evidence that insiders form small clusters in which trade-related information might propagate both vertically (between higher-level and lower-level insiders) and horizontally (among lower-level insiders). The results of this work were presented to SEC.

## 8.1 Challenges Encountered

Next, we discuss the challenges we encountered in our work to provide guidance for future research similar to ours. As we tackled large-scale societal problems in this

thesis, we dealt with very large datasets in some of our work, therefore one of the challenges was to ensure that our algorithms can scale to large amounts of data. To overcome this challenge, we leveraged efficient approximate techniques such as locality-sensitive hashing and belief propagation in our algorithms. Another challenge was to ensure the integrity of the datasets we utilized in our work. For instance, in our work on addiction, our preliminary analysis revealed certain irregularities in the badge sequences of some users, and we defined heuristic rules to filter out those users with noisy badge sequences. Similarly, in our other work, we carefully inspected our datasets to eliminate noise as necessary. The other challenge was the class imbalance present in some of our datasets. Class imbalance occurs when there are significantly fewer data points of one class compared to other classes. This phenomenon is almost unavoidable in some datasets from certain domains, e.g., in the security domain, a dataset containing information about the files appearing on people's computers is likely to contain many more benign files than malicious files. We addressed this challenge by operating on clusters consisting of entities with identical labels (e.g., our Aesop algorithm first clusters the files into buckets consisting of co-occurring files with identical labels and it then establishes guilt-by-association within these clusters to detect malware), or by leveraging robust techniques that can handle class imbalance, such as statistical techniques from the survival analysis literature.

## 8.2 Future Research Directions

There are opportunities to push our research in several interesting directions in the future. We discuss them below.

In the first part of the thesis, we describe the Aesop algorithm for malware detection, the application of the ADAGE algorithm to malware detection, and the EDOCS algorithm for comment spammer detection. There are interesting future directions

for our work in this part. As one example related to Aesop, we would like to investigate whether the users' website co-visitation patterns, as captured by the DNS server queries, could be leveraged to detect malicious websites based on the guilt-by-association principle. Here, the idea would be that an unknown website that is consistently co-visited with the malicious websites by the users might also be malicious, as it might be needed by the latter websites for certain purposes (e.g., luring the users to click on a link directing to the malicious websites). In this setting, we believe Aesop could be used to detect malicious websites effectively and efficiently. As future work for ADAGE, we would like to investigate whether Aesop could be paired with ADAGE to accurately detect malware in a setting where the file-bucket relationships in the file-relation graph of Aesop are streaming. In EDOCS, we currently consider two important effort-requiring features; as future work, we plan to incorporate additional features to the algorithm to extend our definition of "effort" on social media.

In the second part of the thesis, we describe our study on identifying attributes of smoking and drinking abstinence and relapse from an addiction cessation social media community, and our exploratory analysis of how company insiders trade. There are interesting future directions for our work in this part. As one example related to our work on addiction, given the predictive capability of our statistical models, we would like to develop early warning systems that analyze patterns of activity on the social media platform and engage appropriately if the likelihood of relapse in the broader community increases beyond a certain level. If successful, we believe that such early warning systems could further provide scientific and clinical insights into understanding and identifying prospective factors associated with abstinence and relapse over time. Also, we intend to characterize the users' subsequent relapse events that occurred after their first observed relapse event. We acknowledge that a lexicon-driven

approach via Linguistic Inquiry and Word Count (LIWC) can have limitations in characterizing relapse. It is worthwhile to examine alternative lexica (e.g., POMS [117], PANAS [174]) that describe emotional states beyond those described by LIWC. For our work on insider trading, in the future we plan to consider the transactions that occur within a time window to capture additional patterns in our graph-based analysis. Additionally, we intend to incorporate the geographical location information of the insiders' companies into our analyses.

Behavior is likely to change over time. Our behavior characterizations in the second part of the thesis take into account time dimension to an extent; e.g., in our work on addiction, we consider longitudinal badges that indicate abstinence duration to determine the abstinence or relapse status of the individuals, and in our work on insider trading, we consider the transaction dates of the insiders to determine the insiders who consistently trade on similar trades, and therefore, might be sharing nonpublic inside information with each other. The graphs we leverage in this part of the thesis are static in that they aim to reflect all the information present in our datasets in a single snapshot; as such, they help us characterize the behavior of the entities observed within the whole duration of our datasets. A direction we intend to pursue in the future is to investigate if and how certain behavior changes over time. As an example, we are interested in characterizing how the behavior of the abstainers changes over time after failed attempts to abstain from smoking or drinking. For this purpose, we plan to leverage dynamic, time-evolving graph snapshots generated using the ADAGE algorithm in our behavior characterizations. Furthermore, we plan to make use of techniques from the temporal pattern mining literature [119].

It is also worthwhile to investigate how emerging graph databases such as Neo4j[1] and Apache Giraph[2] can be integrated into our algorithms and models. These

[1] `neo4j.com`
[2] `giraph.apache.org`

databases have a potential to assist us with validating the integrity of our datasets [146], and they can also help us further improve the scalability of our algorithms by storing the results of the graph operations that we perform frequently (e.g., finding the immediate neighbors of a node) to facilitate efficient retrieval for future use.

We would like to also explore how our algorithms and models can applied to other domains to tackle large-scale societal problems therein. One particular domain we intend to focus on is energy. The so-called smart grid is emerging in the energy domain as a solution to provide a reliable, efficient, and sustainable energy supply [69]. Traditional studies on smart grid tend to have a "local" view of the grid, focusing on its individual components such as a substation or transformer (see [66] for a survey). Recent work investigates the properties of the smart grid infrastructures from a "global" view obtained by representing the grid as a graph, with the nodes being the components of the grid and the edges representing the physical connections between the components with cables (see [133] for a survey). We believe the application of our graph-based algorithms and models have potential to help solve important problems in this context. One example is the malicious attacks against smart grids, in which an adversary controls a set of meters and is able to alter the measurements from those meters [102]. In this setting, we could first characterize the behavior of normal and anomalous meters using statistical models, and then use our algorithms to detect the anomalous meters controlled by the adversary. Another application area for our algorithms and models could be the new cross-domain paradigm of Internet of Things [16], where billions of sensors and devices are connected to each other, all sharing data via the Internet. For instance, in the telecommunications domain, answering the question "What cell tower is experiencing problems?" based on information from a cell phone network is of great interest to us, and we believe, fundamental importance to many.

# REFERENCES


[1] Abbar, S., Mejova, Y., and Weber, I., "You tweet what you eat: Studying food consumption through twitter," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3197–3206, ACM, 2015.

[2] Aboody, D. and Kasznik, R., "CEO stock option awards and the timing of corporate voluntary disclosures," *Journal of Accounting and Economics*, vol. 29, no. 1, pp. 73–100, 2000.

[3] Abu-Nimeh, S. and Chen, T. M., "Proliferation and detection of blog spam," *IEEE Security & Privacy*, vol. 8, no. 5, pp. 42–47, 2010.

[4] Adamic, L., Brunetti, C., Harris, J. H., and Kirilenko, A. A., "Trading networks," *Available at Social Science Research Network 1361184*, 2010.

[5] Aggarwal, C. C., *Social Network Data Analytics*. Springer Publishing Company, Incorporated, 1st ed., 2011.

[6] Agrawal, A. and Cooper, T., "Insider trading before accounting scandals," *Available at Social Science Research Network 929413*, 2008.

[7] Agrawal, A. and Nasser, T., "Insider trading in takeover targets," *Available at Social Science Research Network 1517373*, 2011.

[8] Akoglu, L. and Faloutsos, C., "RTG: A recursive realistic graph generator using random typing," *Data Mining and Knowledge Discovery*, vol. 19, no. 2, pp. 194–209, 2009.

[9] Akoglu, L., McGlohon, M., and Faloutsos, C., "RTM: Laws and a recursive generator for weighted time-evolving graphs," in *Proceedings of the 8th IEEE International Conference on Data Mining*, pp. 701–706, IEEE, 2008.

[10] Akoglu, L., McGlohon, M., and Faloutsos, C., "OddBall: Spotting anomalies in weighted graphs," in *Advances in Knowledge Discovery and Data Mining*, pp. 410–421, Springer, 2010.

[11] Albert, R., Jeong, H., and Barabási, A.-L., "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[12] An, L., Schillo, B., Saul, J., Wendling, A., Klatt, C., Berg, C., Ahluwalia, J., Kavanaugh, A., Christenson, M., and Luxenberg, M., "Utilization of smoking cessation informational, interactive, and online community resources as predictors of abstinence: Cohort study," *Journal of Medical Internet Research*, vol. 10, no. 5, p. e55, 2008.

[13] Anderson, D. S., Fleizach, C., Savage, S., and Voelker, G. M., "Spamscatter: Characterizing internet scam hosting infrastructure," in *Proceedings of the USENIX Security Symposium*, pp. 1–14, 2007.

[14] Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., and Feamster, N., "Building a dynamic reputation system for DNS," in *Proceedings of the USENIX Security Symposium*, pp. 273–290, 2010.

[15] Atkinson, K., *An Introduction to Numerical Analysis*. Wiley, 1989.

[16] Atzori, L., Iera, A., and Morabito, G., "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[17] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X., "Group formation in large social networks: Membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54, ACM, 2006.

[18] Bay, S., Kumaraswamy, K., Anderle, M. G., Kumar, R., and Steier, D. M., "Large scale detection of irregularities in accounting data," in *Proceedings of the 6th IEEE International Conference on Data Mining*, pp. 75–86, IEEE, 2006.

[19] Bell, K., Salmon, A., Bowers, M., Bell, J., and McCullough, L., "Smoking, stigma and tobacco denormalization: Further reflections on the use of stigma as a public health tool. A commentary on social science & medicine's stigma, prejudice, discrimination and health special issue (67: 3)," *Social Science & Medicine*, vol. 70, no. 6, pp. 795–799, 2010.

[20] Berger-Wolf, T. Y. and Saia, J., "A framework for analysis of dynamic social networks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 523–528, ACM, 2006.

[21] Berlingerio, M., Bonchi, F., Bringmann, B., and Gionis, A., "Mining graph evolution rules," in *Machine Learning and Knowledge Discovery in Databases*, pp. 115–130, Springer, 2009.

[22] Bernhardt, D., Hollifield, B., and Hughson, E., "Investment and insider trading," *The Review of Financial Studies*, vol. 8, no. 2, pp. 501–543, 1995.

[23] Beullens, K. and Schepers, A., "Display of alcohol use on facebook: A content analysis," *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 7, pp. 497–503, 2013.

[24] Bien, T. H. and Burge, R., "Smoking and drinking: A review of the literature," *Substance Use & Misuse*, vol. 25, no. 12, pp. 1429–1454, 1990.

[25] Bilge, L., Kirda, E., Kruegel, C., and Balduzzi, M., "EXPOSURE: Finding malicious domains using passive DNS analysis," in *Proceedings of the Network and Distributed System Security Symposium*, 2011.

[26] Bizjak, J., Lemmon, M., and Whitby, R., "Option backdating and board interlocks," *The Review of Financial Studies*, vol. 22, no. 11, pp. 4821–4847, 2009.

[27] Blanzieri, E. and Bryl, A., "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.

[28] Bleeping Computer, "Cryptolocker ransomware information guide and FAQ." `www.bleepingcomputer.com/virus-removal/cryptolocker-ransomware-information`, October 2013.

[29] Brockman, P., Li, X., and Price, S. M., "Do managers put their money where their mouths are? Evidence from insider trading after conference calls," *Available at Social Science Research Network 2200639*, 2013.

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

[31] Broder, A. Z., "On the resemblance and containment of documents," in *Proceedings of the Compression and Complexity of Sequences*, pp. 21–29, IEEE, 1997.

[32] Bryant, D., "Recent developments in manpower research," *Personnel Review*, vol. 1, no. 3, pp. 14–31, 1972.

[33] CALO, "Enron email dataset." `www.cs.cmu.edu/~./enron/`, 2009.

[34] Carmody, T. P., "Affect regulation, tobacco addiction, and smoking cessation," *Journal of Psychoactive Drugs*, vol. 21, no. 3, pp. 331–342, 1989.

[35] Cavazos-Rehg, P., Krauss, M., Grucza, R., and Bierut, L., "Characterizing the followers and tweets of a marijuana-focused twitter handle," *Journal of Medical Internet Research*, vol. 16, no. 6, p. e157, 2014.

[36] Chakrabarti, D., "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, 2004.

[37] Charikar, M. S., "Similarity estimation techniques from rounding algorithms," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 380–388, ACM, 2002.

[38] Chau, D. H., Nachenberg, C., Wilhelm, J., Wright, A., and Faloutsos, C., "Polonium: Tera-scale graph mining and inference for malware detection," in *Proceedings of the 11th SIAM International Conference on Data Mining*, pp. 131–142, 2011.

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

[40] Chin, A., "Accurate calculation of short-swing profits under section 16(b) of the Securities Exchange Act of 1934," *Delaware Journal of Corporate Law*, vol. 22, no. 32, pp. 587–599, 1997.

[41] Christakis, N. A. and Fowler, J. H., "The collective dynamics of smoking in a large social network," *New England Journal of Medicine*, vol. 358, no. 21, pp. 2249–2258, 2008.

[42] Chum, O., Philbin, J., and Zisserman, A., "Near duplicate image detection: min-hash and tf-idf weighting," in *Proceedings of the British Machine Vision Conference*, pp. 50.1–50.10, 2008.

[43] Chung, C. and Pennebaker, J. W., "The psychological functions of function words," *Social Communication*, pp. 343–359, 2007.

[44] Clark, L., Robbins, T. W., Ersche, K. D., and Sahakian, B. J., "Reflection impulsivity in current and former substance users," *Biological Psychiatry*, vol. 60, no. 5, pp. 515–522, 2006.

[45] Cobb, N. K., Graham, A. L., and Abrams, D. B., "Social network structure of a large online community for smoking cessation," *American Journal of Public Health*, vol. 100, no. 7, pp. 1282–1289, 2010.

[46] Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J. D., and Yang, C., "Finding interesting associations without support pruning," in *Proceedings of the IEEE International Conference on Data Engineering*, pp. 489–500, 2000.

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

[48] Cook, S. H., Bauermeister, J. A., Gordon-Messer, D., and Zimmerman, M. A., "Online network influences on emerging adults alcohol and drug use," *Journal of Youth and Adolescence*, vol. 42, no. 11, pp. 1674–1686, 2013.

[49] Coppersmith, G., Harman, C., and Dredze, M., "Measuring post traumatic stress disorder in twitter," in *Proceedings of the 8th International Conference on Weblogs and Social Media*, 2014.

[50] Cormack, G. V., Gómez Hidalgo, J. M., and Sánz, E. P., "Spam filtering for short messages," in *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, pp. 313–320, ACM, 2007.

[51] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C., *Introduction to algorithms*. MIT press, 2009.

[52] Cox, D. R. and Oakes, D., *Analysis of Survival Data*. Chapman & Hall, 1984.

[53] Das, A. S., Datar, M., Garg, A., and Rajaram, S., "Google news personalization: Scalable online collaborative filtering," in *Proceedings of the 16th International Conference on World Wide Web*, pp. 271–280, ACM, 2007.

[54] Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S., "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the 20th Annual Symposium on Computational Geometry*, pp. 253–262, ACM, 2004.

[55] De Choudhury, M., Counts, S., and Horvitz, E., "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the 2013 ACM Annual Conference on Human Factors in Computing Systems*, pp. 3267–3276, ACM, 2013.

[56] De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A., "Characterizing and predicting postpartum depression from shared facebook data," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 626–638, ACM, 2014.

[57] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E., "Predicting depression via social media," in *Proceedings of the 7th International Conference on Weblogs and Social Media*, 2013.

[58] Dean, J. C. and Poremba, G. A., "The alcoholic stigma and the disease concept," *Substance Use & Misuse*, vol. 18, no. 5, pp. 739–751, 1983.

[59] Dinakar, K., Jones, B., Lieberman, H., Picard, R. W., Rosé, C. P., Thoman, M., and Reichart, R., "You too?! Mixed initiative LDA story-matching to help teens in distress," in *Proceedings of the 6th International Conference on Weblogs and Social Media*, 2012.

[60] Donoho, S., "Early detection of insider trading in option markets," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420–429, ACM, 2004.

[61] Dumitras, T. and Shou, D., "Toward a standard benchmark for computer security research: The worldwide intelligence network environment (WINE)," in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, pp. 89–96, ACM, 2011.

[62] Eberle, W. and Holder, L., "Discovering structural anomalies in graph-based data," in *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*, pp. 393–398, IEEE, 2007.

[63] Edelson, R. and Whisenant, S., "A study of companies with abnormally favorable patterns of executive stock option grant timing," *Available at Social Science Research Network 1326517*, 2009.

[64] Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., Stern, A., and others, "Health related virtual communities and electronic support groups: Systematic review of the effects of online peer to peer interactions," *BMJ*, vol. 328, no. 7449, p. 1166, 2004.

[65] Fan, W., Philip, S. Y., and Wang, H., "Mining extremely skewed trading anomalies," in *Advances in Database Technology-EDBT 2004*, pp. 801–810, Springer, 2004.

[66] Fang, X., Misra, S., Xue, G., and Yang, D., "Smart grid - The new and improved power grid: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 944–980, 2012.

[67] Faught, E., Duh, M. S., Weiner, J. R., Guerin, A., and Cunnington, M. C., "Nonadherence to antiepileptic drugs and increased mortality findings from the RANSOM study," *Neurology*, vol. 71, no. 20, pp. 1572–1578, 2008.

[68] Felzenszwalb, P. F. and Huttenlocher, D. P., "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[69] Fischer, U., Kaulakienė, D., Khalefa, M. E., Lehner, W., Pedersen, T. B., Šikšnys, L., and Thomsen, C., "Real-time business intelligence in the MIRABEL smart grid system," in *Enabling Real-Time Business Intelligence*, pp. 1–22, Springer, 2012.

[70] Fisher, R. and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research*. London, Edinburgh, 3rd ed., 1948.

[71] Fox, S., *The social life of health information*. Pew Internet & American Life Project Washington, DC, 2011.

[72] Galea, S., Nandi, A., and Vlahov, D., "The social epidemiology of substance use," *Epidemiologic Reviews*, vol. 26, no. 1, pp. 36–52, 2004.

[73] Gilpin, E. A., Pierce, J. P., and Farkas, A. J., "Duration of smoking abstinence and success in quitting," *Journal of the National Cancer Institute*, vol. 89, no. 8, pp. 572–576, 1997.

[74] Gionis, A., Indyk, P., and Motwani, R., "Similarity search in high dimensions via hashing," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 518–529, 1999.

[75] Goeman, J. J., "L1 penalized estimation in the Cox proportional hazards model," *Biometrical Journal*, vol. 52, no. 1, pp. 70–84, 2010.

[76] Goldberg, H. G., Kirkland, J. D., Lee, D., Shyr, P., and Thakker, D., "The NASD securities observation, new analysis and regulation system (SONAR)," in *Proceedings of the Conference on Innovative Applications of Artificial Intelligence*, pp. 11–18, 2003.

[77] Grimes, A., Landry, B. M., and Grinter, R. E., "Characteristics of shared health reflections in a local community," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pp. 435–444, ACM, 2010.

[78] Hamilton, J., "The SEC's new approach to fraud." `www.businessweek.com/magazine/the-secs-new-approach-to-fraud-12082011.html`, 2011. Accessed March 22, 2013.

[79] Hanson, C. L., Cannon, B., Burton, S., and Giraud-Carrier, C., "An exploration of social circles and prescription drug abuse through twitter," *Journal of Medical Internet Research*, vol. 15, no. 9, p. e189, 2013.

[80] Harrell, F. E., *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, 2001.

[81] Hartzler, A. and Pratt, W., "Managing the personal side of health: How patient expertise differs from the expertise of clinicians," *Journal of Medical Internet Research*, vol. 13, no. 3, p. e62, 2011.

[82] Harwood, H. J., *Updating estimates of the economic costs of alcohol abuse in the United States: Estimates, update methods, and data*. NIH Publication No. 98-4327, 2000.

[83] Hoerl, A. E. and Kennard, R. W., "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[84] Holme, P., "Epidemiologically optimal static networks from temporal network data," *PLoS Computational Biology*, vol. 9, no. 7, p. e1003142, 2013.

[85] Homan, C. M., Lu, N., Tu, X., Lytle, M. C., and Silenzio, V., "Social structure and depression in trevorspace," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 615–625, ACM, 2014.

[86] Høybye, M. T., Johansen, C., and Tjørnhøj-Thomsen, T., "Online interaction: Effects of storytelling in an internet breast cancer support group," *Psycho-Oncology*, vol. 14, no. 3, pp. 211–220, 2005.

[87] Hu, X., Bhatkar, S., Griffin, K., and Shin, K. G., "MutantX-S: Scalable malware clustering based on static features," in *Proceedings of the USENIX Security Symposium*, pp. 187–198, 2013.

[88] Hu, X., Chiueh, T.-c., and Shin, K. G., "Large-scale malware indexing using function-call graphs," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pp. 611–620, ACM, 2009.

[89] Huh, J. and Ackerman, M. S., "Collaborative help in chronic disease management: Supporting individualized problems," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pp. 853–862, ACM, 2012.

[90] Huh, J., Liu, L. S., Neogi, T., Inkpen, K., and Pratt, W., "Health vlogs as social support for chronic illness management," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 4, p. 23, 2014.

[91] Hutto, C. J. and Gilbert, E., "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the 8th International Conference on Weblogs and Social Media*, 2014.

[92] Idika, N. and Mathur, A. P., "A survey of malware detection techniques," *Purdue University*, vol. 48, 2007.

[93] Indyk, P. and Motwani, R., "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pp. 604–613, ACM, 1998.

[94] Johnson, G. J. and Ambrose, P. J., "Neo-tribes: The power and potential of online communities in health care," *Communications of the ACM*, vol. 49, no. 1, pp. 107–113, 2006.

[95] Kantchelian, A., Ma, J., Huang, L., Afroz, S., Joseph, A., and Tygar, J., "Robust detection of comment spam using entropy rate," in *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence*, pp. 59–70, ACM, 2012.

[96] Karampatziakis, N., Stokes, J. W., Thomas, A., and Marinescu, M., "Using file relationships in malware classification," in *Proceedings of the SIG SIDAR Conference on Detection of Intrusions and Malware & Vulnerability Assessment*, pp. 1–20, 2012.

[97] Kaskutas, L. A., Bond, J., and Humphreys, K., "Social networks as mediators of the effect of alcoholics anonymous," *Addiction*, vol. 97, no. 7, pp. 891–900, 2002.

[98] Khandani, A. E. and Lo, A. W., "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.

[99] Kim, H. and Park, H., "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[100] Kirkos, E., Spathis, C., and Manolopoulos, Y., "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.

[101] Koller, D., Friedman, N., Getoor, L., and Taskar, B., "Graphical models in a nutshell," in *Introduction to Statistical Relational Learning* (Getoor, L. and Taskar, B., eds.), MIT Press, 2007.

[102] Kosut, O., Jia, L., Thomas, R. J., and Tong, L., "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011.

[103] Kottke, T. E., Battista, R. N., DeFriese, G. H., and Brekke, M. L., "Attributes of successful smoking cessation interventions in medical practice: A meta-analysis of 39 controlled trials," *JAMA*, vol. 259, no. 19, pp. 2882–2889, 1988.

[104] Krings, G., Karsai, M., Bernhardsson, S., Blondel, V. D., and Saramäki, J., "Effects of time window size and placement on the structure of an aggregated communication network," *EPJ Data Science*, vol. 1, no. 4, pp. 1–16, 2012.

[105] Kumar, R., Novak, J., and Tomkins, A., "Structure and evolution of online social networks," in *Link Mining: Models, Algorithms, and Applications*, pp. 337–357, Springer, 2010.

[106] Lakonishok, J. and Lee, I., "Are insider trades informative?," *The Review of Financial Studies*, vol. 14, no. 1, pp. 79–111, 2001.

[107] Leskovec, J., Kleinberg, J., and Faloutsos, C., "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 2, 2007.

[108] Lichtenstein, E. and Glasgow, R. E., "Smoking cessation: What have we learned over the past decade?," *Journal of Consulting and Clinical Psychology*, vol. 60, no. 4, p. 518, 1992.

[109] Lorie, J. H. and Niederhoffer, V., "Predictive and statistical properties of insider trading," *Journal of Law and Economics*, vol. 11, no. 1, pp. 35–53, 1968.

[110] MacLean, D., Gupta, S., Lembke, A., Manning, C., and Heer, J., "Forum77: An analysis of an online health forum dedicated to addiction recovery," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 1511–1526, ACM, 2015.

[111] Mamykina, L., Miller, A. D., Mynatt, E. D., and Greenblatt, D., "Constructing identities through storytelling in diabetes management," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1203–1212, ACM, 2010.

[112] Mamykina, L., Nakikj, D., and Elhadad, N., "Collective sensemaking in online health forums," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3217–3226, ACM, 2015.

[113] Mankoff, J., Kuksenok, K., Kiesler, S., Rode, J. A., and Waldman, K., "Competing online viewpoints and models of chronic illness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 589–598, ACM, 2011.

[114] Marlatt, G. and Donovan, D., *Relapse prevention: Maintenance strategies in treatment of addictive behaviors*. Guilford Publications, 2nd ed., 2005.

[115] Marlatt, G. A., Curry, S., and Gordon, J., "A longitudinal analysis of unaided smoking cessation," *Journal of Consulting and Clinical Psychology*, vol. 56, no. 5, p. 715, 1988.

[116] McGlohon, M., Bay, S., Anderle, M. G., Steier, D. M., and Faloutsos, C., "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*, pp. 1265–1274, ACM, 2009.

[117] McNair, D. M., Lorr, M., Droppleman, L. F., and others, *Profile of mood states*. Educational and Industrial Testing Service San Diego, CA, 1981.

[118] Mishne, G., Carmel, D., Lempel, R., and others, "Blocking blog spam with language model disagreement," in *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, pp. 1–6, 2005.

[119] Mitsa, T., *Temporal data mining*. CRC Press, 2010.

[120] Mokdad, A. H., Marks, J. S., Stroup, D. F., and Gerberding, J. L., "Actual causes of death in the United States, 2000," *Journal of the American Medical Association*, vol. 291, no. 10, pp. 1238–1245, 2004.

[121] Moonesinghe, H. and Tan, P.-N., "OutRank: A graph-based outlier detection framework using random walk," *International Journal on Artificial Intelligence Tools*, vol. 17, no. 01, pp. 19–36, 2008.

[122] Moreno, M. A., Christakis, D. A., Egan, K. G., Brockman, L. N., and Becker, T., "Associations between displayed alcohol references on facebook and problem drinking among college students," *Archives of Pediatrics & Adolescent Medicine*, vol. 166, no. 2, pp. 157–163, 2011.

[123] Moreno, M. A., D'Angelo, J., Kacvinsky, L. E., Kerr, B., Zhang, C., and Eickhoff, J., "Emergence and predictors of alcohol reference displays on facebook during the first year of college," *Computers in Human Behavior*, vol. 30, pp. 87–94, 2014.

[124] Murnane, E. L. and Counts, S., "Unraveling abstinence and relapse: Smoking cessation reflected in social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1345–1354, ACM, 2014.

[125] Myslín, M., Zhu, S.-H., Chapman, W., and Conway, M., "Using twitter to examine smoking behavior and perceptions of emerging tobacco products," *Journal of Medical Internet Research*, vol. 15, no. 8, p. e174, 2013.

[126] National Economic Council, Council of Economic Advisers, and Office of Science and Technology Policy, "A Strategy for American Innovation: Securing Our Economic Growth and Prosperity." `www.whitehouse.gov/sites/default/files/uploads/InnovationStrategy.pdf`, 2011. Accessed August 24, 2015.

[127] Newman, M. W., Lauterbach, D., Munson, S. A., Resnick, P., and Morris, M. E., "It's not that i don't have problems, i'm just not putting them on facebook: Challenges and opportunities in using online social networks for health," in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pp. 341–350, ACM, 2011.

[128] Niaura, R. S., Rohsenow, D. J., Binkoff, J. A., Monti, P. M., Pedraza, M., and Abrams, D. B., "Relevance of cue reactivity to understanding alcohol and smoking relapse," *Journal of Abnormal Psychology*, vol. 97, no. 2, pp. 133–152, 1988.

[129] Odean, T., "Are investors reluctant to realize their losses?," *The Journal of Finance*, vol. 53, no. 5, pp. pp. 1775–1798, 1998.

[130] Ossip-Klein, D. J., Bigelow, G., Parker, S. R., Curry, S., Hall, S., and Kirkland, S., "Task force 1: Classification and assessment of smoking behavior," *Health Psychology*, 1986.

[131] Ott, M., Cardie, C., and Hancock, J. T., "Negative deceptive opinion spam," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2013.

[132] Ott, M., Choi, Y., Cardie, C., and Hancock, J. T., "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 309–319, Association for Computational Linguistics, 2011.

[133] Pagani, G. A. and Aiello, M., "The power grid as a complex network: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688–2700, 2013.

[134] Pagano, M. E., Friend, K. B., Tonigan, J. S., and Stout, R. L., "Helping other alcoholics in alcoholics anonymous and drinking outcomes: Findings from project MATCH," *Journal of Studies on Alcohol*, vol. 65, no. 6, pp. 766–773, 2004.

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

[136] Papalexakis, E. E., Dumitras, T., Chau, D. H. P., Prakash, B. A., and Faloutsos, C., "Spatio-temporal mining of software adoption & penetration," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 878–885, ACM, 2013.

[137] Park, M., McDonald, D. W., and Cha, M., "Perception differences between the depressed and non-depressed users in twitter," in *Proceedings of the 7th International Conference on Weblogs and Social Media*, 2013.

[138] Partner, J., Vukotic, A., Watt, N., Abedrabbo, T., and Fox, D., *Neo4j in Action*. Manning Publications, 2014.

[139] Paul, M. J. and Dredze, M., "You are what you tweet: Analyzing twitter for public health," in *Proceedings of the 5th International Conference on Weblogs and Social Media*, 2011.

[140] Pennebaker, J. W. and Chung, C. K., "Expressive writing, emotional upheavals, and health," *Foundations of Health Psychology*, pp. 263–284, 2007.

[141] Preece, J. and Maloney-Krichmar, D., "Online communities: Design, theory, and practice," *Journal of Computer-Mediated Communication*, vol. 10, no. 4, pp. 00–00, 2005.

[142] Rajaraman, A. and Ullman, J. D., *Mining of Massive Datasets*. Cambridge University Press, 2012.

[143] Ramachandran, A. and Feamster, N., "Understanding the network-level behavior of spammers," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 4, pp. 291–302, 2006.

[144] Reis, V. L. D. and Culotta, A., "Using matched samples to estimate the effects of exercise on mental health from twitter," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 182–188, AAAI Press, 2015.

[145] Robert S. Witte, J. S. W., *Statistics*. Wiley, 9 ed., 2009.

[146] Robinson, I., Webber, J., and Eifrem, E., *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, Inc., 2015.

[147] Rodgers, S. and Chen, Q., "Internet community group participation: Psychosocial benefits for women with breast cancer," *Journal of Computer-Mediated Communication*, vol. 10, no. 4, pp. 00–00, 2005.

[148] Rowe, P., *Essential statistics for the pharmaceutical sciences*. John Wiley & Sons, 2007.

[149] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A., "Correlating financial time series with micro-blogging activity," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 513–522, ACM, 2012.

[150] Schoenborn, C., Adams, P., and Peregoy, J., "Health behaviors of adults: United States, 2008-2010," *Vital and Health Statistics*, vol. 10, no. 257, pp. 1–184, 2013.

[151] Sculley, D. and Wachman, G. M., "Relaxed online SVMs for spam filtering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 415–422, ACM, 2007.

[152] Sedgewick, R. and Wayne, K., *Algorithms*. Pearson Education, 4th ed., 2011.

[153] Shiffman, S., "Relapse following smoking cessation: A situational analysis," *Journal of Consulting and Clinical Psychology*, vol. 50, no. 1, pp. 71–86, 1982.

[154] Shiffman, S., "Reflections on smoking relapse research," *Drug and Alcohol Review*, vol. 25, no. 1, pp. 15–20, 2006.

[155] Shiffman, S., Shumaker, S. A., Abrams, D. B., Cohen, S., Garvey, A., Grunberg, N. E., and Swan, G. E., "Task force 2: Models of smoking relapse," *Health Psychology*, 1986.

[156] Skeels, M. M., Unruh, K. T., Powell, C., and Pratt, W., "Catalyzing social support for breast cancer patients," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 173–182, ACM, 2010.

[157] Soundarajan, S., Tamersoy, A., Khalil, E., Eliassi-Rad, T., Chau, D. H., Gallagher, B., and Roundy, K., "Generating graph snapshots from streaming edge data," in *Proceedings of the 25th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, 2016.

[158] Suler, J., "The online disinhibition effect," *Cyberpsychology & Behavior*, vol. 7, no. 3, pp. 321–326, 2004.

[159] Sulo, R., Berger-Wolf, T., and Grossman, R., "Meaningful selection of temporal resolution for dynamic networks," in *Proceedings of the 8th Workshop on Mining and Learning with Graphs*, pp. 127–136, ACM, 2010.

[160] Summers, S. L. and Sweeney, J. T., "Fraudulently misstated financial statements and insider trading: An empirical analysis," *Accounting Review*, pp. 131–146, 1998.

[161] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S., "GraphScope: Parameter-free mining of large time-evolving graphs," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 687–696, ACM, 2007.

[162] Sun, J., Qu, H., Chakrabarti, D., and Faloutsos, C., "Neighborhood formation and anomaly detection in bipartite graphs," in *Proceedings of the 5th IEEE International Conference on Data Mining*, pp. 418–425, IEEE, 2005.

[163] Symantec, "Internet security threat report." `www.symantec.com/security_response/publications/archives.jsp`, 2013.

[164] Tamersoy, A., De Choudhury, M., and Chau, D. H., "Characterizing smoking and drinking abstinence from social media," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 139–148, ACM, 2015.

[165] Tamersoy, A., Khalil, E., Xie, B., Lenkey, S. L., Routledge, B. R., Chau, D. H., and Navathe, S. B., "Large-scale insider trading analysis: patterns and discoveries," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–17, 2014.

[166] Tamersoy, A., Ouyang, H., and Chau, D. H., "Effort-based detection of comment spammers," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2015.

[167] Tamersoy, A., Roundy, K., and Chau, D. H., "Guilt by association: Large scale malware detection by mining file-relation graphs," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1524–1533, ACM, 2014.

[168] Tamersoy, A., Xie, B., Lenkey, S. L., Routledge, B. R., Chau, D. H., and Navathe, S. B., "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*, pp. 797–804, IEEE, 2013.

[169] Tantipathananandh, C., Berger-Wolf, T., and Kempe, D., "A framework for community identification in dynamic social networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 717–726, ACM, 2007.

[170] Tonigan, J. S. and Rice, S. L., "Is it beneficial to have an alcoholics anonymous sponsor?," *Psychology of Addictive Behaviors*, vol. 24, no. 3, p. 397, 2010.

[171] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H., "Recognizing depression from twitter activity," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3187–3196, ACM, 2015.

[172] U.S. Executive Office of the President, "Big data: Seizing opportunities, preserving values." `www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf`, 2014. Accessed July 26, 2015.

[173] U.S. Security and Exchange Comission, "Electronic data gathering, analysis, and retrieval (EDGAR) system." `www.sec.gov/edgar.shtml`.

[174] Watson, D., Clark, L. A., and Tellegen, A., "Development and validation of brief measures of positive and negative affect: The PANAS scales," *Journal of Personality and Social Psychology*, vol. 54, no. 6, p. 1063, 1988.

[175] Whitworth, A., Oberbauer, H., Fleischhacker, W., Lesch, O., Walter, H., Nimmerrichter, A., Platz, T., Fischer, F., and Potgieter, A., "Comparison of acamprosate and placebo in long-term treatment of alcohol dependence," *The Lancet*, vol. 347, no. 9013, pp. 1438–1442, 1996.

[176] Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., and Heywood, J., "Sharing health data for better outcomes on PatientsLikeMe," *Journal of Medical Internet Research*, vol. 12, no. 2, p. e19, 2010.

[177] Yedidia, J., Freeman, W., and Weiss, Y., *Understanding belief propagation and its generalizations*, pp. 239–270. Morgan Kaufmann Publishers Inc., 2003.

[178] Zhou, X., Nonnemaker, J., Sherrill, B., Gilsenan, A. W., Coste, F., and West, R., "Attempts to quit smoking and relapse: Factors associated with success or failure from the ATTEMPT cohort study," *Addictive Behaviors*, vol. 34, no. 4, pp. 365–373, 2009.