Understanding Propagation of Malicious Information Online

by

Hamidreza Alvari

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved February 2020 by the
Graduate Supervisory Committee:

Paulo Shakarian, Chair
Hasan Davulcu
Hanghang Tong
Scott Ruston

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

The recent proliferation of online platforms has not only revolutionized the way people communicate and acquire information but has also led to propagation of malicious information (e.g., online human trafficking, spread of misinformation, etc.). Propagation of such information occurs at unprecedented scale that could ultimately pose imminent societal-significant threats to the public. To better understand the behavior and impact of the malicious actors and counter their activity, social media authorities need to deploy certain capabilities to reduce their threats. Due to the large volume of this data and limited manpower, the burden usually falls to automatic approaches to identify these malicious activities. However, this is a subtle task facing online platforms due to several challenges: (1) malicious users have strong incentives to disguise themselves as normal users (e.g., intentional misspellings, camouflaging, etc.), (2) malicious users are high likely to be key users in making harmful messages go viral and thus need to be detected at their early life span to stop their threats from reaching a vast audience, and (3) available data for training automatic approaches for detecting malicious users, are usually either highly imbalanced (i.e., higher number of normal users than malicious users) or comprise insufficient labeled data.

To address the above mentioned challenges, in this dissertation I investigate the propagation of online malicious information from two broad perspectives: (1) *content* posted by users and (2) *information cascades* formed by resharing mechanisms in social media. More specifically, first, non-parametric and semi-supervised learning algorithms are introduced to discern potential patterns of human trafficking activities that are of high interest to law enforcement. Second, a time-decay causality-based framework is introduced for early detection of “Pathogenic Social Media (PSM)” accounts (e.g., terrorist supporters). Third, due to the lack of sufficient annotated data for training PSM detection approaches, a semi-supervised causal framework is proposed that utilizes causal-related attributes from unlabeled instances to compensate for the lack of enough labeled data. Fourth, a feature-driven approach for PSM detection is introduced that leverages different sets of attributes from users' causal activities, account-level and content-related information as well as those from URLs shared by users.

To my loving parents, wife and brother.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Paulo Shakarian, for his guidance, encouragement and support throughout my Ph.D. studies. I am very fortunate to be his student and will always be indebted to him. I would also like to thank my thesis committee, Dr. Hasan Davulcu, Dr. Hanghang Tong, and Dr. Scott Ruston, for their insightful feedback.

I would like to thank my lab mates at Cyber-Socio Intelligent Systems (CySIS) Lab, Eric Nunes, Ericsson Santana Marin, Elham Shaabani, Soumajyoti Sarkar, Mohammed Almukaynizi, Ahmad Diab, Ruocheng Guo, Ashkan Aleali, Abhinav Bhatnager and Vivin Paliath. I am also thankful to Jana Shakarian, Narges Masoumi, Rouzbeh Khodadadeh, Alireza Hajibagheri, Rahmatollah Beheshti, Erfan Davami and many more friends for this companionship.

Last but not least, this dissertation would not have been possible without company and support of my family. I owe the deepest gratitude to my beloved parents, Roya Joudaki and Abdolreza Alvari, for their unconditional support, and always believing in me. I would like to especially acknowledge my lovely, gorgeous and smart wife, Ghazaleh Beigi, whose eternal love, encouragement, support, patience, wit and delicious foods made this long journey possible and plausible. Words cannot express how much I owe to her as she literally scarified her life to support me during the hard times of the PhD life, while she was herself pursuing her PhD studies. I am extremely lucky to have her in life as my best friend and will always be indebted to her. Finally, many thanks to my brother and friend, Alireza Alvari, for his supports over the years and my little niece, Elma Alvari for sending me beautiful pictures of herself and her paintings. This dissertation is dedicated to all of them.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Recent years have witnessed an exponential growth of online platforms such as online social networks (OSNs), microblogging websites and other Web platforms. Nowadays, these platforms play a major role in online communication and information sharing as they have become massive-scale and real-time communication tools. This leads to huge user-generated data produced on a daily basis and in different forms, that are rich sources of information and can be used in different tasks from marketing to research. On the negative side, online platforms have become widespread tools exploited by various malicious actors who can orchestrate large-scale and societal-significant threats to public, ranging from online human trafficking Alvari *et al.* (2017) to misinformation spread Shao *et al.* (2017).

To better understand the behavior and impact of the malicious actors and counter their activity, social media and other online platforms' authorities need to deploy certain capabilities to reduce their threats. Due to the large volume of information published online and because of the limited manpower, the burden usually falls to algorithms that are designed to automatically identifying these bad actors. However, this is a subtle task facing online platforms due to several challenges: (1) malicious users have strong incentives to disguise themselves as normal users (e.g., intentional misspellings, camouflaging, etc.), (2) malicious users are high likely to be key users in making harmful messages go viral and thus need to be detected at their early life span to stop their threats from reaching a vast audience, and (3) available data for training automatic approaches for detecting malicious users, are usually either highly imbalanced (i.e., higher number of normal users than malicious users) or comprise insufficient labeled data.

In this dissertation, we address the aforementioned challenges by investigating the propagation of online malicious information from two broad aspects: (1) *content* posted by users and (2) *information cascades* formed by resharing mechanisms in social media. In particular, for the former, the problem of online human trafficking and potential countermeasures to combat them are studied. We present non-parametric and semi-supervised learning algorithms for detecting online human trafficking. For the latter, we study and understand "Pathogenic Social Media" (PSM) accounts who are likely to be key users in making malicious campaigns. Various machine learning-based algorithms are then presented to detect PSM accounts. In the followings, we first briefly explain each problem separately and then present research challenges. We conclude this chapter by providing the major contributions of this dissertation.

According to the United Nation uno (2011), human trafficking is defined as the modern slavery or the trade of humans mostly for the purpose of sexual exploitation and forced labor, via different improper ways including force, fraud and deception. The United States' Trafficking Victim Protection Act of 2000 tvp (2000) was the first U.S. legislation passed against human trafficking. Human trafficking has ever since received increased national and societal concern HTr (2015) but still demands persistent fight from all over the globe. Before the Internet, human traffickers were under risks of being arrested by law enforcement while advertising their victims on streets Desplaces (2012). However, move to the Internet has made it easier and less dangerous for sex sellers Nicholas D. (2012) as they no longer needed to advertise on the streets. There are numerous websites such as Backpage that host and provide sexual services under categories of escort, adult entertainment, massage services, etc., and help sex sellers and buyers maintain their anonymity.

Despite the above mentioned challenges facing law enforcement and presented by the Internet, on the positive side, it has provided readily and publicly available rich source of information which could be gleaned from online sex advertisements for fighting this crime Kennedy (2012). However, we lack the ground truth and obtaining the labels through hand-labeling is expensive even for a small subset of data. To overcome the issue of lacking ground truth, in Chapter 3 we rely on law enforcement experts for hand-labeling a small set of data. We then utilize the labeled and unlabeled data crawled from the adult entertainment section of the website Backpage and propose a non-parametric learner and a semi-supervised Laplacian SVM framework to detect escort advertisements of high interest to law enforcement. Here, we only focus on the textual *content* posted by users on Backpage and leave investigating other forms of user-generated data to the next chapters.

On the other hand, resharing mechanisms on social media such as retweeting in Twitter allow massive spread of harmful disinformation to viral proportions. Manipulating public opinion and political events on the Web can be attributed to accounts dedicated to spreading malicious information, referred to as "Pathogenic Social Media" (PSM) accounts (e.g., terrorist supporters, or fake news writers) Alvari *et al.* (2018). PSMs are users who seek to promote or degrade certain ideas by utilizing large online communities of supporters to reach their goals. Identifying PSMs has applications including countering terrorism Khader (2016); Klausen *et al.* (2016), fake news detection Gupta *et al.* (2014, 2013) and water armies detection Chen *et al.* (2011).

Identifying PSMs in social media is crucial as they are likely to be *key* users to malicious campaigns Varol *et al.* (2017b). This is a challenging task for several reasons. First, these platforms are primarily based on reports they receive from their own users [1] to manually shut down PSMs. This straightforward solution is not

[1] https://bit.ly/2Dq5i4M

necessarily a timely approach since despite efforts to suspend these accounts, many of them simply return to social media with different accounts which makes their manual suspension a non-trivial task. Second, available data for training automatic PSM detection approaches is often imbalanced and social network structure, which is at the core of many techniques Weng *et al.* (2014); Kempe *et al.* (2003); Zhang *et al.* (2013), is not readily available. Third, PSMs often seek to utilize and cultivate large number of online communities of passive supporters to spread as much harmful information as they can while disguising themselves as normal users. To address the aforementioned challenges, we propose several methods and algorithms to detect PSM accounts in their early life span.

1.1 Research Challenges

This dissertation addresses the following challenges facing online platforms in identifying malicious actors:

- Malicious users have strong incentives to disguise themselves as normal users (e.g., intentional misspellings Alvari *et al.* (2016b), camouflaging Hooi *et al.* (2016)). This makes the task of malicious users identification a daunting task. Later in Chapter 3, we observe that human traffickers would deploy techniques to generate diverse information to make their posts look more complicated and ensure their anonymity. We utilize *Kolmogorov complexity* Li and Vitányi (2008) from complexity theory to approximate the complexity of an advertisement content on Backpage.

- Malicious users are high likely to be key users in making harmful messages go "viral"– where "viral" is defined as an order-of-magnitude increase. Mechanisms are thus required to stop their threats from reaching vast audience early enough to stop formation of malicious campaigns Alvari *et al.* (2018). Accordingly, in Chapter 4, causal inference is tailored to identify PSMs since they are key users in making a harmful message viral. We propose *time-decay* causal metrics to distinguish PSMs from normal users within a *short* time around their activity. Our metrics alone can achieve high classification performance in identification of PSMs soon after they perform actions.

- Malicious users often seek to utilize and cultivate large number of online communities of passive supporters to spread as much harmful information. Consequently, in Chapter 4, we investigate whether or not causality scores of PSM users within same communities are higher than those across different communities? We propose a causal community detection-based classification method (C^2DC), that takes causality attribute vectors of users and the community structure of their action log.

- Available data for training automatic approaches for detecting malicious users, are usually either highly imbalanced (i.e., higher number of normal users than malicious users) or comprise insufficient labeled data Alvari *et al.* (2017, 2018). To overcome the issue of lack of enough annotated data, we present several semi-supervised based approaches for detecting human trafficking Alvari *et al.* (2016b, 2017) in Chapter 3 and PSMs Alvari *et al.* (2019b) in Chapter 5.

- Despite malicious users' incentives to disguise themselves as normal users, they still behave significantly different than normal users on many levels Xia *et al.* (2019). In Chapter 6, we take a closer look at the differences between malicious and normal behavior in terms of their posted URLs. We then leverage several characteristics of URLs as source-level information as well as other attributes from causal, profile and content-related information as input attributes to a supervised setting for detecting PSMs.

1.2 Contributions

Overall, this dissertation makes the following major contributions:

- We use the user-generated content posted on Backpage and present a semi-supervised Laplacian SVM to identify human trafficking-related posts that are of high interest to law enforcement. We trained our model on both of the *labeled* and *unlabeled* data from the website Backpage and sent back the identified human trafficking related advertisements to an expert from law enforcement for further verification. We finally validated our approach on a small subset of the unlabeled data (i.e. *unseen* data) with further verification of the expert.

- We leverage the rich information from cascade structure embedded in users' re-sharing interactions on Twitter and present *time-decay* causal metrics for early identification of PSMs, based on the Suppes' probabilistic causal theorem Suppes (1970). We further investigate the role of community structure in early detection of PSMs by demonstrating that users within a community establish stronger causal relationships compared to the rest. To account for this, we propose a causal community detection-based classification. We conduct a suit of experiments on a real-world dataset from Twitter. Our metrics reached F1-score of 0.6 in identifying PSMs, half way their activity, and identified 71% of PSMs based on first 10 days of their activity, via supervised settings. The community detection approach achieved precision of 0.84 based on first 10 days of users activity; the misclassified accounts were identified based on their activity of 10 more days.

- We frame the problem of detecting PSM accounts in the presence of far less number of labeled instances than unlabeled data, as an optimization problem and present a Laplacian semi-supervised causal inference SemiPsm for solving it. The unlabeled data are utilized via manifold regularization. Manifold regularization used in the resultant optimization formulation is built upon causality-based features created on a notion of Suppes' theory. We conduct a suite of experiments using different supervised and semi-supervised methods. Empirical experiments on a real-world ISIS-related dataset from Twitter suggests the effectiveness of the proposed semi-supervised causal inference over the existing methods.

- We study differences between malicious and normal behavior in terms of the URLs and platforms referenced. We then incorporate characteristics of URLs and their associated referenced Websites, as source-level attributes in a feature-driven approach for detecting PSMs in social media. More specifically, we assess the extent to which causal-level, account-level, source-level and content-level attributes contribute to identification of PSM accounts. Our causal and profile-related attributes investigate signals in causal users along with their profile information. For the source-level attributes, we explore different characteristics in URLs content that users share (e.g., underlying themes, complexity of text, etc.). For the content-level attributes, we examine attributes from tweets posted by users. We conduct a suite of experiments on three real-world Twitter datasets from different countries, using several classifiers. Using all of the attributes, we achieve average F1 scores of 0.81, 0.76 and 0.74 for Sweden, Latvia and UK datasets, respectively. Our observations suggest the effectiveness of the proposed method in identifying PSM accounts in real-world Twitter data.

1.3 Organization

The rest of this dissertation is organized as follows. Chapter 2 provides a review of the literature for identifying human trafficking and PSM accounts. In Chapter 3, we present semi-supervised methods and algorithms for identifying online human trafficking. Chapter 4 will detail our probabilistic causality-based methods for early identification of PSM accounts. Chapter 5 will present semi-supervised causal-based learning algorithms for detecting PSMs that uses far less number of labeled data than unlabeled examples for training. In Chapter 6, we take one more step further and present a hybrid feature-driven approach that using as little as four groups of attributes on causal, profile, source and content levels, outperforms baselines in detecting PSM accounts in Twitter data. Finally, Chapter 7 concludes the dissertation by presenting future directions for the algorithms proposed throughout the dissertation.

Chapter 2

RELATED WORK

This dissertation investigates the propagation of online malicious information from two broad aspects: (1) *content* posted by users and (2) *information cascades* formed by resharing mechanisms in social media. In particular, for the former, the problem of online human trafficking and potential countermeasures to combat them are studied. We present non-parametric Alvari *et al.* (2016b) and semi-supervised learning Alvari *et al.* (2017) algorithms for detecting online human trafficking. For the latter, we study and understand "Pathogenic Social Media" (PSM) accounts who are likely to be key users in making malicious campaigns. Various machine learning-based algorithms Alvari *et al.* (2018, 2019b, 2020) are then presented to detect PSM accounts. Our works on human trafficking and PSM detection are related to several research directions. Below, we discuss some of the state-of-the-art works in each category while highlighting the differences.

2.1 Human Trafficking

Recently, several studies have examined the role of the Internet and related technology in facilitating human trafficking Hughes *et al.* (2005); Hughes (2002); Latonero (2011). For example, the work of Hughes *et al.* (2005) studied how closely sex trafficking is intertwined with new technologies. According to Hughes (2002), sexual exploitation of women and children is a global human right crisis that is being escalated by the use of new technologies. Researchers have studied relationships between new technologies and human trafficking and advantages of the Internet for sex traffickers. For instance, findings from a group of experts from the Council of Europe demonstrated that the Internet and sex industry are closely interlinked and volume and content of the material on the Internet promoting human trafficking are unprecedented Latonero (2011).

One of the earliest works which leveraged data mining techniques for online human trafficking was Latonero (2011), wherein the authors conducted data analysis on the adult section of the website Backpage.com. Their findings confirmed that female escort post frequency would increase in Dallas, Texas, leading up to the Super Bowl 2011 event. In a similar attempt, other studies Roe--Sepowitz *et al.* (2015); Miller *et al.* (2016) have investigated impact of large public events such as the Super Bowl on sex trafficking by exploring advertisement volume, trends and movement of advertisements along with the scope and volume of demand associated with such events. The work of Roe--Sepowitz *et al.* (2015) for instance, concluded that large events such as the Super Bowl which attract significant amount of concentration of people in a relatively short period of time and in a confined urban area, could be a desirable location for sex traffickers to bring their victims for commercial sexual exploitation. Similarly, the data-driven approach of Miller *et al.* (2016) showed that in some but not all events, one can see a correlation between occurrence of the event and statistically significant evidence of an influx of sex trafficking activity. Also, certain studies Szekely *et al.* (2015) have tried to build large distributed systems to store and process available online human trafficking data in order to perform entity resolution and create ontological relations between entities.

Beyond these works, the work of Nagpal *et al.* (2015) studied the problem of isolating sources of human trafficking from online advertisements with a pairwise entity resolution approach. Specifically, they used phone number as a strong feature and trained a classifier to predict if two ads are from the same source. This classifier was then used to perform entity resolution using a heuristically learned value for the score of classifier. Another work of Kennedy (2012) used Backpage.com data and extracted most likely human trafficking spatio-temporal patterns with the help of law enforcement. Note that unlike our method, this work did not employ any machine learning methodologies for automatically identifying human trafficking related advertisements. The work of Dubrawski *et al.* (2015) also deployed machine learning for the advertisement classification problem, by training a supervised learning classifier on labeled data (based on phone numbers of known traffickers) provided by a victim advocacy group. We note that while phone numbers can provide a very precise set of positive labeled data, there are clearly many posts with previously unseen phone numbers.

In contrast, we do not solely rely on phone numbers for labeling our data. Instead, our expert analyze each post's content to identify whether it is human trafficking related or not. To do so, we first filter out most likely advertisements using several feature groups and pass a small sample to the expert for hand-labeling. Then, we train our semi-supervised learner on both of the labeled and unlabeled data which in turn lets us evaluate our approach on new coming (unseen) data later. We note that our semi-supervised approach can also be used as a complementary method to procedures such as those described in Dubrawski *et al.* (2015) as we can significantly expand the training set for use with supervised learning.

Finally, note that our semi-supervised approach Alvari *et al.* (2017) is different from our non-parametric method Alvari *et al.* (2016b) and we list the key nuances here:

- In Alvari *et al.* (2017) we experiment with a much larger dataset. To obtain such dataset, we use the same raw data from Alvari *et al.* (2016b), but this time with slight modifications of the thresholds that were used for filtering out less likely human trafficking related advertisements.

- As opposed to our previous research which deployed only one feature space, in this work, two feature spaces that have complementary roles to each other are used.

- In Alvari *et al.* (2017) we present a new framework based on the existing Laplacian SVM Belkin *et al.* (2006), by adding a regularization term to the standard optimization problem and solving the new optimization equation derived from there. In contrast, Alvari *et al.* (2016b) utilized the off-the-shelf graph based semi-supervised learner, LabelSpreading method Zhou *et al.* (2004b), without any further manipulation of the original approach.

- Unlike Alvari *et al.* (2016b) in which we did not compare our method with other approaches, Alvari *et al.* (2017) compares our proposed framework against other semi-supervised and supervised learners. Also unlike our previous work in which only *one* group of human trafficking related advertisements were passed to *two* experts for validation, here in order to reduce the inconsistency, *two* control groups of advertisements–those of interest to law enforcement and those of not–are sent to only *one* expert for verification.

2.2 Pathogenic Social Media Accounts

The explosive growth of the Web has raised numerous security and privacy issues. Mitigating these concerns has been studied from several aspects Beigi *et al.* (2018); Alvari *et al.* (2016b); Cao *et al.* (2014); Beigi and Liu (2018a); Cui *et al.* (2013); Beigi *et al.* (2014); Broniatowski *et al.* (2018); Beigi *et al.* (2019a); Alvari *et al.* (2019a); Beigi *et al.* (2020, 2019c). Our work is related to a number of research directions. Below, we will summarize some of the state-of-the-art methods in each category while highlighting their differences with our work.

Identifying PSM accounts. Compared to Shaabani *et al.* (2018) which uses causal inference to detect PSM accounts, works of Alvari *et al.* (2018); Alvari and Shakarian (2018) utilize time-decay causal inference (using sliding-time window) which allows early detection of PSM. Furthermore, in contrast to Alvari *et al.* (2018) where a causal community detection algorithm is proposed to leverage communities of PSM accounts in order to achieve higher performance,the work of Alvari *et al.* (2019b) proposes a semi-supervised causal inference algorithm that achieves reasonable performance using much less labeled data by utilizing unlabeled data. Also, a recent work of Shaabani *et al.* (2019) addresses the problem of detecting PSM accounts using a variety of supervised and semi-supervised algorithms using causality-based and graph-based metrics as attributes.

Social Spam/Bot Detection. Recently, DARPA organized a Twitter bot challenge to detect "influence bots" Subrahmanian *et al.* (2016). Among the participants, the work of Cao *et al.* (2014), used similarity to cluster accounts and uncover groups of malicious users. The work of Varol *et al.* (2017a) presented a supervised framework for bot detection which uses more than thousands features. In a different attempt, the work of Green and Spezzano (2017) studied the problem of spam detection in Wikipedia using different spammers behavioral features. There also exist some studies in the literature that have addressed (1) differences between humans and bots Chu *et al.* (2012), (2) different natures of bots Varol *et al.* (2017a) or (3) differences between bots and human trolls Broniatowski *et al.* (2018). For example the work of Chu *et al.* (2012) conducted a series of measurements in order to distinguish humans from bots and cyborgs, in term of tweeting behavior, content, and account properties. To do so, they used more than 40 million tweets posted by over 500 K users. Then, they performed analysis and find groups of features that are useful for classifying users into human, bots and cyborgs. They concluded that entropy and certain account properties can be very helpful in differentiating between those accounts. In a different attempt, some other studies have tried to differentiate between several natures of bots. For instance, in the work of Varol *et al.* (2017a), authors performed clustering analysis and revealed specific behavioral groups of accounts. Specifically, they identified different types of bots such as *spammers*, *self promoters*, and *accounts that post content from connected applications*, using manual investigation of samples extracted from clusters. Their cluster analysis emphasized that Twitter hosts a variety of users with diverse behaviors; that is in some cases the boundary between human and bot users is not sharp, i.e. some account exhibit characteristics of both.

Also, the work of Broniatowski *et al.* (2018), uses Twitter data to quantify the impact of Russian trolls and bots on amplifying polarizing and anti-vaccine tweets. They first used the Botometer API to assign bot probabilities to the users in the dataset and divided the whole dataset into 3 categories: those with scores less than 20% (very likely to be human), between 20% and 80% (e.g., cyborgs with uncertain provenance) and above 80% (high likely to be bots). Then, they posed two research questions: (1) Are bots and trolls more likely to tweet about vaccines?, and (2) Are bots and trolls more likely to tweet polarizing and anti-vaccine content? Their analysis demonstrated that Twitter bots and trolls significantly impact on online discussion about vaccination and this differs by account type. For example, Russian trolls and bots post content about vaccination at higher rates compared to an average user. Also, according to this study, troll accounts and content polluters (e.g., dissemination of malware, unsolicited commercial content, etc.) post anti-vaccine tweets 75% more than average users. In contrast, spambots which can be easily distinguished from humans, are less likely to promote anti-vaccine messages. Their closing remarks suggest strongly that distinguishing between malicious actors (bots, trolls, cyborgs, and human users) is difficult and thus anti-vaccine messages may be disseminated at higher rates by a combination of these malicious actors.

In contrast to the above works, our proposed PSM detection methods Alvari *et al.* (2018, 2019b) do not deploy any extra information (e.g., user-related attributes or network-based features) other than users' resharing actions (i.e., cascade with timestamps). It is also worthwhile to note that most of the existing well-known bot detection algorithms such as Botometer Davis *et al.* (2016) leverage over one thousand features in order to detect high-likely bots.

Fake News Identification. A growing body of research is addressing the impact of bots in manipulating political discussion, including the 2016 U.S. presidential election Shao *et al.* (2017) and the 2017 French election Ferrara (2017). For example, Shao *et al.* (2017) analyzes tweets following recent U.S. presidential election and found evidences that bots played key roles in spreading fake news.

Identifying Instigators. Given a snapshot of the diffusion process at a given time, these works aim to detect the source of the diffusion. For instance, Zhu and Ying (2016) designed an approach for information source detection and in particular initiator of a cascade. In contrast, we are focused on a set of users who *might* or *might not* be initiators. Other similar works on finding most influential spreaders of information such as Pei *et al.* (2014); Fu and Sun (2015) and outbreak prediction such as Cui *et al.* (2013) also exist in the literature. For example, the work of Konishi *et al.* (2016) performed classification to detect users who adopt popular items. In Zhu and Ying (2016), authors designed an approach for information source detection and in particular initiator of a cascade. Our proposed PSM detection methods Alvari *et al.* (2018, 2019b) are different from these works since we leverage causality analysis to detect causes of popularity of messages that go viral.

Extremism and Water Armies Detection. Several studies have focused on understanding extremism in social networks Benigni *et al.* (2017); Klausen *et al.* (2016);

Scanlon and Gerber (2014, 2015); Alvari *et al.* (2019a). The work of Klausen *et al.* (2016) uses Twitter and proposes an approach to predict new extremists, determine if the newly created account belongs to a suspended extremist, and predict the ego-network of the suspended extremist upon creating her new account. Authors in Benigni *et al.* (2017) performed iterative vertex clustering and classification to identify Islamic Jihadists on Twitter. The term "Internet water armies" refers to a special group of online users who get paid for posting comments for some hidden purposes such as influencing other users towards social events or business markets. Therefore, they are also called "hidden paid posters". The works of Chen *et al.* (2011, 2013); Wang *et al.* (2014) use user behavioral and domain-specific attributes and designed approaches to detect Internet water armies. The works of Chen *et al.* (2011); Wang *et al.* (2014) also used user behavioral and domain-specific attributes to detect water armies. Our proposed PSM detection methods Alvari *et al.* (2018, 2019b) also differ from these works as we do not use any features such as network/user attributes.

Causal Reasoning. As opposed to Kleinberg and Mishra (2012); Stanton *et al.* (2015); Kleinberg (2011) which deal with preconditions as single atomic propositions, in this dissertation, we use rules with preconditions of more than one atomic propositions.

Point Processes. In Chapter 6 we use point processes to differentiate between malicious and normal behaviors. When dealing with timestamped events in continuous time such as the activity of users on social media, point process could be leveraged for modeling such events. Point processes have been extensively used to model activities in networks Xiao *et al.* (2017). Hawkes process is a special form of point processes which models complicated event sequences with historical events influencing future ones. Hawkes processes have been applied to a variety of problems including financial analysis Bacry *et al.* (2016), seismic analysis Daley and Vere-Jones (2007) and social network modeling Zhou *et al.* (2013), community detection Tran *et al.* (2015), and causal inference Xu *et al.* (2016).

Chapter 3

SEMI-SUPERVISED LEARNING FOR DETECTING HUMAN TRAFFICKING

According to the United Nation uno (2011), human trafficking is defined as the modern slavery or the trade of humans mostly for the purpose of sexual exploitation and forced labor, via different improper ways including force, fraud and deception. The United States' Trafficking Victim Protection Act of 2000 (TVPA 2000) tvp (2000) was the first U.S. legislation passed against human trafficking. Human trafficking has ever since received increased national and societal concern HTr (2015) but still demands persistent fight against from all over the globe. No country is immune and the problem is rapidly growing with little to no law enforcement addressing the issue. This problem is amongst the challenging ones facing law enforcement as it is difficult to identify victims and counter traffickers.

Before the advent of the Internet, human traffickers were under risks of being arrested by law enforcement while advertising their victims on streets Desplaces (2012). However, move to the Internet has made it easier and less dangerous for sex sellers Nicholas D. (2012) as they no longer needed to advertise on the streets. There are now a plethora of websites that host and provide sexual services under categories of escort, adult entertainment, massage services, etc., which help sex sellers and buyers maintain their anonymity. Although some services such as the Craiglist's adult section and myredbook.com were shut down recently, there are still many websites such as the Backpage.com that provide such services and many new are frequently created. Traffickers even use dating and social networking websites such as the Twitter, Facebook, Instagram and Tinder to reach out to sex buyers and their followers. Although the Internet has presented new trafficking related challenges for law enforcement, it has also provided readily and publicly available rich source of information which could be gleaned from online sex advertisements for fighting this crime Kennedy (2012). However, the problem is we lack the ground truth and obtaining the labels through hand-labeling is indeed tedious and expensive even for a small subset of data– this is the point where the semi-supervised setting comes in handy.

Despite considerable attention which has been devoted to studying supervised, unsupervised and semi-supervised learning settings via different applications Mitchell (2006); Beigi *et al.* (2016a); Backstrom and Leskovec (2011); Alvari *et al.* (2016a); Beigi *et al.* (2016b); Mitchell *et al.* (1997); Beigi *et al.* (2014), semi-supervised learning, i.e., learning from labeled and unlabeled examples, is still one of the most interesting yet challenging problems in the machine learning community Belkin *et al.* (2006). The idea is simple though– we shall have an approach that makes a better use of unlabeled data to boost performance. This is pretty close to the most natural learning that occurs in the world. For the most part, we as humans are exposed only to a small number of labeled instances; yet we successfully generalize well by effective utilization of a large amount of unlabeled data. This motivates us to use unlabeled samples to improve recognition performance while developing classifiers.

In this chapter, we present results from our works on detecting human trafficking Alvari *et al.* (2016b, 2017). We use the data crawled from the adult entertainment section of the website Backpage.com and extend the existing Laplacian SVM framework Belkin *et al.* (2006) to detect escort advertisements of high interest to law enforcement. Here, we merely focus on the online advertisements in the form of *content* posted by users, and leave investigating other forms of data to the subsequent chapters. We thus highlight several contributions of the current research as follows.

1. Based on the literature, we created different groups of features that capture the characteristics of potential human trafficking activities. The less likely human

trafficking related posts were then filtered out using these features. We also conducted a feature importance analysis to demonstrate how these features contribute to the proposed learner.

2. We extended the Laplacian SVM Belkin *et al.* (2006) and proposed the semi-supervised support vector machine learning algorithm, $S^3VM - R$. In particular, we incorporated additional information of our feature space as a regularization term into the standard optimization formulation with regard to the Laplacian SVM. We also used geometry of the underlying data as an intrinsic regularization term in Laplacian SVM.

3. We trained our model on both of the *labeled* and *unlabeled* data and sent back the identified human trafficking related advertisements to an expert from law enforcement for further verification. We then validated our approach on a small subset of the unlabeled data (i.e. *unseen* data) with further verification of the expert.

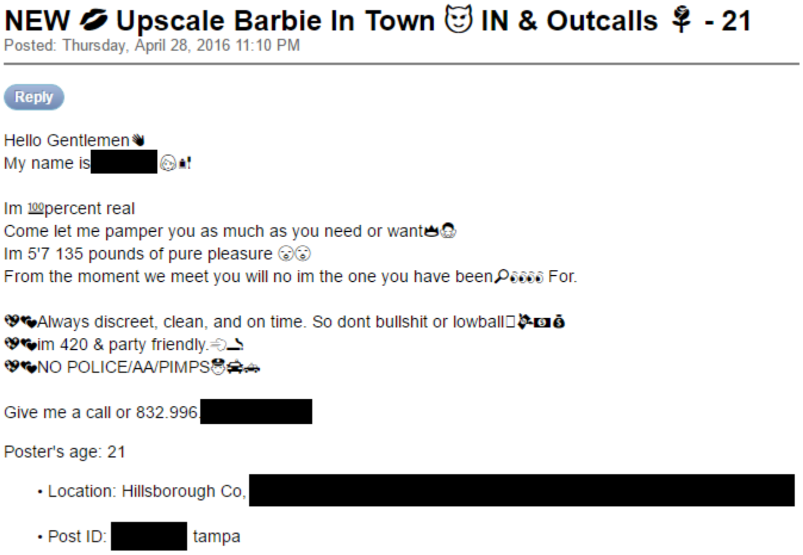
4. We performed comparisons between our approach and several semi-supervised and supervised baselines on both of the labeled and unseen data (so-called *blind* evaluation).

5. We demonstrated the effect of varying different hyperparameters used in our learner on its performance.

3.1 Data Preparation

We collected about 20K publicly available listings from the U.S. posted on Backpage.com in March, 2016. Each post includes a title, description, time stamp, poster's age, poster's ID, location, image, and sometimes video and audio. The description usually lists the attributes of the individual(s) and contact phone numbers. In this work, we only focus on the textual component of the data. This free-text data required significant cleaning due to a variety of issues common to textual analytics (i.e. misspellings, format of phone numbers, etc.). We also acknowledge that the information in data could be intentionally inaccurate, such as poster's name, age and even physical appearance (e.g. bra cup size, weight). Figure 3.1 shows an actual post from Backpage.com. To illustrate geographic diversity of the listings, we use the Tableau [1] software to visualize choropleth map of phone frequency with respect to the different states in Figure 3.2, wherein darker colors mean higher frequencies.

Figure 3.1: A Real Post from Backpage.com. To Ensure Anonymity, the Personal Information has been Intentionally Obfuscated.



NEW 💋 Upscale Barbie In Town 😼 IN & Outcalls 🌹 - 21

Posted: Thursday, April 28, 2016 11:10 PM

Hello Gentlemen👋
My name is[redacted]😊💁!

Im 💯percent real
Come let me pamper you as much as you need or want👑😘
Im 5'7 135 pounds of pure pleasure 😜😛
From the moment we meet you will no im the one you have been🔍👀👀 For.

💖💘Always discreet, clean, and on time. So dont bullshit or lowball□💸💵💰
💖💘im 420 & party friendly.💨🚬
💖💘NO POLICE/AA/PIMPS👮🚔🚗

Give me a call or 832.996[redacted]

Poster's age: 21

- Location: Hillsborough Co, [redacted]
- Post ID: [redacted] tampa

Next, we will explain most important characteristics of potential human trafficking advertisements which are captured by our feature groups.

[1] https://www.tableau.com/

Figure 3.2: Choropleth Map of Phone Frequency w.r.t the Different States. Darker Colors Show Higher Frequencies.

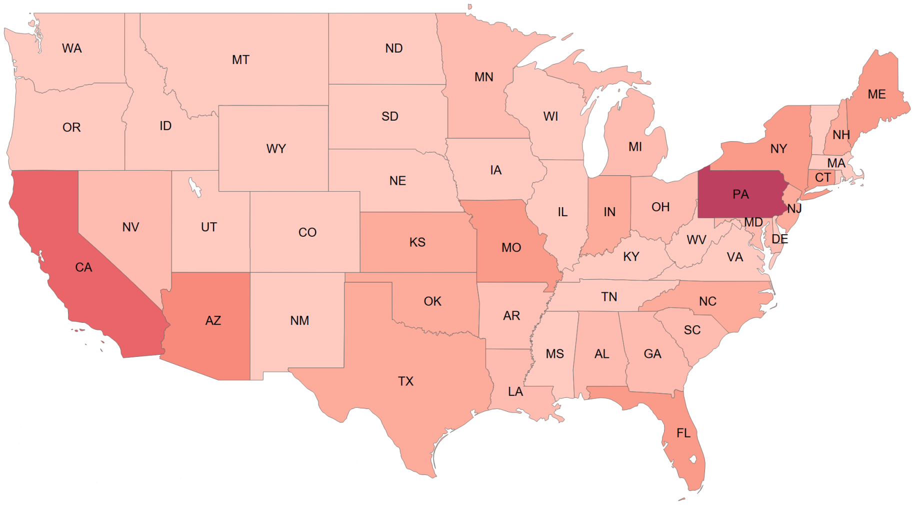
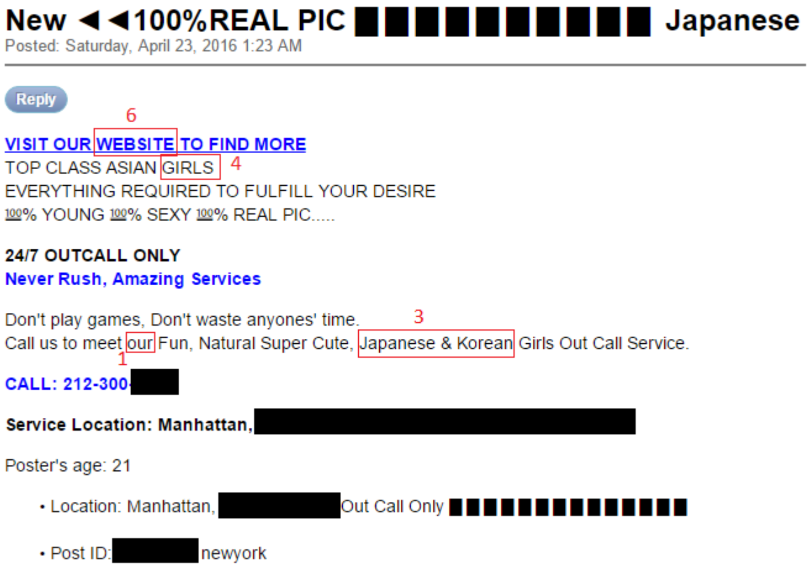


Figure 3.3: An Evidence of Human Trafficking. The Boxes and Numbers in Red, Indicate the Features and their Corresponding Group Numbers (See Also Table 3.1).



New ◄◄100%REAL PIC ██████████ Japanese

Posted: Saturday, April 23, 2016 1:23 AM

VISIT OUR WEBSITE TO FIND MORE
TOP CLASS ASIAN GIRLS
EVERYTHING REQUIRED TO FULFILL YOUR DESIRE
100% YOUNG 100% SEXY 100% REAL PIC.....

24/7 OUTCALL ONLY
Never Rush, Amazing Services

Don't play games, Don't waste anyones' time.
Call us to meet our Fun, Natural Super Cute, Japanese & Korean Girls Out Call Service.

CALL: 212-300[redacted]

Service Location: Manhattan,[redacted]

Poster's age: 21

- Location: Manhattan, [redacted]Out Call Only ██████████████
- Post ID:[redacted]newyork

3.1.1 Feature Engineering

Though many advertisements on Backpage.com are posted by posters selling their own services without coercion and intervention of traffickers, some do exhibit many common trafficking triggers. For example, in contrast to Figure 3.1, Figure 3.3 shows an advertisement that could be an evidence of human trafficking. This advertisement indicates several potential properties of human trafficking, including advertising for multiple escorts with the first individual coming from Asia and very young. In what follows, such common properties of human trafficking related advertisements are discussed in more detail.

Inspired by the literature, we define and extract 6 groups of features from advertisements (see Table 3.1). These features could be amongst the strong indicators of human trafficking. Let us now briefly describe each group of features used in our work. Note each feature listed here is ultimately treated as a *binary* variable.

Advertisement Language Pattern

The first group consists of different language related features. For the first and second features, we identify posts which have third person language (more likely to be written by someone other than the escort) and posts which contain first person plural pronouns such as 'we' and 'our' (more likely to be an organization) Kennedy (2012).

To ensure their anonymity, traffickers would deploy techniques to generate diverse information and hence make their posts look more complicated. They usually do this to avoid being identified by either human analysts or automated programs. Thus, to obtain the third feature we take an approach from complexity theory, namely *Kolmogorov complexity*, which is defined as length of shortest program to reproduce a string of characters on a universal machine such as the Turing Machine Li and

Vitányi (2008). Since the Kolmogorov complexity is not computable, we approximate the complexity of an advertisement content by first removing stop words and then computing entropy of the content Li and Vitányi (2008). To illustrate this, let X denote the content and x_i be a given word in the content. We use the following equation Shannon (2001) to calculate the entropy of the content and thus approximate the Kolmogorov complexity of X:

$$K(X) \approx -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{3.1}$$

We expect higher values of the entropy correspond to human trafficking. Finally, we discretize the result by using the threshold of 4 which was found empirically in our experiments.

For the next features, we use word-level n-grams to find common language patterns of advertisements. This particular choice is because of the fact that character-level n-grams have already shown to be useful in detecting unwanted content for spam detection Kanaris *et al.* (2006). We set $n = 4$ and use the range of (4,4) to compute normalized n-grams (using TF-IDF) of each advertisement content. We ultimately create a matrix whose rows and columns correspond to the advertisements contents and their associated 4-grams, respectively. We rank all elements of this matrix in a descending order and pick the top 3 ones. Finally for each advertisement content, 3 elements with the column numbers associated with the top elements are chosen. This way, 3 more features will be added to our feature set. Overall, we have 6 features related to the language of the advertisement.

Words and Phrases of Interest

Despite the fact that advertisements on Backpage.com do not directly mention sex with children, customers who prefer children know to look for words and phrases such as "*sweet*, *candy*, *fresh*, *new in town*, *new to the game*" Hetter (2012); Lloyd (2012); Dickinson Goodman and Holmes (2011). We thus investigate within the posts to see if they contain such words as they could be highly related with human trafficking in general.

Countries of Interest

We identify if the individual being escorted is coming from other countries such as those in Southeast Asia (especially from China, Vietnam, Korea and Thailand, as we observed in our data) HTr (2015).

Multiple Victims Advertised

Some advertisements advertise for multiple women at the same time. We consider the presence of more than one victim as a potential evidence of organized human trafficking Kennedy (2012).

Victim Weight

We take into account the weight of the individual being escorted as a feature (if it is available). This information is particularly useful assuming that for the most part, lower body weights (under 115 lbs) correlate with smaller and underage girls tvp (2000); wei (2017) and thereby human trafficking.

Reference to Website or Spa Massage Therapy

The presence of a link in the advertisement either referencing to an outside website (especially infamous ones) or spa massage therapy could be an indicator of more elaborate organization Kennedy (2012). In particular, in case of spa therapy, we observed many advertisements interrelated with advertising for young Asian girls and their erotic massage abilities. Therefore, the last group of features has two binary features for presence of any website and spa.

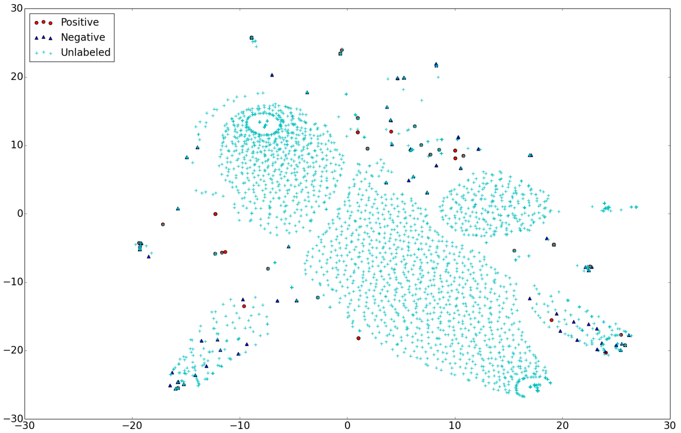
Finally, in order to extract all of the above features, we first clean the original data and conduct preprocessing. By applying these features, we draw a random sample of 3,543 instances out of our dataset for further analysis to see if they are evidences of human trafficking– this is described in the next section.

3.1.2 Unsupervised Filtering

Having detailed our feature set, we now construct a feature vector for each instance by creating a vector of 12 binary features that correspond to the important characteristics of human trafficking. Hereafter, we refer to this feature space, as our *first* feature space and denote it with $\mathcal{F}_1$. As mentioned earlier, we draw 3,543 instances from our raw data by filtering out those that do not posses any of the binary features. We will refer to this as our *filtered* dataset. For the sake of visualization, a 2-D projection (using the t-SNE transformation van der Maaten and Hinton (2008)) of the filtered dataset is depicted in Figure 3.4. The purpose of this figure is to demonstrate how hard it is for basic clustering techniques such as the K-means, to correctly assign labels to unlabeled instances using only few existing labeled ones.

Now, we shall define our *second* feature space, namely $\mathcal{F}_2$, which will be used to compute geometry of the underlying data. Note that our proposed framework will utilize both of the feature spaces in the form of regularization terms, to detect advertisements of high interest to law enforcement. After conducting standard preprocessing techniques on the filtered dataset, we build $\mathcal{F}_2$ by transforming the filtered data into a 3,543×3,543 matrix of TF-IDF similarity features. Each entry in this matrix simply shows the similarity between a pair of advertisements in our filtered dataset.

Figure 3.4: 2-D Projection of the Entire Set of the Filtered Data.



Note that since we lack the ground truth, we would rely on a human analyst (expert) for labeling the listings as either 'of interest' or 'of not interest' to law enforcement. In the next section, we select a smaller yet finer grain subset of this data to be sent to the expert. This alleviates the burden of the tedious work of hand-labeling.

3.1.3 Expert Assisted Labeling

We first obtain a sample of 200 listings from the filtered dataset. This set of listings was labeled by our expert from law enforcement who is specialized in this type of crime. From this subset, the law enforcement professional identified 70 instances to be of interest to law enforcement and the rest to be not human trafficking related. However, we are still left with a large amount of the unlabeled examples (3,343 instances) in our dataset. The ratio of the labeled to unlabeled instances in our dataset is very small (about 0.06). The statistics of our dataset is summarized in Table 6.6.

3.2 A Non-Parametric Learning Approach

We use the Python package *scikit-learn* for training semi-supervised learner on the filtered dataset. There are two label propagation semi-supervised (non-parametric) based models in this package, namely, LabelPropagation and LabelSpreading Bengio *et al.* (2006). These models rely on the geometry of the data induced by both labeled and unlabeled instances as opposed to the supervised models which only use the labeled data Bengio *et al.* (2006). This geometry is usually represented by a graph $G = (V, E)$, with the nodes V represent the training data and edges E represent the similarity between them Bengio *et al.* (2006) in the form of weight matrix $\mathbf{W}$. Given the graph G, a basic approach for semi-supervised learning is through propagating labels on the graph Bengio *et al.* (2006). Due to the higher performance achieved, we chose to use LabelSpreading model.

3.2.1 Experiments

We conducted experiment with the two built-in kernels radial basis function (RBF) and K-nearest neighbor (KNN) in label propagation models and report the results in Table 3.3. Note that we only reported the precision when 119 negative samples (labeled by either of the experts) were used in the learning process. We did so because of the reasonable number of the positive labels assigned by either of the kernels in presence of these negative instances (our experts had limited time to validate the labels of the data).

As we see from this table, out of 849 unlabeled data, our learner with RBF and KNN kernels assigned positive labels to the 145 and 188 instances, respectively. Next, we pass the identified positive labels to the experts for further verification. Our approach with RBF and KNN correctly identified 134 and 170 labels out of 145 and

188 positive instances and achieved precision of 92.41% and 90.42%, respectively. We further demonstrate the word clouds for the positive instances assigned by RBF and KNN, in Figure 3.5 and Figure 3.6, respectively.

Figure 3.5: Word Cloud for the Positive Instances Assigned by RBF.

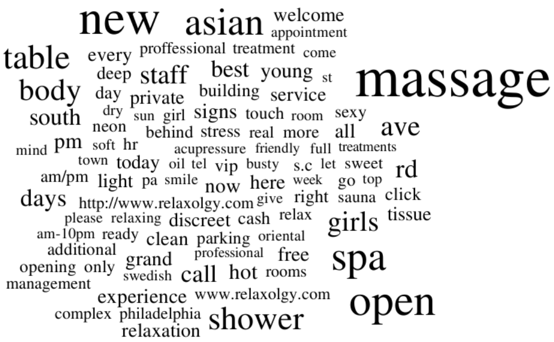
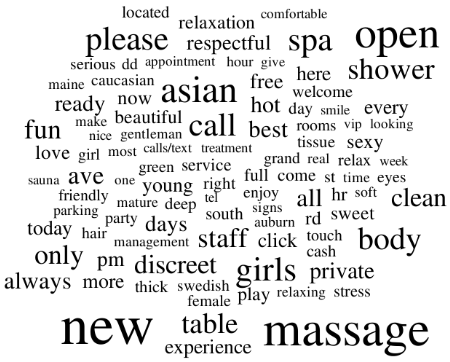


Figure 3.6: Word Cloud for the Positive Instances Assigned by KNN.



3.3 Semi-Supervised Learning Framework

Our framework is an extension to the existing Laplacian SVM Belkin *et al.* (2006). In particular, we incorporated another regularization term into the standard Laplacian SVM to leverage the additional information of our first feature space and then solved the associated optimization problem. Consequently, similar notation is adopted throughout the following section. Furthermore, we shall once again note that our current research does not utilize any off-the-shelf graph based semi-supervised leaner in contrast to our previous research Alvari *et al.* (2016b).

3.3.1 Technical Preliminaries

We assume a set of l labeled pairs $\{(x_i, y_i)\}_{i=1}^{l}$ and an unlabeled set of u instances $\{x_{l+i}\}_{i=1}^{u}$, where $x_i \in \mathbb{R}^n$ and $y_i \in \{+1, -1\}$. Recall for the standard soft-margin support vector machine, the following optimization problem is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||_k^2 + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) \tag{3.2}$$

In the above equation, $f_\theta(\cdot)$ is a decision function of the form $f_\theta(\cdot) = w.\mathbf{\Phi}(\cdot) + b$ where $\theta = (w, b)$ are the parameters of the model, and $\mathbf{\Phi}(\cdot)$ is the feature map which is usually implemented using the kernel trick Cortes and Vapnik (1995). Also, the function $H_1(\cdot) = \max(0, 1 - \cdot)$ is the Hinge Loss function.

The classical Representer theorem Belkin *et al.* (2005) suggests that solution to the optimization problem exists in a Hilbert space $\mathcal{H}_k$ and is of the following form:

$$f_\theta^*(x) = \sum_{i=1}^{l} \alpha_i^* \mathbf{K}(x, x_i) \tag{3.3}$$

where $\mathbf{K}$ is the $l \times l$ Gram matrix over labeled samples. Equivalently, the above problem can be written as:

$$\min_{w,b,\epsilon} \frac{1}{2}||w||_2^2 + C_l \sum_{i=1}^{l} \epsilon_i \tag{3.4}$$

$$\begin{aligned} s.t. \quad & y_i(w.\mathbf{\Phi}(x_i) + b) \geq 1 - \epsilon_i, \ i = 1, ..., l \\ & \epsilon_i \geq 0, \ i = 1, ..., l \end{aligned} \tag{3.5}$$

We will use the above optimization equation as our basis to derive the formulations for our proposed semi-supervised learner.

3.3.2 The Proposed Method

The basic assumption behind semi-supervised learning methods is to leverage unlabeled instances in order to restructure hypotheses during the learning process. In this work, exogenous information extracted from both of our feature spaces is further exploited to make a better use of the unlabeled examples. To do so, we first introduce matrix $\mathbf{F}$ in $\mathcal{F}_1$ and over both of the labeled and unlabeled samples with $\mathbf{F}_{ij}$ defined as follows:

$$\mathbf{F}_{ij} = \frac{1}{n_f}(\mathbf{\Phi}(x_i) \cdot \mathbf{\Phi}(x_j)) \tag{3.6}$$

where n_f is the number of features in $\mathcal{F}_1$ (here, $n_f = 12$). We force the instances x_i and x_j in our dataset to have same label if they both possess same features. To account for this, a regularization term is added to the standard equation and the following optimization is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \frac{1}{2} \sum_{i=1}^{l} \mathbf{F}_{ij} ||f_\theta(x_i) - f_\theta(x_j)||_2^2 = \mathbf{f}_\theta^T \mathcal{L}^T \mathbf{f}_\theta \tag{3.7}$$

where $\mathbf{f} = [f(x_1), ..., f(x_{l+u})]^T$ and $\mathcal{L}$ is the Laplacian matrix based on $\mathbf{F}$ given by $\mathcal{L} = \mathbf{D} - \mathbf{F}$, and $\mathbf{D}_{ii} = \sum_{j=1}^{l+u} \mathbf{F}_{ij}$. The intuition here is that any two instances which are composed of same features are more likely to have same labels than others. Next, by solving a similar optimization problem, we are able to capture data geometry in $\mathcal{F}_2$ as $\mathbf{f}_\theta^T \mathcal{L}'^T \mathbf{f}_\theta$ (also referred to as the intrinsic smoothness penalty term Belkin *et al.* (2006)). Here, $\mathcal{L}'$ is the Laplacian of matrix $\mathbf{A}$ associated with the data adjacency graph $\mathbf{G}$ in $\mathcal{F}_2$.

We construct $\mathbf{G}$ with $(l+u)$ nodes in $\mathcal{F}_2$, and by adding an edge between each pair of nodes $\langle i,j\rangle$, if the edge weight W_{ij} exceeds a given threshold. For computing the edge weights, we use the heat kernel Grigor'yan (2006) as a function of the Euclidean distance between two samples in $\mathcal{F}_2$, hence we set $W_{ij} = \exp^{-||x_i - x_j||^2/4t}$.

Following the notations used in Belkin *et al.* (2006) and by including our regularization term as well as the intrinsic smoothness penalty term, we would extend the standard equation by solving the following optimization:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||_k^2 + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) + C_r \mathbf{f}_\theta^T \boldsymbol{\mathcal{L}} \mathbf{f}_\theta + C_s \mathbf{f}_\theta^T \boldsymbol{\mathcal{L}}' \mathbf{f}_\theta \tag{3.8}$$

Note one typical value for the smoothness penalty coefficient C_s is $\frac{\gamma_I}{(l+u)^2}$, where $\frac{1}{(l+u)^2}$ is a natural scale factor for empirical estimate of the Laplace operator and γ_I is a regularization term Belkin *et al.* (2006). Again, solution in $\mathcal{H}_k$ would be in the following form:

$$f_\theta^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(x, x_i) \tag{3.9}$$

Here $\mathbf{K}$ is the $(l+u) \times (l+u)$ Gram matrix over all samples. The equation 5.6 could be then written as follows:

$$\min_{\alpha,b,\epsilon} \frac{1}{2}\alpha^T \mathbf{K}\alpha + C_l \sum_{i=1}^{l} \epsilon_i + \frac{C_r}{2}\alpha^T \mathbf{K}\boldsymbol{\mathcal{L}}\mathbf{K}\alpha + \frac{\gamma_I}{2(l+u)^2}\alpha^T \mathbf{K}\boldsymbol{\mathcal{L}}'\mathbf{K}\alpha \tag{3.10}$$

$$\begin{aligned} s.t. \quad & y_i(\sum_{j=1}^{l+u} \alpha_j \mathbf{K}(x_i, x_j) + b) \geq 1 - \epsilon_i,\ i = 1, ..., l \\ & \epsilon_i \geq 0,\ i = 1, ..., l \end{aligned} \tag{3.11}$$

With introduction of the Lagrangian multipliers β and γ, we write the Lagrangian function of the above equation as follows:

$$L(\alpha, \epsilon, b, \beta, \gamma) = \frac{1}{2}\alpha^T \mathbf{K}(I + C_r \mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')\alpha + C_l \sum_{i=1}^{l} \epsilon_i$$
$$- \sum_{i=1}^{l} \beta_i (y_i (\sum_{j=1}^{l+u} \alpha_j \mathbf{K}(x_i, x_j) + b) - 1 + \epsilon_i) - \sum_{i=1}^{l} \gamma_i \epsilon_i \tag{3.12}$$

Obtaining the dual representation, requires taking the following steps:

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l} \beta_i y_i = 0 \tag{3.13}$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \rightarrow C_l - \beta_i - \gamma_i = 0 \rightarrow 0 \leq \beta_i \leq C_l \tag{3.14}$$

With the above equations, we formulate the reduced Lagrangian as a function of only α and β as follows:

$$L^R(\alpha, \beta) = \frac{1}{2}\alpha^T \mathbf{K}(I + C_r \mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')\alpha$$
$$- \sum_{i=1}^{l} \beta_i (y_i (\sum_{j=1}^{l+u} \alpha_j \mathbf{K}(x_i, x_j) + b) - 1 + \epsilon_i) \tag{3.15}$$

This equation is further simplified as follows:

$$L^R(\alpha, \beta) = \frac{1}{2}\alpha^T \mathbf{K}(I + C_r \mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')\alpha$$
$$-\alpha^T \mathbf{K}\mathbf{J}^T \mathbf{Y} \beta + \sum_{i=1}^{l} \beta_i \tag{3.16}$$

In the above equation, $\mathbf{J} = [\mathbf{I}\ \mathbf{0}]$ is a $l \times (l+u)$ matrix, $\mathbf{I}$ is the $l \times l$ identity matrix and $\mathbf{Y}$ is a diagonal matrix consisting of the labels of the labeled examples.

In the followings, we first take the derivative of L^R with respect to α and then set $\frac{\partial L^R(\alpha,\beta)}{\partial \alpha} = 0$:

$$\mathbf{K}(I + C_r\mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')\alpha - \mathbf{K}\mathbf{J}^T\mathbf{Y}\beta = 0 \tag{3.17}$$

Accordingly, we obtain α^* by solving the following equation:

$$\alpha^* = (I + C_r\mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')^{-1}\mathbf{J}^T\mathbf{Y}\beta^* \tag{3.18}$$

Next, we obtain the dual problem in the form of a quadratic programming problem by substituting α back in the reduced Lagrangian function:

$$\beta^* = \mathrm{argmax}_{\beta \in \mathbb{R}^l} \ -\frac{1}{2}\beta^T\mathbf{Q}\beta + \sum_{i=1}^{l}\beta_i \tag{3.19}$$

$$\begin{aligned} s.t. \quad & \sum_{i=1}^{l}\beta_i y_i = 0 \\ & 0 \le \beta_i \le C_l \end{aligned} \tag{3.20}$$

where $\beta = [\beta_1, ..., \beta_l]^T \in \mathbb{R}^l$ are the Lagrangian multipliers and $\mathbf{Q}$ is obtained as follows:

$$\mathbf{Q} = \mathbf{Y}\mathbf{J}\mathbf{K}(I + (C_r\mathcal{L} + \frac{\gamma_I}{(l+u)^2}\mathcal{L}')\mathbf{K})^{-1}\mathbf{J}^T\mathbf{Y} \tag{3.21}$$

We summarize the proposed semi-supervised framework in Algorithm 1. Our optimization problem is very similar to the standard optimization problem solved for SVMs, hence we use a standard optimizer for SVMs to solve our problem.

Algorithm 1 The Proposed Semi-Supervised Framework

```
Input: $\{(x_i, y_i)\}_{i=1}^{l}$, $\{x_{l+i}\}_{i=1}^{u}$, $\mathcal{F}_1$, $\mathcal{F}_2$, $C_l$, $C_r$, $C_s$.
Output: Estimated function $f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$
Construct matrix $\mathbf{F}$ based on the features in $\mathcal{F}_1$
Compute the corresponding Laplacian matrix $\mathcal{L}$.
Construct $\mathbf{A}$ according to the features in $\mathcal{F}_2$.
Compute the graph Laplacian matrix $\mathcal{L}'$.
Construct the gram matrix over all examples using $\mathbf{K}_{ij} = k(x_i, x_j)$ where $k$ is a kernel function.
Compute $\alpha^*$ and $\beta^*$ using Eq. 3.18 and Eq. 5.7 and a standard QP solvers.
Compute function $f_\theta^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(x, x_i)$
```

3.3.3 Experimental Study

In this section, we provide a comprehensive analysis of the proposed framework by designing a series of experiments on the filtered dataset. First, we explain several approaches used in this study. Next, various results are discussed: (1) comparisons on the *labeled* data were made between our method and other approaches, (2) experiments were performed on a fraction of the unlabeled data (i.e., unseen data), and the results were further verified by our expert to see what fraction is of interest to law enforcement, (3) *blind* evaluation was conducted to examine other approaches on the unseen data, and finally, (4) experiments were designed to analyze effect of varying different hyperparameters on our method as well as impact of different groups of features in $\mathcal{F}_1$ on our approach. We present results for the following methods:

- **Semi-Supervised**: $S^3VM{-}R$, Laplacian support vector machines Belkin *et al.* (2006), graph inference based label spreading approach Zhou *et al.* (2004b) with radial basis function (RBF) and K-nearest neighbors (KNN) kernels, and co-training learner Blum and Mitchell (1998) with two support vector machines classifiers (SVM).

- **Supervised**: SVM, KNN, Gaussian naïve Bayes, logistic regression, adaboost and random forest.

For the sake of fair comparison, all algorithms were implemented and run in Python. More specifically, the Python package CVXOPT [2] was used to implement $S^3VM - R$ and Laplacian support vector machines, and all other approaches were implemented with the help of the Scikit-learn [3] package in Python. Note for those methods that require special tuning of parameters, we performed grid search to choose the best set of parameters. Before going any further, we first define main parameters used in each method and then demonstrate their best values picked by our grid search. The discussion on the effect of varying the hyperparameters on our learner is provided in the section 3.3.3.

- $S^3VM - R$: we set the penalty parameter as $C_l = 0.6$ and the regularization parameters $C_r = 0.2$ and $C_s = 0.2$. Linear kernel was used in our approach.

- *Laplacian SVM*: we used linear kernel and set the parameters $C_l = 0.6$ and $C_s = 0.6$.

- *LabelSpreading (RBF)*: RBF Kernel was used and γ was set to the default value of 20.

- *LabelSpreading (KNN)*: KNN kernel was used and the number of neighbors was set to 5.

[2] http://cvxopt.org/

[3] http://scikit-learn.org/stable/

- *Co-training (SVM)*: we followed the algorithm introduced in Blum and Mitchell (1998) and used two SVM as our classifiers. For both SVMs we set the tolerance for stopping criteria to 0.001 and the penalty parameter $C = 1$.
- *SVM*: tolerance for stopping criteria was set to the default value of 0.001. Penalty parameter C was set to 1 and linear kernel was used.
- *KNN*: number of neighbors was set to 5.
- *Gaussian NB*: there were no specific parameter to tune.
- *Logistic regression*: we used the '$l2$' penalty. We also set the parameter $C = 1$ (the inverse of regularization strength) and tolerance for stopping criteria to 0.01.
- *Adaboost*: number of estimators was set to 200 and we also set the learning rate to 0.01.
- *Random forest*: we used 200 estimators and the 'entropy' criterion was used.

Classification Results

Here, we first evaluate the entire set of approaches on a small portion of the data for which we already know the labels, i.e., the *labeled* examples. We note that expert-generated judgmental labeling might be error-prone, though it is served as a surrogate to the ground truth problem.

We used 10-fold cross-validation on the labeled data in the following way. We first divided the set of the labeled samples into 10 different sets of approximately equal size. Each time we held one set out for validation (by removing their labels and adding them to the unlabeled samples) and used the remaining along with the unlabeled samples for the training–this was performed for all approaches for the sake of fair comparison. Finally, we reported the average of 10 different runs, using different combinations of the feature spaces and various evaluation metrics, including the area under curve (AUC), accuracy, precision, recall and F1-score. In table 3.4, we reported the average AUC and accuracy for each method and each feature space. On the other hand, for precision, recall and F1-score, we reported separate results for each feature space, in tables 3.5-3.7, respectively. Note, each of these tables includes separate scores for the positive and negative classes. In general, we observed the followings:

- Overall, our approach achieved highest performance on $\mathcal{F}_1$ (tables 3.4 and 3.5) and $\{\mathcal{F}_1, \mathcal{F}_2\}$ (table 3.7), in terms of all metrics. However it did not perform well using solely $\mathcal{F}_2$ (table 3.6), i.e. when $C_r = 0$. This clearly demonstrates the importance of using C_r over C_s.

- When the feature space used is $\mathcal{F}_2$, Co-training (SVM) is the best method. Next best methods are supervised learners KNN and Gaussian NB. Three remarks can be made here. First, our approach could not always defeat supervised learners as it is seen from tables 3.4 and 3.6. This is not surprising and in fact lies at the inherent difference between semi-supervised and supervised methods– unlabeled examples could make the trained model susceptible to error propagation and thus wrong estimation. Second, as it is seen in tables 3.5-3.7, achieving very high recall on the negative examples and low score on the positive ones shall not be treated as a potent property, otherwise a trivial classifier which always assigns negative labels to all samples would be the best learner. Third, using C_r always improves the performance over C_s. One point that needs to be clarified is, our ultimate goal is not to achieve high performance on the labeled data, but rather to detect the suspicious (unlabeled) advertisements which could be human trafficking related– this will be explained in more details in 3.3.3.

- Compared to the other semi-supervised approaches, our approach either achieved higher or comparable AUC scores. The reason we performed exactly the same as the Laplacian SVM, is because by setting $C_r = 0$, the two approaches are inherently the same.

- For the Laplacian SVM to be able to run on $\mathcal{F}_1$, the Laplacian $\mathcal{L}'$ has to be constructed using $\mathcal{F}_1$ while inherently is supposed to be made using $\mathcal{F}_2$. This is because C_r is essentially associated with $\mathcal{F}_1$, and C_s corresponds to $\mathcal{L}'$ and correspondingly $\mathcal{F}_2$. The same holds for $\{\mathcal{F}_1, \mathcal{F}_2\}$, where we need to construct a new feature space by concatenating $\mathcal{F}_1$ and $\mathcal{F}_2$ as the Laplacian SVM does not inherently use $\mathcal{F}_1$ at all. The new feature space is then used to construct the Laplacian $\mathcal{L}'$.

- Since our approach inherently incorporates both of the Laplacian matrices corresponding to the two feature spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, all other baselines were also run using the concatenation of these two feature spaces for the sake of fair comparison. Unlike our approach which used the wise combination of $\mathcal{F}_1$ and $\mathcal{F}_2$, other methods do not gain high AUC by simply combining the feature spaces.

Blind Evaluation

For the next set of experiments, we first run our method on the *entire* filtered dataset and without cross-validation. Recall from the previous sections that this is to make a better use of the unlabeled examples. Then the following *control* experiment was conducted. Our learner was tested on the whole set of the unlabeled examples. Out of 3,343 instances, our approach identified two sets of positive and negative instances. The positive set contained 394 advertisements which were likely to be of interest to law enforcement, whereas the negative set included the remaining 2,962 unlabeled advertisements of probably less interest to law enforcement. Next, to precisely determine the correctly identified fractions of these two sets, we randomly picked two subsets (control groups) of 100 examples from each set for further validation by our expert.

We passed these two control groups to our expert for further verification. The expert-validated results demonstrated that all of the examples in the positive group were of interest to law enforcement, while only two examples from the negative group were not correctly classified as of not being of any interest to law enforcement. Thus, both results support the effectiveness of our framework in identifying highly human trafficking advertisements. Using the same two control groups and AUC metric, we now perform so-called blind evaluation (see table 3.8) of other baselines. Note, we call this blind since actual labels are not provided and the expert-generated labels might convey uninformative information. In general, supervised methods failed to achieve good results in the blind evaluation compared to most of the semi-supervised methods.

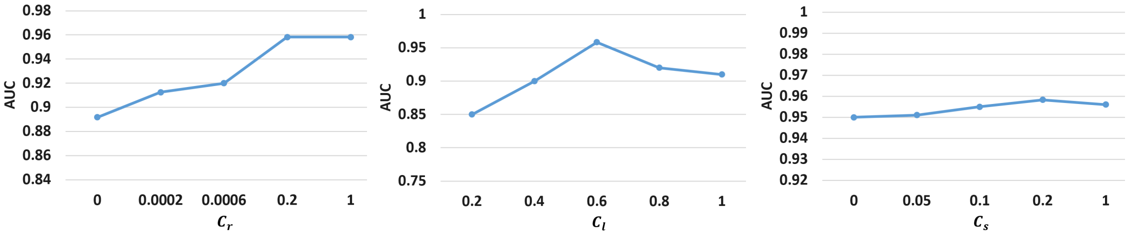
Hyperparameter Sensitivity

Here, we discuss how altering the hyperparameters C_l, C_r and C_s may affect the performance of $S^3VM - R$. We start off by fixing the value of C_l to 0.6, which was empirically found to work well in our experiments. Also, recall from the previous sections that one typical choice for C_s is $\frac{\gamma_I}{(l+u)^2}$ Belkin *et al.* (2006). Here, we set $C_s = 0.2$ and varied the values of C_r as $\{0, 0.0002, 0.0006, 0.2, 1.0\}$ and plotted the results in Figure 3.7. We used the same 10-fold cross-validation setting from the previous section.

We made the following observation. With the slight increase of C_r, the performance of our approach increased, peaked and then stabilized, i.e., further increase of C_r did not change the performance. This suggests significance of deploying the additional information from our first feature space $\mathcal{F}_1$, over $\mathcal{F}_2$ and its corresponding smoothness penalty parameter C_s which is used by $S^3VM - R$ and the standard Laplacian SVM.

Next, to see the impact of C_l on the performance, we set $C_r = 0.2$ and varied C_l as $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. The results are depicted in Figure 3.7. We note that setting $C_l = 0$ is meaningless and thus we do not have any performance corresponding to that– otherwise each β_i in Eq. 5.7 would be zero. In general, the performance was not particular sensitive to this parameter– varying by 0.2 for values of 0.4 and greater.

Figure 3.7: Effect of Varying Different Parameters on the Performance.



Finally, having fixed $C_l = 0.6$ and $C_r = 0.2$, we also tried other values for C_s including $\sum_{i,j=1}^{l+u} W_{ij}$ suggested by Belkin *et al.* (2006) and depicted the results in Figure 3.7. The results suggest that our approach is less sensitive to this parameter compared to C_r and C_l.

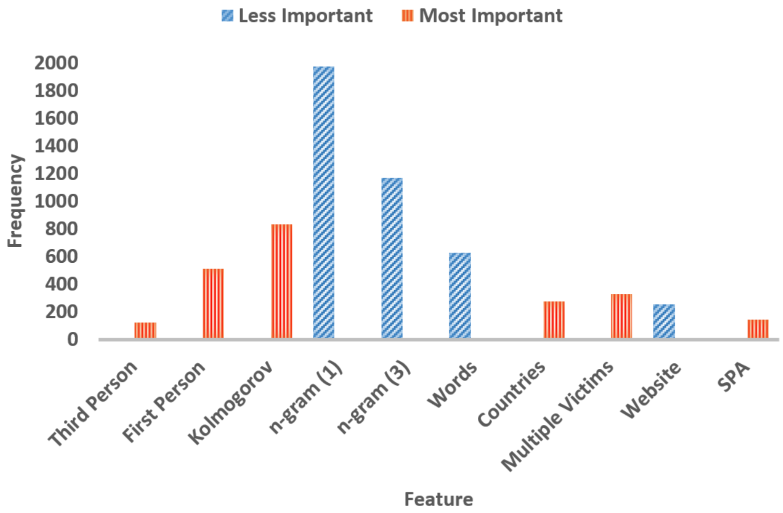
Significance of Features

To examine how much discriminative our feature groups in $\mathcal{F}_1$ are, we further conducted an analysis using the labeled examples and the standard feature selection measure χ^2 to find the top features– only half of the features with scores greater than a given threshold (0.5) were selected (see table 3.9 for the complete set of features and their corresponding χ^2 scores).

From this list, we noticed that 'countries of interest' and 'reference to spa massage therapy' were the most discriminative feature groups, while 'advertisement language pattern' group (with 3 important features) appeared to be the most dominant feature group.

Figure 3.8 compares the top features against the less important subset of the features (denoted by $\overline{\mathcal{F}_1^*}$) in the filtered dataset, in terms of frequency values. Note for clarity, we have removed from this figure, the features with frequency less than 20. According to this figure, our most discriminative features are not necessarily those that appear more often.

Figure 3.8: Frequency of Each Feature in $\mathcal{F}_1$ in the Filtered Dataset. Features are Grouped into the Two Groups, Most Important ($\mathcal{F}_1^*$) and Less Important Features ($\overline{\mathcal{F}_1^*}$), According to χ^2.



To further investigate the importance of each of the top features, we performed classification using the labeled examples and the previous setting, on basis of these two subsets of the features and their combination, i.e., $\mathcal{F}_1^*$, $\overline{\mathcal{F}_1^*}$ and $\mathcal{F}_1$. The classification results are shown in Table 3.10. We made the following observations:

- Considering only the feature space $\mathcal{F}_1$, our approach achieved higher performance compared to all other baselines by either using the whole feature space or the most discriminative features $\mathcal{F}_1^*$.
- Deploying only the features from $\mathcal{F}_1^*$, we were able to achieve comparable results as if we used the whole feature space $\mathcal{F}_1$.

3.4 Conclusion

Readily available online data from escort advertisements could be leveraged in favor of fight against human trafficking. In this study, having focused on textual information from the available data crawled from Backpage.com, we identified if an escort advertisement can be reflective of human trafficking activities. In particular, we first proposed an unsupervised filtering approach to filter out the data which are more likely involved in human trafficking. We then proposed a non-parametric learner and a semi-supervised framework, and trained them on a small portion of the data which was hand-labeled by a human trafficking expert. We used the trained models to identify labels of unseen data. Results suggest our methods are effective at identifying potential human trafficking related advertisements.

Table 3.1: Different Features and their Corresponding Groups.

No.	Feature Group	Ref.
1	**Advertisement Language Pattern**	Kennedy (2012); Li and Vitányi (2008); Kanaris *et al.* (2006)
	- Third person language	
	- First person plural pronouns	
	- Kolmogorov complexity	
	- n-grams (1)	
	- n-grams (2)	
	- n-grams (3)	
2	**Words and Phrases of Interest**	Hetter (2012); Lloyd (2012); Dickinson Goodman and Holmes (2011)
3	**Countries of Interest**	HTr (2015)
4	**Multiple Victims Advertised**	Kennedy (2012)
5	**Victim Weight**	tvp (2000); wei (2017)
6	**Reference to Website or Spa Massage Therapy**	Kennedy (2012)
	- Reference to a website	
	- Reference to a Spa Massage Therapy	

Table 3.2: Description of the Dataset.

Name	**Value**	
Raw	20,822	
Filtered	3,543	
Unlabeled	3,343	
Labeled	**Positive**	**Negative**
	70	130

Table 3.3: Validated Results on Unlabeled Data for both Kernels.

Name	**Value**			
	Positive	Negative	Positive	Precision
Kernel	(Learner)	(Learner)	(Experts)	(Positive)
RBF (Union)	145	704	134	92.41%
RBF (Intersection)	848	1	-	-
KNN (Union)	188	661	170	90.42%
KNN (Intersection)	849	0	-	-

Table 3.4: AUC and Accuracy Results with 10-Fold Cross-Validation on the Labeled Data. The Best Performance is in Bold.

Learner	**AUC**			**Accuracy**		
	$\mathcal{F}_1$	$\mathcal{F}_2$	$\{\mathcal{F}_1, \mathcal{F}_2\}$	$\mathcal{F}_1$	$\mathcal{F}_2$	$\{\mathcal{F}_1, \mathcal{F}_2\}$
$S^3VM - R$	**0.91**	0.9	**0.96**	**0.91**	0.9	**0.97**
Laplacian SVM	0.9	0.9	0.9	0.91	0.9	0.92
LabelSpreading (RBF)	0.78	0.87	0.84	0.8	0.85	0.86
LabelSpreading (KNN)	0.68	0.80	0.74	0.71	0.8	0.8
Co-Training (SVM)	0.82	**0.94**	0.92	0.85	**0.94**	0.93
SVM	0.82	0.9	0.91	0.85	0.92	0.93
KNN	0.76	0.91	0.81	0.79	0.92	0.84
Gaussian NB	0.78	0.91	0.9	0.82	0.9	0.9
Logistic Regression	0.82	0.89	0.88	0.85	0.92	0.92
AdaBoost	0.82	0.85	0.85	0.85	0.88	0.88
Random Forest	0.81	0.89	0.89	0.83	0.91	0.92

Table 3.5: Precision, Recall and F1-Score for the Positive and Negative Classes using $\mathcal{F}_1$. Experiments were Run using 10-Fold Cross-Validation on the Labeled Data. The Best Performance is in Bold.

Learner	**Precision**		**Recall**		**F1-score**	
	$class_p$	$class_n$	$class_p$	$class_n$	$class_p$	$class_n$
S^3VM-R	**0.91**	**0.92**	**0.91**	**0.93**	**0.91**	**0.92**
Laplacian SVM	0.86	0.89	0.88	0.9	0.87	0.89
LabelSpreading (RBF)	0.76	0.78	0.77	0.73	0.76	0.75
LabelSpreading (KNN)	0.65	0.7	0.71	0.68	0.68	0.69
Co-Training (SVM)	0.81	0.84	0.71	0.92	0.76	0.88
SVM	0.86	0.83	0.68	0.91	0.76	0.87
KNN	0.72	0.8	0.63	0.88	0.67	0.84
Gaussian NB	0.79	0.81	0.72	0.85	0.75	0.83
Logistic Regression	0.81	0.85	0.71	0.93	0.76	0.89
AdaBoost	0.86	0.83	0.68	0.95	0.76	0.89
Random Forest	0.77	0.85	0.73	0.89	0.75	0.87

Table 3.6: Precision, Recall and F1-Score for the Positive and Negative Classes using $\mathcal{F}_2$. Experiments were Run using 10-Fold Cross-Validation on the Labeled Data. The Best Performance is in Bold.

Learner	**Precision**		**Recall**		**F1-score**	
	$class_p$	$class_n$	$class_p$	$class_n$	$class_p$	$class_n$
S^3VM-R	0.91	0.91	0.91	0.91	0.91	0.91
Laplacian SVM	0.9	0.9	0.9	0.9	0.9	0.9
LabelSpreading (RBF)	0.8	0.86	0.82	0.83	0.81	0.84
LabelSpreading (KNN)	0.7	0.75	0.73	0.78	0.71	0.76
Co-Training (SVM)	**0.96**	0.91	0.91	**0.97**	0.93	**0.94**
SVM	0.93	0.91	0.84	**0.97**	0.88	**0.94**
KNN	0.87	0.92	0.88	0.94	0.87	0.93
Gaussian NB	0.78	**0.96**	**0.94**	0.87	0.85	0.91
Logistic Regression	0.98	0.89	0.81	0.98	0.89	0.93
AdaBoost	0.88	0.88	0.75	0.95	0.81	0.91
Random Forest	0.93	0.89	0.81	**0.97**	0.87	0.93

Table 3.7: Precision, Recall and F1-Score for the Positive and Negative Classes using $\{\mathcal{F}_1, \mathcal{F}_2\}$. Experiments were Run using 10-Fold Cross-Validation on the Labeled Data. The Best Performance is in Bold.

Learner	**Precision**		**Recall**		**F1-score**	
	$class_p$	$class_n$	$class_p$	$class_n$	$class_p$	$class_n$
$S^3VM - R$	**0.97**	**0.97**	**0.95**	**0.98**	**0.96**	**0.97**
Laplacian SVM	0.96	0.94	0.91	0.96	0.93	0.95
LabelSpreading (RBF)	0.83	0.86	0.82	0.84	0.82	0.85
LabelSpreading (KNN)	0.71	0.74	0.75	0.78	0.73	0.76
Co-Training (SVM)	0.92	0.9	0.9	0.94	0.91	0.92
SVM	0.96	0.92	0.84	0.97	0.9	0.94
KNN	0.84	0.83	0.67	0.95	0.75	0.89
Gaussian NB	0.77	0.96	0.94	0.87	0.85	0.91
Logistic Regression	0.95	0.9	0.79	0.97	0.86	0.93
AdaBoost	0.88	0.88	0.75	0.95	0.81	0.91
Random Forest	0.93	0.9	0.82	0.97	0.87	0.93

Table 3.8: Blind Evaluation of the Baselines on the Two Control Groups. The Best Performance is in Bold.

Learner	**AUC**		
	$\mathcal{F}_1$	$\mathcal{F}_2$	$\{\mathcal{F}_1, \mathcal{F}_2\}$
Laplacian SVM	**0.9**	**0.92**	**0.93**
LabelSpreading (RBF)	0.75	0.85	0.87
LabelSpreading (KNN)	0.7	0.82	0.79
Co-Training (SVM)	0.8	0.9	0.91
SVM	0.8	0.65	0.69
KNN	0.74	0.62	0.77
Gaussian NB	0.77	0.51	0.52
Logistic Regression	0.76	0.62	0.75
AdaBoost	0.77	0.74	0.74
Random Forest	0.8	0.8	0.8

Table 3.9: Significance of the Features in $\mathcal{F}_1$. The Check-marked Features Show the Top Features.

No.	Feature Group	χ^2	Selected
1	**Advertisement Language Pattern**		
	- Third person language	8.4	✓
	- First person plural pronouns	9.5	✓
	- Kolmogorov complexity	0.7	✓
	- *n*-grams (1)	0.4	
	- *n*-grams (2)	0.0	
	- *n*-grams (3)	0.4	
2	**Words and Phrases of Interest**	0.0	
3	**Countries of Interest**	59.3	✓
4	**Multiple Victims Advertised**	14.1	✓
5	**Victim Weight**	0.2	
6	**Reference to Website or Spa Massage Therapy**		
	- Reference to website	0.1	
	- Reference to Spa Massage Therapy	33.5	✓

Table 3.10: Classification Results (AUC) using 10-Fold Cross-Validation and Different Subsets of the Features on the Labeled Data.

Name	Value		
	$\overline{\mathcal{F}_1^*}$	$\mathcal{F}_1^*$	$\mathcal{F}_1$
S^3VM-R	0.82	0.87	0.91

Chapter 4

EARLY IDENTIFICATION OF PATHOGENIC SOCIAL MEDIA ACCOUNTS

The unregulated nature and rapid growth of the Web have raised numerous challenges, including hate speech Badjatiya *et al.* (2017), human trafficking Alvari *et al.* (2017) and disinformation spread Shaabani *et al.* (2018) which ultimately pose threats to users privacy Beigi *et al.* (2018); Beigi and Liu (2018a). Take disinformation spread as an example where "Pathogenic Social Media" (PSM) accounts (e.g., terrorist supporters, or fake news writers) Shaabani *et al.* (2018) seek to promote or degrade certain ideas by utilizing large online communities of supporters to reach their goals. Identifying PSMs has applications including countering terrorism Khader (2016); Klausen *et al.* (2016), fake news detection Gupta *et al.* (2014, 2013) and water armies detection Chen *et al.* (2011).

Early detection of PSMs in social media is crucial as they are likely to be *key* users to malicious campaigns Varol *et al.* (2017b). This is a challenging task for three reasons. First, these platforms are primarily based on reports they receive from their own users [1] to manually shut down PSMs which is not a timely approach. Despite efforts to suspend these accounts, many of them simply return to social media with different accounts. Second, the available data is often imbalanced and social network structure, which is at the core of many techniques Weng *et al.* (2014); Kempe *et al.* (2003); Beigi and Liu (2018b); Zhang *et al.* (2013); Beigi *et al.* (2019b); Sarkar *et al.* (2019); Alvari *et al.* (2016a), is not readily available. Third, PSMs often seek to utilize and cultivate large number of online communities of passive supporters to spread as much harmful information as they can.

[1] https://bit.ly/2Dq5i4M

In this chapter, causal inference is tailored to identify PSMs since they are key users in making a harmful message "viral"– where "viral" is defined as an order-of-magnitude increase. We propose *time-decay* causal metrics to distinguish PSMs from normal users within a *short* time around their activity. Our metrics alone can achieve high classification performance in identification of PSMs soon after they perform actions. Next, we pose the following research question: *Are causality scores of users within a community higher than those across different communities?* We propose a causal community detection-based classification method (C^{2}DC), that takes causality of users and the community structure of their action log.

Contributions. We make the following major contributions:

- We enrich the causal inference framework of Kleinberg and Mishra (2012) and present *time-decay* extensions of the causal metrics in Shaabani *et al.* (2018) for early identification of PSMs.
- We investigate the role of community structure in early detection of PSMs, by demonstrating that users within a community establish stronger causal relationships compared to the rest.
- We conduct a suit of experiments on a dataset from Twitter. Our metrics reached F1-score of 0.6 in identifying PSMs, half way their activity, and identified 71% of PSMs based on first 10 days of their activity, via supervised settings. The community detection approach achieved precision of 0.84 based on first 10 days of users activity; the misclassified accounts were identified based on their activity of 10 more days.

4.1 Technical Preliminaries

Following the convention of Goyal *et al.* (2010), we assume an *action log* $\mathbf{A}$ of the form *Actions(User,Action,Time)*, which contains tuples (u, a_u, t_u) indicating that user u has performed action a_u at time t_u. For ease of exposition, we slightly abuse the notation and use the tuple (u, m, t) to indicate that user u has posted (tweeted/retweeted) message m at time t. For a given message m we define a *cascade* of actions as $\mathbf{A}_m = \{(u, m', t) \in \mathbf{A} | m' = m\}$.

User u is said to be an m-participant if there exists t_u such that $(u, m, t_u) \in \mathbf{A}$. For users who have adopted a message in the early stage of its life span, we define *key users* as follows.

Definition 1 (Key Users). *Given message m, m-participant u and cascade $\mathbf{A}_m$, we say user u is a key user iff user u precedes at least ϕ fraction of other m-participants where $\phi \in (0, 1)$. In other words, $|\mathbf{A}_m| \times \phi \leq |\{j | \exists t' : (j, m, t') \in \mathbf{A} \wedge t < t'\}|$, where $|.|$ is the cardinality of a set.*

Next, we define viral messages as follows.

Definition 2 (Viral Messages). *Given a threshold θ, we say a message $m \in \mathbf{M}$ is viral iff $|\mathbf{A}_m| \geq \theta$. We denote a set of all viral messages by $\mathbf{M}_{vir}$.*

The prior probability of a message going viral is $\rho = |\mathbf{M}_{vir}|/|\mathbf{M}|$. The probability of a message going viral given key user u has participated in, is computed as follows:

$$\rho_u = \frac{|\{m | m \in \mathbf{M}_{vir} \wedge u \textit{ is a key user}\}|}{|\{m | m \in \mathbf{M} \wedge u \textit{ is a key user}\}|} \tag{4.1}$$

The probability that key users i and j tweet/retweet message m chronologically and make it viral, is computed as:

$$p_{i,j} = \frac{|\{m \in \mathbf{M}_{vir} | \exists t, t' : t < t' \wedge (i, m, t), (j, m, t') \in \mathbf{A}\}|}{|\{m \in \mathbf{M} | \exists t, t' : t < t' \wedge (i, m, t), (j, m, t') \in \mathbf{A}\}|} \tag{4.2}$$

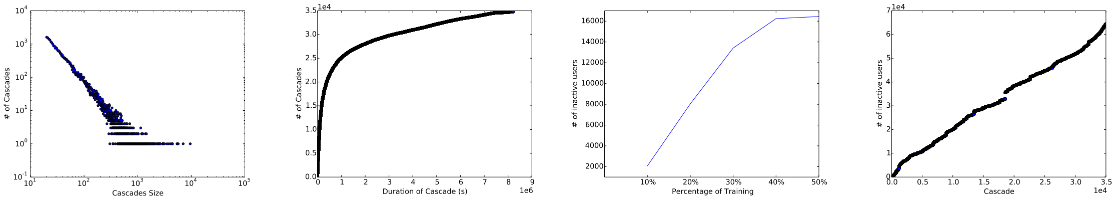


Figure 4.1: From Left to Right: Log-Log Distribution of Cascades vs. Cascade Size. Cumulative Distribution of Duration of Cascades. Number of Inactive Users in Different Subsets of the Training Set. Total Inactive Users in each Cascade.

To examine how causal user i was in helping a message m going viral, we shall explore what will happen if we exclude user i from m. We define the probability that *only* key user j has made a message m viral, i.e. user i has not posted m or does not precede j as:

$$p_{\neg i,j} = \frac{|\{m \in \mathbf{M}_{vir} | \exists t' : (j, m, t') \in \mathbf{A} \wedge \nexists t : t < t', (i, m, t) \in \mathbf{A}\}|}{|\{m \in \mathbf{M} | \exists t' : (j, m, t') \in \mathbf{A} \wedge \nexists t : t < t', (i, m, t) \in \mathbf{A}\}|} \tag{4.3}$$

Next, we adopt the notion of Prima Facie causes Suppes (1970):

Definition 3 (Prima Facie Causal Users). *A user u is said to be Prima Facie causal user for cascade $\mathbf{A}_m$ iff: (1) user u is a key user of m, (2) $m \in \mathbf{M}_{vir}$, and (3) $\rho_u > \rho$.*

We borrow the concept of *related users* from a rule-based system Stanton *et al.* (2015) which was an extension to the causal inference framework in Kleinberg and Mishra (2012). We say users i and j are m-related if (1) both are Prima Facie causal for m, and (2) i precedes j. We then define a set of user i's related users as $\mathbf{R}(i) = \{j | j \neq i \textit{ and } i, j \textit{ are } m\textit{-related}\}$.

4.1.1 Dataset

We collect a dataset (Table 6.6) of 53M ISIS related tweets/retweets in Arabic, from Feb 22, 2016 to May 27, 2016. The dataset has different fields including user ID, retweet ID, hashtags, content, posting time. The tweets were collected using 290 different hashtags such as #Terrorism and #StateOfTheIslamicCaliphate. We use a subset of this dataset which contains 35K cascades of different sizes and durations. There are ∼2.8M tweets/retweets associated with the cascades. After pre-processing and removing duplicate users from cascades, cascades sizes (i.e. number of associated postings) vary between 20 to 9,571 and take from 10 seconds to 95 days to finish. The log-log distribution of cascades vs. cascade size and the cumulative distribution of duration of cascades are depicted in Figure 4.1.

Based on the content of tweets in our dataset, PSMs are terrorism-supporting accounts who have participated in viral cascades. We chose to use $\theta = 100$ and take ∼6K viral cascades with at least 100 tweets/retweets. We demonstrate number of PSMs that have been suspended by the Twitter over time and total number of suspended users in each cascade, in Figure 4.1. We experiment the effectiveness of our proposed approach on subsets of the training set with different sizes. Note we use no more than 50% of original dataset to ensure our approach is able to identify PSMs early enough. The dataset does not have any underlying network. We only focus on the non-textual information in the form of an *action log.* We set $\phi = 0.5$ to select *key users* and after the data collection, we check through Twitter API whether they have been suspended (PSM) or they are active (normal) Thomas *et al.* (2011). According to Table 6.6, 11% of the users in our dataset are PSM and others are normal.

Table 4.1: Description of the Dataset.

Name	**Value**	
# of Cascades	35K	
# of Viral Cascades	6,602	
# of Tweets/Retweets	2,808,878	
# of Users	Suspended	Active
	64,484	536,609

4.1.2 Causal Measures

Causal inference framework was first introduced in Kleinberg and Mishra (2012). Later, Shaabani *et al.* (2018) adopted the framework and extended it to suite the problem of identifying PSMs. They extend the Kleinberg-Mishra causality ($\epsilon_{K\&M}$) to a series of causal metrics. To recap, we briefly explain them in the following discussion. Before going any further, $\epsilon_{K\&M}$ is computed as follows:

$$\epsilon_{K\&M}(i) = \frac{\sum_{j \in \mathbf{R}(i)} (p_{i,j} - p_{\neg i,j})}{|\mathbf{R}(i)|} \tag{4.4}$$

This metric measures how causal user i is, by taking the average of $p_{i,j} - p_{\neg i,j}$ over $\mathbf{R}(i)$. The intuition here is user i is more likely to be cause of message m to become viral than user j, if $p_{i,j} - p_{\neg i,j} > 0$. The work of Shaabani *et al.* (2018) devised a suit of the variants, namely relative likelihood causality (ϵ_{rel}), neighborhood-based causality (ϵ_{nb}) and its weighted version (ϵ_{wnb}). Note that none of these metrics were originally introduced for *early* identification of PSMs. Therefore, we shall make slight modifications to their notations to adjust our temporal formulations, using calligraphic uppercase letters. We define $\mathcal{E}_{K\&M}$ over a given time interval I as follows:

$$\mathcal{E}^{I}_{K\&M}(i) = \frac{\sum_{j\in\mathcal{R}(i)}(\mathcal{P}_{i,j} - \mathcal{P}_{\neg i,j})}{|\mathcal{R}(i)|} \tag{4.5}$$

where $\mathcal{R}(i)$, $\mathcal{P}_{i,j}$, and $\mathcal{P}_{\neg i,j}$ are now defined over I. Authors in Shaabani *et al.* (2018) mention that this metric cannot spot all PSMs. They define another metric, relative likelihood causality $\mathcal{E}_{rel}$, which works by assessing relative difference between $\mathcal{P}_{i,j}$, and $\mathcal{P}_{\neg i,j}$. We use its temporal version over I, $\mathcal{E}^{I}_{rel}(i) = \frac{\mathcal{S}(i,j)}{|\mathcal{R}(i)|}$.
where $\mathcal{S}(i,j)$ is defined as follows and α is infinitesimal:

$$\mathcal{S}(i,j) = \begin{cases} \frac{\mathcal{P}_{i,j}}{\mathcal{P}_{\neg i,j}+\alpha} - 1, & \mathcal{P}_{i,j} > \mathcal{P}_{\neg i,j} \\ 1 - \frac{\mathcal{P}_{\neg i,j}}{\mathcal{P}_{i,j}}, & \mathcal{P}_{i,j} \leq \mathcal{P}_{\neg i,j} \end{cases} \tag{4.6}$$

Two other neighborhood-based metrics were also defined in Shaabani *et al.* (2018), whose temporal variants are computed over I as $\mathcal{E}^{I}_{nb}(j) = \frac{\sum_{i\in\mathcal{Q}(j)}\mathcal{E}^{I}_{K\&M}(i)}{|\mathcal{Q}(j)|}$, where $\mathcal{Q}(j) = \{i|j \in \mathcal{R}(i)\}$ is the set of all users that user j belongs to their related users sets. Similarly, the second metric is a weighted version of the above metric and is called weighted neighborhood-based causality and is calculated as $\mathcal{E}^{I}_{wnb}(j) = \frac{\sum_{i\in\mathcal{Q}(j)} w_i \times \mathcal{E}^{I}_{K\&M}(i)}{\sum_{i\in\mathcal{Q}(j)} w_i}$. This is to capture different impacts that users in $Q(j)$ have on user j. We apply a threshold-based selection approach that selects PSMs from normal users, based on a given threshold. Following Shaabani *et al.* (2018), we use a threshold of 0.7 for all metrics except $\mathcal{E}^{I}_{rel}$ for which we used a threshold of 7 (Table 4.2).

4.2 The Proposed Framework

4.2.1 *Leveraging Temporal Aspects of Causality*

Previous causal metrics do not take into account time-decay effect. They assume a steady trend for computing causality scores. This is an unrealistic assumption, as causality of users may change over time. We introduce a generic decay-based metric.

Table 4.2: F1-Score Results for PSM Accounts using Causal Metrics Shaabani *et al.* (2018).

Metric	F1-score				
	10%	20%	30%	40%	50%
$\mathcal{E}^I_{K\&M}$	0.41	0.42	0.45	0.46	0.49
$\mathcal{E}^I_{rel}$	0.3	0.31	0.33	0.35	0.37
$\mathcal{E}^I_{nb}$	0.49	0.51	0.52	0.54	0.55
$\mathcal{E}^I_{wnb}$	**0.51**	**0.52**	**0.55**	**0.56**	**0.59**

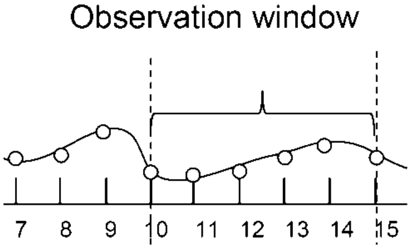


Figure 4.2: An Illustration of How Decay-based Causality Works. To Compute $\xi^I_k(i)$ over $I = [t_0, t]$, We Use a Sliding Window $\Delta = [t' - \delta, t']$ and Take the Average Between the Resultant Causality Scores $e^{-\sigma(t-t')} \times \mathcal{E}^{\Delta}_k(i)$.

Our metric assigns different weights to different time points of a given time interval, inversely proportional to their distance from t (i.e., smaller distance is associated with higher weight). Specifically, it performs the following: it (1) breaks down the given time interval into shorter time periods using a sliding time window, (2) deploys an exponential decay function of the form $f(x) = e^{-\alpha x}$ to account for the time-decay effect, and (3) takes average of the causality values computed over each sliding time window. Formally, ξ^I_k is defined as follows, where $k \in \{K\&M, rel, nb, wnb\}$:

$$\xi^I_k(i) = \frac{1}{|\mathcal{T}|} \sum_{t' \in \mathcal{T}} e^{-\sigma(t-t')} \times \mathcal{E}^{\Delta}_k(i) \qquad (4.7)$$

Table 4.3: F1-Score Results for PSM Accounts using each Decay-based Metric with and without Communities.

Metric	F1-score (without/with communities)				
	10%	20%	30%	40%	50%
$\xi^I_{K\&M}$	0.44/0.49	0.46/0.51	0.47/0.52	0.5/0.54	0.53/0.57
ξ^I_{rel}	0.36/0.4	0.38/0.43	0.39/0.46	0.41/0.49	0.42/0.5
ξ^I_{nb}	0.52/0.56	0.53/0.57	0.54/0.58	0.56/0.6	0.59/0.61
ξ^I_{wnb}	**0.54/0.57**	**0.55/0.58**	**0.57/0.6**	**0.58/0.62**	**0.6/0.63**

where σ is a scaling parameter of the exponential decay function, $\mathcal{T} = \{t'|t' = t_0 + j \times \delta, j \in \mathbb{N} \wedge t' \leq t - \delta\}$ is a sequence of sliding-time windows, and δ is a small fixed amount of time, which is used as the length of each sliding-time window $\Delta = [t' - \delta, t']$ (Figure 4.2). To apply the threshold-based approach, we once again use a threshold of 0.7 for all metrics except ξ^I_{rel} for which we used a threshold of 7 (Table 4.3).

Early Detection of PSMs. *Given action log* $\mathbf{A}$*, and user* u *where* $\exists t$ *s.t.* $(u, m, t) \in \mathbf{A}$*, our goal is to determine if u's account shall be suspended given its causality vector* $\mathbf{x}_u \in \mathbb{R}^d$ *(here,* $d = 4$*) computed using any of the causality metrics over* $[t - \delta, t + \delta]$*.*

4.2.2 Leveraging Community Structure Aspects of Causality

To answer the research question posed earlier, since network structure is not available, we need to build a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ from $\mathbf{A}$ by connecting any pairs of users who have posted same message *chronologically.* In this graph, $\mathbf{V}$ is a set of vertices (i.e. users) and $\mathbf{E}$ is a set of directed edges between users. For the sake of simplicity and without loss of generality, we make the edges of this graph undirected. Next, we leverage the Louvain algorithm Blondel *et al.* (2008) to find the partitions

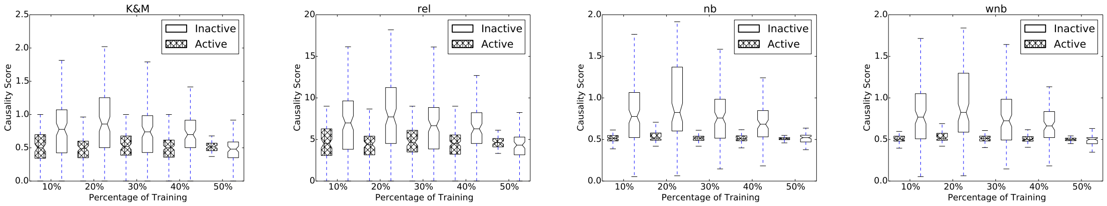


Figure 4.3: From Left to Right: Distributions of Active and Inactive Users using Communities and ξ_k^I when $k \in \{K\&M, rel, nb, wnb\}$.

$\mathbf{C} = \{C_1, C_2, ..., C_k\}$ of k communities over $\mathbf{G}$. Among a myriad of the community detection algorithms Alvari *et al.* (2016a); Lancichinetti *et al.* (2011); Alvari *et al.* (2011), we chose Louvain due to its fast runtime and scalability– we leave examining other community detection algorithms to future work. Next, we perform the two-sample t-test $H_0 : v_a \geq v_b,\ H_1 : v_a < v_b$. The null hypothesis is: *users in a given community establish weak causal relations with each other as opposed to the other users in other communities.* We construct two vectors v_a and v_b as follows. We create v_a by computing Euclidean distances between causality vectors $(\mathbf{x}_i, \mathbf{x}_j)$ corresponding to each pair of users (u_i, u_j) who are from same community $C_l \in \mathbf{C}$. Therefore, v_a contains exactly $\frac{1}{2}\sum_{l=1}^{|\mathbf{C}|} |C_l|.(|C_l| - 1)$ elements. We construct v_b of size $\sum_{l=1}^{|\mathbf{C}|} |C_l|$ by computing Euclidean distance between each user u_i in community $C_l \in \mathbf{C}$, and a random user u_k chosen from the rest of the communities, i.e., $\mathbf{C} \setminus C_l$. The null hypothesis is rejected at significance level $\alpha = 0.01$ with the p-value of 4.945e-17. We conclude that users in same communities are more likely to establish stronger causal relationships with each other than the rest of the communities. The answer to the question is thus positive. For brevity, we only reported results for 10% of the training set, while making similar arguments for other percentages is straightforward. Figure 4.3 shows box plots of the distributions of users using the decay-based metrics and the communities and same set of thresholds as before. We observe a clear distinction between active/suspended accounts, using the community structure. Results in Table 4.3 show improvements over previous ones.

Algorithm 2 Causal Community Detection-based Classification Algorithm ($\mathbf{C^2dc}$)

```
Input: Training samples $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ , tests $\{\mathbf{x}'_1, ..., \mathbf{x}'_n\}$, $\mathbf{G}$, $k$
Output: Predicted labels $\{y'_1, ..., y'_n\}$
$\mathbf{C} \leftarrow$ Louvain($\mathbf{G}$)
for each $\mathbf{x}'_i$ do
    $C_l \leftarrow C' \in \mathbf{C}$ s.t. $\mathbf{x}'_i \in C'$
    $\mathbf{D} \leftarrow \{\}$
    for each $\mathbf{x}_j \in C_l$ do
        $d_{ij} \leftarrow ||\mathbf{x}'_i - \mathbf{x}_j||_2$
        $\mathbf{D} \leftarrow \mathbf{D} \cup \{d_{ij}\}$
    end for
    $\mathbf{K} \leftarrow$ Knn($\mathbf{D}$, $k$)
    $y'_i \leftarrow$ Dominant-Label($\mathbf{K}$)
end for
```

First step of the proposed algorithm (Algorithm 3) involves finding the communities. In the second step, each unlabeled user is classified based on the available labels of her nearby peers in the same community. We use the K-Nearest Neighbors (Knn) algorithm to compute her k nearest neighbors in the same community, based on Euclidean distances between their causality vectors. We label her based on the majority class of her k nearest neighbors in the community. The merit of using community structure over merely using Knn is, communities can give finer-grained and more accurate sets of neighbors sharing similar causality scores.

4.3 Experiments

We use different subsets of size $x\%$ of the entire time-line (from Feb 22, 2016 to May 27, 2016) of the action log $\mathbf{A}$, by varying x as $\{10, 20, 30, 40, 50\}$. For each subset and user i in the subset, we compute feature vector $\mathbf{x}_i \in \mathbb{R}^4$ of the corresponding causality scores. The feature vectors are then fed into supervised classifiers and the community detection-based algorithm. For the sake of fair comparison, we perform this for both causal and decay-based metrics. For both metrics, we empirically found that $\rho = 0.1$ and $\alpha = 0.001$ work well. For the decay-based causality metric we shall also assume a sliding window of size of 5 days (i.e. $\delta = 5$) and set $\sigma = 0.001$ which were found to work well in our experiments. Note we only present results for PSMs. Among many other supervised classifiers such as AdaBoost , Logistic Regression and Support Vector Machines (Svm), Random Forest (RF) with 200 estimators and 'entropy' criterion, achieved the best performance. Therefore, for brevity we only report results when RF is used as the classifier.

We present results for the proposed community detection-based framework and causal and decay-based metrics. For computing k nearest neighbors, we set $k = 10$ as it was found to work well for our problem. By reporting the results of Knn trained on the decay-based causality features, we stress that using Knn alone does not yield a good performance. For the sake of fair comparison, all approaches were implemented and run in Python 2.7x, using the scikit-learn package. For any approach that requires special tuning of parameters, we conducted grid search to choose the best set of parameters.

4.3.1 Baseline Methods

Causal Shaabani *et al.* (2018)

We compare our metrics against the ones in Shaabani *et al.* (2018) via supervised and community detection settings.

SentiMetrix-Dbscan Subrahmanian *et al.* (2016)

This was the winner of the DARPA challenge. It uses several features such as tweet syntax (e.g., average number of hashtags, average number of links), tweet semantics (e.g., LDA topics), and user behavior (e.g., tweet frequency). We perform 10-fold cross validation and use a held-out test set for evaluation. This baseline uses a seed set of 100 active and 100 inactive accounts, and then use Dbscan clustering algorithm to find the associated clusters. Available labels are propagated to nearby unlabeled users in each cluster based on the Euclidean distance metric, and labels of the remaining accounts are predicted using Svm.

SentiMetrix-RF Subrahmanian *et al.* (2016)

This is a variant of Subrahmanian *et al.* (2016) where we excluded the Dbscan part and instead trained RF classifier using only the above features to evaluate the feature set.

4.3.2 Identification of PSM Accounts

For each subset a separate 10-fold cross validation was performed (Figure 4.4). We observe the following:

- Community detection achieves the best performance using several metrics. This

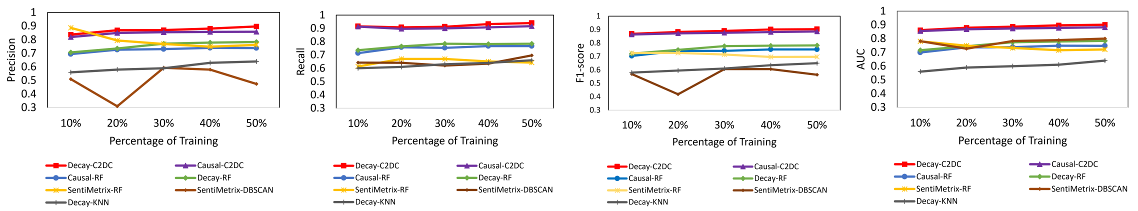


Figure 4.4: Precision, Recall, F1-Score and AUC Results for each Classifier. Experiments were Run using 10-Fold Cross-Validation.

aligns well with the t-test results discussed earlier: *taking into account community structure of PSMs can boost the performance.*

- Causal and decay-based metrics mostly achieve higher performance than other approaches via both settings.

- Decay-based metrics are effective at identifying PSMs at different intervals via both settings. This lies at the inherent difference between decay-based and causal metrics– our metrics take into account time-decay effect.

- Although both variants of SentiMetrix-Dbscan use many features, they were unable to defeat our approach.

4.3.3 Timeliness of PSM Accounts Identification

For each approach, we would like to see *how many* of PSMs who were active in the first 10 days of the dataset, are correctly classified (i.e., true positives) as time goes by. Also, we need to keep track of false positives to ensure given approach does not merely label each instance as positive– otherwise a trivial approach that always label each instance as PSM would achieve highest performance. We are also interested to figure *how many* days need to pass to find these accounts. We train each classifier using 50% of the first portion of the dataset, and use a held-out set of the rest for evaluation. Next, we pass along the misclassified PSMs to the next portions to see how many of them are captured over time. We repeat the process until reaching 50% of the dataset– each time we increase the training set by adding new instances of the current portion.

There are 14,841 users in the first subset from which 3,358 users are PSMs. Table 4.4 shows the number of users from the first portion that (1) are correctly classified as PSM (out of 3,358), (2) are incorrectly classified as PSM (out of 29,617), over time. Community detection approaches were able to detect all PSMs who were active in the first 10 days of our dataset, no later than a month from their first activity. DECAY-C^{2}DC, identified all of these PSMs in about 20 days since the first time they posted a message. Also, both causal and decay-based metrics when fed to RF classifier, identified all of the PSMs in the first period. SENTIMETRIX-DBSCAN and SENTIMETRIX-RF failed to detect all PSMs from the first portion, even after passing 50 days since their first activity. Furthermore, these two baselines generated much higher rates of false positives compared to the rest. The observations we make here are in line with the previous ones: *the proposed community detection-based framework is more effective and efficient than the rivals.*

4.4 Conclusion

We enriched the existing causal inference framework to suite the problem of early identification of PSMs. We proposed time-decay causal metrics which reached F1-score of 0.6 and via supervised learning identified 71% of the PSMs from the first 10 days of the dataset. We proposed a causal community detection-based classification algorithm, by leveraging community structure of PSMs and their causality. We achieved the precision of 0.84 for detecting PSMs within 10 days around their activity; the misclassified accounts were then detected 10 days later. Our future plan

Table 4.4: True/False Positives for PSM Accounts. Numbers are out of 3,358/29,617 PSM/Normal Accounts from the First Period. Last Column Shows the Number of PSM Accounts From the First Period which were not Caught.

Learner	**True Positives/False Positives**					**Remaining**
	02/22-03/02	03/02-03/12	03/12-03/22	03/22-03/31	03/31-04/09	
DECAY-C^{2}DC	3,072/131	286/0	0/0	0/0	0/0	0
CAUSAL-C^{2}DC	3,065/156	188/20	105/0	0/0	0/0	0
DECAY-KNN	2,198/459	427/234	315/78	109/19	96/0	213
DECAY-RF	2,472/307	643/263	143/121	72/68	28/0	0
CAUSAL-RF	2,398/441	619/315	221/169	89/70	51/0	0
SENTIMETRIX-RF	2,541/443	154/0	93/0	25/0	14/0	531
SENTIMETRIX-DBSCAN	2,157/2,075	551/5,332	271/209	92/118	72/696	215

includes exploring other community detection algorithms and other forms of causal metrics.

Chapter 5

SEMI-SUPERVISED CAUSAL INFERENCE FOR DETECTING PATHOGENIC USERS IN SOCIAL MEDIA

Over the past years, social media play major role in massive dissemination of misinformation online. Political events and public opinion on the Web and social networks have been allegedly manipulated by different forms of accounts including real users and automated software (a.k.a social bots or sybil accounts). Pathogenic Social Media (PSM) accounts are among those that are responsible for such a massive spread of disinformation online and swaying normal users' opinion Alvari *et al.* (2018); Shaabani *et al.* (2019). These accounts (1) are usually owned by either normal users or automated bots, (2) seek to promote or degrade certain ideas; and (3) can appear in many forms such as terrorist supporters (e.g., ISIS supporters), water armies or fake news writers. Understanding the behavior of PSMs will allow social media to take countermeasures against their propaganda at the early stage and reduce their threat to the public.

The problem of identification of PSMs has long been addressed in the past by the research community mostly in the form of bot detection. Several approaches especially supervised learning methods have been proposed in the literature and they have shown promising results Kudugunta and Ferrara (2018). However, for the most part, these approaches are often based on labeled data and exhaustive feature engineering. Examples of such feature groups include but are not limited to content, sentiment of posts, profile information and network features. These approaches are thus very expensive as they require significant amount of efforts to design features and annotate large labeled datasets. In contrast, unlabeled data is ubiquitous and cheap to collect thanks to the massive user-generated data produced on a daily basis. Thus, in this work we set out to examine if unlabeled instances can be utilized to compensate for the lack of enough labeled data.

In this chapter, semi-supervised causal inference is tailored to detect PSMs who are promoters of misinformation online. We cast the problem of identifying PSMs as an optimization problem and propose a semi-supervised causal learning framework which utilizes unlabeled examples through manifold regularization Belkin *et al.* (2006). In particular, we incorporate causality-based features extracted from users' activity log (i.e., cascades of retweets) as regularization terms into the optimization problem. In this work, causal inference is leveraged in an effort to capture whether or not PSMs exert causal influences while making a message viral. Our causality-based features are built upon *Suppes' theory of probabilistic causation* Suppes (1970) whose central concept is *prima facie causes*: an event to be recognized as a cause, must occur before the effect and must lead to an increase of the likelihood of observing the effect. While there exists a prolific literature on causality and their great impact in the computer-science community (see Pearl (2009) for instance), we build our foundation on *Suppes' theory* as it is computationally less complex.

Key idea and highlights. To summarize, this chapter makes the following main contributions:

- We frame the problem of detecting PSM accounts as an optimization problem and present a Laplacian semi-supervised causal inference SemiPsm for solving it. The unlabeled data are utilized via manifold regularization.

- Manifold regularization used in the resultant optimization formulation is built upon causality-based features created on a notion of *Suppes' theory of probabilistic causation.*

- We conduct a suite of experiments using different supervised and semi-supervised methods. Empirical experiments on a real-world ISIS-related dataset from Twitter suggests the effectiveness of the proposed semi-supervised causal inference over the existing methods.

5.1 The Proposed Method

We leverage the time-decay causal inference introduced in Chapter 4 built on Suppes' theory, to compute causality-based features for users. Then, we detail the proposed semi-supervised causal inference, namely SemiPsm for detecting PSM accounts.

5.1.1 *Causality-based Attributes*

The time-decay causal metrics Alvari *et al.* (2018) will be fed as features to the semi-supervised framework– this will be described in the next section. The final set of features is in the following generic form ξ_k^I where $k \in \{K\&M, rel, nb, wnb\}$:

$$\xi_k^I(i) = \frac{1}{|\mathcal{T}|} \sum_{t' \in \mathcal{T}} e^{-\sigma(t-t')} \times \mathcal{E}_k^{\Delta}(i) \tag{5.1}$$

Here, σ is a scaling parameter of the exponential decay function, $\mathcal{T} = \{t' | t' = t_0 + j \times \delta, j \in \mathbb{N} \wedge t' \leq t - \delta\}$ is a sequence of sliding-time windows, and δ is a small fixed amount of time, which is used as the length of each sliding-time window $\Delta = [t' - \delta, t']$.

In essence, this metric assigns different weights to different time points of a given time interval, inversely proportional to their distance from t (i.e., smaller distance is associated with higher weight). Specifically, it performs the following: it (1) breaks down the given time interval into shorter time periods using a sliding time window, (2) deploys an exponential decay function of the form $f(x) = e^{-\alpha x}$ to account for the time-decay effect, and (3) takes average of the causality values computed over each sliding time window Alvari *et al.* (2018).

5.1.2 Semi-Supervised Causal Inference

Having defined the causality-based features, we now proceed to present the proposed semi-supervised Laplacian SVM framework, SemiPsm. For the rest of the discussion, we shall assume a set of l labeled pairs $\{(x_i, y_i)\}_{i=1}^{l}$ and an unlabeled set of u instances $\{x_{l+i}\}_{i=1}^{u}$, where $x_i \in \mathbb{R}^n$ denotes the causality vector $\xi_k^I(i)$ of user i and $y_i \in \{+1, -1\}$ (PSM or not).

Recall for the standard soft-margin support vector machines, the following optimization problem is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||_k^2 + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) \tag{5.2}$$

In the above equation, $f_\theta(\cdot)$ is a decision function of the form $f_\theta(\cdot) = w.\mathbf{\Phi}(\cdot) + b$ where $\theta = (w, b)$ are the parameters of the model, and $\mathbf{\Phi}(\cdot)$ is the feature map which is usually implemented using the kernel trick Cortes and Vapnik (1995). Also, the function $H_1(\cdot) = \max(0, 1 - \cdot)$ is the Hinge Loss function. The classical Representer theorem Belkin *et al.* (2005) suggests that solution to the optimization problem exists in a Hilbert space $\mathcal{H}_k$ and is of the form $f_\theta^*(x) = \sum_{i=1}^{l} \alpha_i^* \mathbf{K}(x, x_i)$. Here, $\mathbf{K}$ is the $l \times l$ Gram matrix over labeled samples. Equivalently, the above problem can be written as:

$$\min_{w,b,\epsilon} \frac{1}{2} ||w||_2^2 + C_l \sum_{i=1}^{l} \epsilon_i \tag{5.3}$$

$$s.t. \quad y_i(w.\mathbf{\Phi}(x_i) + b) \geq 1 - \epsilon_i, \; i = 1, ..., l$$

$$\epsilon_i \geq 0, \; i = 1, ..., l \tag{5.4}$$

Next, we will use the above optimization equation as our basis to derive the formulations for our proposed semi-supervised learner.

The basic assumption behind semi-supervised learning methods is to leverage unlabeled instances in order to restructure hypotheses during the learning process Alvari *et al.* (2017). Here, exogenous information extracted from causality-based features of users is exploited to make a better use of the unlabeled examples. To do so, we first introduce matrix $\mathbf{F}$ over both of the labeled and unlabeled samples with $\mathbf{F}_{ij} = ||\mathbf{\Phi}(x_i) - \mathbf{\Phi}(x_j)||_2$ in $||.||_2$ norm. This way, we force instances x_i and x_j in our dataset to be relatively 'close' to each other Beigi and Liu (2018b), i.e., having a same label, if their corresponding causal-based feature vectors are close. To account for this, a regularization term is added to the standard equation and the following optimization is solved:

$$\min_{f_\theta \in \mathcal{H}_k} \frac{1}{2} \sum_{i=1}^{l} \mathbf{F}_{ij} ||f_\theta(x_i) - f_\theta(x_j)||_2^2 = \mathbf{f}_\theta^T \mathcal{L}^T \mathbf{f}_\theta \tag{5.5}$$

where $\mathbf{f} = [f(x_1), ..., f(x_{l+u})]^T$ and $\mathcal{L}$ is the Laplacian matrix based on $\mathbf{F}$ given by $\mathcal{L} = \mathbf{D} - \mathbf{F}$, and $\mathbf{D}_{ii} = \sum_{j=1}^{l+u} \mathbf{F}_{ij}$. The intuition here is that causal pairs are more likely to have same labels than others.

Following the notations used in Chapter 3, and by including our regularization term, we would extend the standard equation by solving the following optimization:

$$\min_{f_\theta \in \mathcal{H}_k} \gamma ||f_\theta||_k^2 + C_l \sum_{i=1}^{l} H_1(y_i f_\theta(x_i)) + C_r \mathbf{f}_\theta^T \mathcal{L} \mathbf{f}_\theta \tag{5.6}$$

Where solution in $\mathcal{H}_k$ is in the following form $f_\theta^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(x, x_i)$. Here $\mathbf{K}$ is the $(l+u) \times (l+u)$ Gram matrix over all samples.

Next, we follow the procedure explained in Chapter 3 to obtain the dual problem in the form of a quadratic programming problem:

$$\beta^* = \text{argmax}_{\beta \in \mathbb{R}^l} \; -\frac{1}{2}\beta^T \mathbf{Q}\beta + \sum_{i=1}^{l} \beta_i \tag{5.7}$$

$$s.t. \quad \sum_{i=1}^{l} \beta_i y_i = 0$$

$$0 \leq \beta_i \leq C_l \tag{5.8}$$

where $\beta = [\beta_1, ..., \beta_l]^T \in \mathbb{R}^l$ are the Lagrangian multipliers and $\mathbf{Q}$ is obtained as follows:

$$\mathbf{Q} = \mathbf{YJK}(I + (C_r\mathcal{L})\mathbf{K})^{-1}\mathbf{J}^T\mathbf{Y} \tag{5.9}$$

We summarize the proposed semi-supervised framework in Algorithm 1. Our optimization problem is very similar to the standard optimization problem solved for SVMs, hence we use a standard optimizer for SVMs to solve our problem.

Algorithm 3 Semi-Supervised Causal Inference for PSM detection (SemiPsm)

```
Input: $\{(x_i, y_i)\}_{i=1}^{l}$, $\{x_{l+i}\}_{i=1}^{u}$, $C_l$, $C_r$.
Output: Estimated function $f_\theta : \mathbb{R}^n \to \mathbb{R}$
1: Construct matrix $\mathbf{F}$ based on the causality-based features
2: Compute the corresponding Laplacian matrix $\mathcal{L}$.
3: Construct the Gram matrix over all examples using $\mathbf{K}_{ij} = k(x_i, x_j)$ where $k$ is a kernel function.
4: Compute $\alpha^*$ and $\beta^*$ using Eq. 3.18 and Eq. 5.7 and a standard QP solvers.
5: Compute function $f_\theta^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(x, x_i)$
```

Table 5.1: Description of the Dataset.

Name	**Value**	
# of Cascades	35 K	
# of Viral Cascades	6,602	
# of Tweets/Retweets	10,823,168	
# of Users	PSM	Normal
	19,859	65,417

5.1.3 Computational Complexity

Here, we will explain the scalability of the algorithm in terms of big-$\mathcal{O}$ notation for both constituents of the proposed framework separately. For the first part of the approach, given a set of $\mathcal{A}$ cascades, and average number of $avg(\tau)$ users' actions (i.e., timestamps) in each cascade where $\tau \in \mathcal{A}$, the complexity of computing causality scores is $\mathcal{O}(|\mathcal{A}|.(avg(\tau))^2)$ (note on average there are $(avg(\tau))^2$ pairs of users in each cascade). For the second part, i.e., learning the semi-supervised algorithm, the most time-consuming part is calculating the inverse of a dense Gram matrix which leads to $\mathcal{O}((l+u)^3)$ complexity, where l and u are number of labeled and unlabeled instances Belkin *et al.* (2006).

5.2 Experiments

In this section we conduct experiments on the Twitter ISIS-related dataset described in Chapter 4 and present results for several supervised and semi-supervised approaches. We first explain the dataset and provide some data analysis. Then, we will present the baseline methods. Finally, results and discussion are provided.

Baseline Methods

We compare the proposed method SemiPsm against the following baseline methods. Note for all methods, we only report results when their best settings are used.

- **LabelSpreading (RBF Kernel) Zhou *et al.* (2004a).** This is a graph inference-based label spreading approach with radial basis function (RBF) kernel.

- **LabelSpreading (KNN Kernel) Zhou *et al.* (2004a).** Similar to the previous approach with K-nearest neighbor (KNN) kernel.

- **LSTM Kudugunta and Ferrara (2018).** The word-level LSTM approach here is similar to the deep neural network models used for sequential word predictions. We adapt the neural network to a sequence classification problem where the inputs are the vector of words in each tweet and the output is the predicted label of the tweet. We first use the word2vec Mikolov *et al.* (2013) embedding pre-trained from a set of tweets similar to the data representation in our Twitter dataset.

- **Account-Level (RF Classifier) Kudugunta and Ferrara (2018)** This approach uses the following features of the user profiles: *Statuses Count, Followers Count, Friends Count, Favorites Count, Listed Count, Default Profile, Geo Enables, Profile Uses Background Image, Verified, Protected.* We chose this method over Botometer Varol *et al.* (2017a) as it achieved comparable results with far less number of features (Varol *et al.* (2017a) uses over 1,500 features)(see also Ferrara *et al.* (2016)). According to Kudugunta and Ferrara (2018), we report the best results when Random Forest (RF) is used.

- **Tweet-Level (RF Classifier) Kudugunta and Ferrara (2018).** Similar to the previous baseline, this method uses only a handful of features extracted from tweets: *retweet count, reply count, favorite count, number of hashtags, number of URLs, number of mentions.* Likewise, we use RF as the classification algorithm.

- **SentiMetrix Subrahmanian *et al.* (2016).** This approach was proposed by the top-ranked team in the DARPA Twitter Bot Challenge. We consider all features that we could extract from our dataset. Our features include tweet syntax (average number of hashtags, average number of user mentions, average number of links, average number of special characters), tweet semantics (LDA topics), and user behaviour (tweet spread, tweet frequency, tweet repeats). The proposed approach starts with a small seed set and propagates the labels. Since we have enough labeled data for the training part, we use Random Forest as the learning approach.

- **C^{2}DC Alvari *et al.* (2018).** This approach uses time-decay causal community detection-based classification to detect PSM accounts Alvari *et al.* (2018). For community detection, this approach uses Louvain algorithm.

Results and Discussion

All experiments were implemented in Python 2.7x and run on a machine equipped with an Intel(R) Xeon(R) CPU of 3.50 GHz with 200 GB of RAM running Linux. The proposed approach was implemented using CVXOPT package. Furthermore, we split the whole dataset into 50% training and 50% test sets for all experiments. We report results in terms of F1-score in tables 5.2 and 5.3. For any approach that requires special tuning of parameters, we conducted grid search to choose the best set of parameters. Specifically, for the proposed approach, we set the penalty

Table 5.2: F1-Score Results of Various Methods on the Labeled Data. For Semi-Supervised Learners, the Size of the Unlabeled Data is Fixed to 10% of the Training Set. The Best Performance is in Bold.

Learner	**F1-score**
SemiPSM (Causal Features)	**0.94**
SemiPSM (Account-Level Features)	0.89
SemiPSM (Tweet-Level Features)	0.88
LabelSpreading (KNN/Causal Features)	0.89
LabelSpreading (RBF/Causal Features)	0.88
Account-Level (RF Classifier)	0.88
Tweet-Level (RF Classifier)	0.82
SentiMetrix	0.54
LSTM	0.41
C^{2}DC	0.4

parameter as $C_l = 0.6$ and the regularization parameter $C_r = 0.2$, and used linear kernel. For LabelSpreading (Rbf), the default vale of $\gamma = 20$ was used and for LabelSpreading (Knn), number of neighbors was set to 5. Also, for random forest we used 200 estimators and the 'entropy' criterion was used. For computing k nearest neighbors in C2dc, we set $k = 10$.

Furthermore for LSTM, we preprocessed the individual tweets in line with the steps mentioned in Soliman *et al.* (2017). Since the content of the tweets are in Arabic, we replaced special characters that were present in the text with their Arabic counterparts if they were present. We used word vectors of dimensions 100 and deployed the skip-gram technique for obtaining the word vectors where the input is the target word, while the outputs are the words surrounding the target words.

Table 5.3: F1-Score Results of the Semi-Supervised Approaches when Causality-based Features are Used. Results are Reported on Different Portions of the Unlabeled Data. The Best Performance is in Bold.

	Percentage of Unlabeled Data				
	10%	**20%**	**30%**	**40%**	**50%**
SemiPsm	**0.94**	**0.93**	**0.91**	**0.9**	**0.88**
LabelSpreading (Knn)	0.89	0.88	0.87	0.85	0.81
LabelSpreading (Rbf)	0.88	0.86	0.85	0.82	0.80

To model the tweet content in a manner that uses it to predict whether an account is PSM or not, we used Long Short Term Memory (LSTM) models Hochreiter and Schmidhuber (1997). For the LSTM architecture, we used the first 20 words in the tokenized Arabic text of each tweet and use padding in situations where the number of tokens in a tweet are less than 20. We used 30 units in the LSTM architecture (many to one). The output of the LSTM layer was fed to a dense layer of 32 units with ReLU activations. We added dropout regularization following this layer to avoid overfitting and the output was then fed to a dense layer which outputs the category of the tweets.

We depict in Table 5.2 classification performance of all approaches on the labeled data. For the proposed framework SemiPsm, we examine three sets of features (1) causality-based features, (2) account-level features Kudugunta and Ferrara (2018); and (3) tweet-level features Kudugunta and Ferrara (2018). For the graph inference-based semi-supervised algorithms, i.e., LabelSpreading (Rbf) and LabelSpreading (Knn), we only report results where causality-based features are used as they achieved best performance with them. As it is observed from the table, the best results in terms of F1-score belong to SemiPsm where causality-based features are used. The runner-up is SemiPsm with account-level features and the next best approach is SemiPsm where tweet-level features are deployed. This clearly demonstrates the significance of using manifold regularization in the Laplacian semi-supervised framework over using other semi-supervised methods, LabelSpreading (Rbf) and LabelSpreading (Knn).

We further note that the supervised classifier Random Forest using both of the account-level and tweet-level features and the whole labeled dataset achieve worse or comparable results to the semi-supervised learners. The fact that obtaining several tweet and account-level features is not trivial and do not necessarily lead to the best classification performance, motivates us to use semi-supervised algorithms which use less number of labeled examples, and yet achieve competing performance. We also obtain an F1-score of 0.41 when LSTM is used– the poor performance of the this neural network model can be attributed to the raw Arabic text content. It suggests that the Arabic tokens as a representation might not be very informative about the category of accounts it has been generated from and some kind of weighting might be necessary before the LSTM module is used.

Also, Table 5.3 shows the classification performance of the semi-supervised approaches with causality-based features. The results are achieved using different portions of the unlabeled data, i.e., $\{10\%, 20\%, 30\%, 40\%, 50\%\}$ of the training set. As it is seen in the table, SemiPsm achieves the best performance on different portions of the unlabeled data compared to the other semi-supervised learners, while performances of all approaches deteriorate with increasing the percentage of the unlabeled data. Furthermore, SemiPsm still outperforms all other supervised methods as well as Lstm and C2dc when up to 50% of the data has been made unlabeled.

Observations. Overall, this work makes the following observations:

- Among the semi-supervised learners used in this study, SemiPsm achieves the best classification performance suggesting the significance of using unlabeled instances in the form of manifold regularization. Manifold regularization is shown effective in boosting the classification performance, with three different sets of features confirming this.

- Causality-based features achieve the best performance via both Laplacian and graph inference-based semi-supervised settings. This lies at the inherent property of the causality-based features– they are designed to show whether or not user i exerts a causal influence on j. This is effective in capturing PSMs as they are key users in making a message viral.

- Compared to the supervised methods Account-Level (Rf) and Tweet-Level (Rf), semi-supervised learners achieve either comparable or best results, suggesting promising results with less number of labeled examples.

- Among the supervised methods Account-Level (Rf) and Tweet-Level (Rf), the former achieves higher F1-score indicating that account-level features are more useful in boosting the performance, although they are harder to obtain Kudugunta and Ferrara (2018).

- Semi-supervised learners achieve best or comparable results with supervised learners, even with up to 50% of the data made unlabeled. This clearly shows the superiority of using unlabeled examples over labeled ones.

5.3 Conclusion

We presented a semi-supervised Laplacian SVM to detect PSM users in social media who are promoters of misinformation spread. We cast the problem of identifying PSMs as an optimization problem and introduced a Laplacian semi-supervised SVM via utilizing unlabeled examples through manifold regularization. We examined different sets of features extracted from users activity log (in the form of cascades of retweets) as regularization terms: (1) causality-based features; and (2) LSTM-based features. Our causality-based features were built upon *Suppes' theory of probabilistic causation*. The LSTM-based features were extracted via LSTM which has shown promising results for different tasks in the literature.

Chapter 6

FEATURE-DRIVEN APPROACH TO DETECT PATHOGENIC SOCIAL MEDIA ACCOUNTS

Following the PSM detection works presented in Chapters 4 and 5, here, we first set out to understand differences between Pathogenic Social Media (PSM) accounts and normal users in terms of URLs they share online. We then incorporate various characteristics of URLs shared online (e.g., URL address, content of the associated website, etc.) Baly *et al.* (2018); Phuong *et al.* (2014); Entman (1993); Morstatter *et al.* (2018); Kincaid *et al.* (1975) as source-level attributes into a holistic feature-driven approach that uses supervised setting for identifying PSM users– this is discussed in the second part of this chapter.

6.1 Hawkes Process for Understanding Differences Between Pathogenic Social Media Accounts and Normal Users

In this section, we aim to understand PSM accounts by (1) analyzing their behavior in terms of their posted URLs, and (2) estimate their influence by conducting experiments on a real-world dataset from Twitter. We deploy a mathematical technique known as "Hawkes process" Hawkes (1971) to quantify the impact of PSMs on normal users and the greater Web, by looking at their posted URLs on Twitter. Hawkes processes are special forms of point processes and have shown promising results in many problems that require modeling complicated event sequences where historical events have impact on future ones, including financial analysis Bacry *et al.* (2016), seismic analysis Daley and Vere-Jones (2007) and social network modeling Zhou *et al.* (2013) to name a few. This study uses an ISIS-related dataset from Twitter Alvari *et al.* (2018). The dataset contains an *action log* of users in the form of cascades of retweets. In this work, we consider URLs posted by two groups of users: (1) PSM accounts and (2) normal users. The URLs can belong to any platform including the major social media (e.g., Facebook.com), mainstream news (e.g., nytimes.com) and alternative news outlets (e.g., rt.com). For each group of users, we fit a multi-dimensional Hawkes processes model wherein each process correspond to a platform referenced in at least one tweet. Furthermore, every process can influence all the others including itself, which allows estimating the strength of connections between each of the social media platforms and news sources, in terms of how likely an event (i.e., the posted URL) can cause subsequent events in each of the groups. In other words, in this study we are interested to investigate if a given URL u_1 has influence on another URL u_2 (i.e., $u_1 \rightarrow u_2$) and thus can trigger subsequent events.

Main Findings. This work makes the following main observations:

- Among all platforms studied here, URLs shared from Facebook.com and alternative news media contributed the most to the dissemination of malicious information from PSM accounts. Simply put, they had the largest impact on making a message viral and causing the subsequent events.

- Posts that were tweeted by the PSM accounts and contained URLs from Facebook.com, demonstrated more influence on the subsequent retweets containing URLs from Youtube.com, in contrary to the other way around. This means that ultimately tweets with URLs from Facebook will high likely end up inducing more external impulse on YouTube than YouTube might have on Facebook.

- URLs posted by the normal users have nearly the same impact on the subsequent events regardless of the social media or news outlet used. This basically means that normal users do not often prefer specific social media or news sources over the others.

6.1.1 Dataset

We collect a dataset of 2.8M ISIS related tweets/retweets in Arabic between February 22, 2016 and May 27, 2016. The dataset contains different fields including user ID, retweet ID, hashtags, content, posting time. The dataset also contains user profile information including name, number of followers/followees, description, location, etc. The tweets were collected using different ISIS-related hashtags such as #stateoftheislamiccaliphate. In this dataset, about 600K tweets have at least one URL (i.e., event) referencing one of the social media platforms or news outlets. There are about 1.4M of paired URLs which we denote by $u_1 \rightarrow u_2$ and indicates a retweet (with the URL u_2) of the original tweet (with the URL u_1).

In this study, we are interested in investigating the impact of the URL u_1 on u_2. Accordingly, the dataset contains 35K cascades (i.e., sequences of events) of different sizes and duration, some of which contain paired URLs in the aforementioned form.

The statistics of the dataset are presented in Table 6.6. For labeling, we check through Twitter API to examine whether the users have been suspended (labeled as PSM) or they are still active (labeled as normal) Thomas *et al.* (2011). According to Table 6.6, 11% of the users in our dataset are PSMs and others are normal.

Social Media Platforms and News Outlets

Twitter deploys a URL shortener technique to leave more space for content and protect users from malicious sites [1] . To obtain the original URLs, we use a URL unshortening tool [2] to obtain the original links contained in the tweets in our dataset.

[1] https://help.twitter.com/en/using-twitter/url-shortener

[2] https://github.com/skevas/unshorten

Table 6.1: Description of the Dataset.

Name	**Value**	
# of Cascades	35K	
# of Tweets/Retweets	2.8M	
	PSM	**Normal**
# of Users	64,484	536,609
# of Single URLs	104,948	536,046
# of Paired URLs	200,892	1,123,434

We consider a number of major and well-known social media platforms including Twitter, Facebook, Instagram, Google and Youtube. About the dichotomy of mainstream and alternative media, it is notable to mention that most criteria for determining whether a news source counts as either of them, are based on a number of factors including but not limited to the content and whether or not it is corporate owned [3] . However, a key difference between these two sources of media comes from the fact that all of mainstream media is profit-oriented, in contrast to the alternative media. We further note that for the most part, mainstream media is considered as a more credible source than alternative media, although the reputation has been recently tainted by the fake news.

In this work, following the commonsense, we consider popular news outlets such as The New York Times, and The Wall Street Journal as mainstream and less popular ones as alternatives. In Table 6.2, we summarize the total number of paired URLs (i.e, $u_1 \rightarrow u_2$) in which the original URL (i.e., u_1) corresponds to each social media platform with at least one event in our dataset. We also summarize in Table 6.3, the total number of paired URLs whose original URL belongs to the mainstream

[3] https://smallbusiness.chron.com/mainstream-vs-alternative-media-21113.html

Table 6.2: Social Media Platform's Total Number of Paired URLs of the Form $u_1 \rightarrow u_2$ with at least One Event in the Dataset for the PSM and Normal Users.

Platform	**PSM**	**Normal**
Twitter	139,940	918,803
Facebook	878	4,017
Instagram	0	2,857
Google	163	132
Youtube	24,724	72,890

Table 6.3: News Sources' Total Paired URLs ($u_1 \rightarrow u_2$) with at least One Event in the Dataset for the PSM and Normal Users.

News Source	**PSM**	**Normal**
Mainstream	0	286
Alternatives	35,187	124,449

and alternative news sources. In Table 6.4, we see the break down of number of paired URLs for the PSM and normal users. We further demonstrate in Table 6.5 some examples of the mainstream and alternative news URLs occurrence used in this work.

Temporal Analysis

Here, we present the differences between the PSM accounts in our dataset with their counterparts, normal users through temporal analysis of their posted URLs.

In Figure 6.1, we depict the daily occurrence of the paired URLs over the span of 43 days for both PSM and normal users. Recall from the previous section that our dataset has a larger number of normal users and higher number of the posted URLs

Table 6.4: Total Number of the Paired URLs of the Form $u_1 \rightarrow u_2$ with at least One Event for PSM/Normal Users and For all Platforms.

	→ Twitter	→ Facebook	→ Instagram	→ Google	→ Youtube	→ Mainstream	→ Alternatives
Twitter →	109,354/766,617	598/3,843	229/2,461	120/382	11,992/59,889	90/688	17,557/84,923
Facebook →	655/3,108	4/41	3/9	2/1	87/281	0/1	127/576
Instagram →	0/2,362	0/11	0/25	0/2	0/161	0/2	0/294
Google →	134/74	0/0	0/1	0/0	12/53	0/0	17/4
Youtube →	14,004/56,545	132/312	23/211	22/32	6,799/7,529	13/48	3,731/8,213
Mainstream →	0/189	0/1	0/1	0/0	0/13	0/1	0/81
Alternatives →	21,047/95,641	145/767	45/318	59/64	3,862/9,199	26/122	10,003/18,338

Table 6.5: Examples of Mainstream and Alternative News.

Mainstream	Alternatives
https://www.nytimes.com	https://www.rt.com
https://www.reuters.com	https://www.arabi21.com
https://www.wsj.com	https://www.7adramout.net
https://www.nbcnews.com	https://www.addiyar.com
https://www.ft.com	https://zamnpress.com

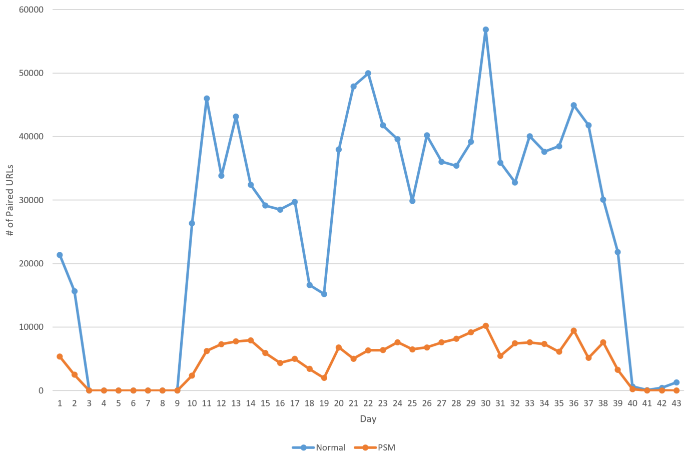


Figure 6.1: Number of Paired URLs Posted by the PSM and Normal users in our Dataset. Note that Number of Normal Users in our Dataset is Higher than the PSM Accounts.

compared to the PSM accounts. Therefore, it is reasonable to observe more activity from normal users than PSMs. For both groups of users, we observe a similar trend in occurrence of spikes and their durations. As it is seen, distinguishing between PSMs and normal users merely based on the occurrence of URLs and their patterns is not reliable. Therefore, we set out to conduct experiments using a more sophisticated statistical technique known as "Hawkes Process" in the next section.

6.1.2 Framework

In the previous section, we presented our data analysis to demonstrate differences between PSM accounts and normal users in terms of URLs they usually post on Twitter. We now set out to assess their impact via a well-known mathematical technique called “Hawkes process”.

Hawkes Processes

In many scenarios, one needs to deal with timestamped events such as the activity of users on a social network recorded in continuous time. An important task then is to estimate the influence of the nodes based on their timestamp patterns Gomez-Rodriguez *et al.* (2013). Point process is a principled framework for modeling such event data, where the dynamic of the point process can be captured by its conditional intensity function as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N(t+\Delta t) - N(t)|\mathcal{H}_t)}{\Delta t} = \frac{\mathbb{E}(dN(t)|\mathcal{H}_t)}{dt} \tag{6.1}$$

where $\mathbb{E}(dN(t)|\mathcal{H}_t)$ is the expectation of the number of events happend in the interval $(t, t+dt]$ given the historical observations $\mathcal{H}_t$ and $N(t)$ records the number of events before time t. Point process can be equivalently represented as a counting process $N = \{N(t)|t \in [0,T]\}$ over the time interval $[0,T]$.

The Hawkes process framework Hawkes (1971) has been used in many problems that require modeling complicated event sequences where historical events have impact on future ones. Examples include but are not limited to financial analysis Bacry *et al.* (2016), seismic analysis Daley and Vere-Jones (2007) and social network modeling Zhou *et al.* (2013). One-dimensional Hawkes process is a point process N_t with the following particular form of intensity function:

$$\lambda(t) = \mu + a \int_{-\infty}^{t} g(t-s)dN_s = \mu + a \sum_{i:t_i<t} g(t-t_i) \tag{6.2}$$

where $\mu > 0$ is the exogenous base intensity (i.e., background rate of events), t_i are the time of events in the point process before time t, and $g(t)$ is the decay kernel.

In this work, we use exponential kernel of the form $g(t) = we^{-wt}$, but adapting to the other positive forms is straightforward. The second part of the above formulation captures the self-exciting nature of the point processes– the occurrence of events in the past has a positive impact on the future ones. Given a sequence of events $\{t_i\}_{i=1}^{n}$ observed in $[0, T]$ and generated from the above intensity function, the log-likelihood function can be obtained as follows Zhou *et al.* (2013):

$$\mathcal{L} = \log \frac{\prod_{i=1}^{n} \lambda(t_i)}{\exp \int_0^T \lambda(t)dt} = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)dt \tag{6.3}$$

Here, we focus on multi-dimensional Hawkes processes which is defined by a U-dimensional point process $N_t^u, u = 1, \ldots, U$. In other words, we have U Hawkes processes coupled with each other– each Hawkes process correspond to one of the platforms and the influence between them is modeled using the mutually-exciting property of the multi-dimensional Hawkes processes. We formally define the following formulation to model the influence of different events on each other:

$$\lambda_u(t) = \mu_u + \sum_{i:t_i<t} a_{uu_i} g(t-t_i) \tag{6.4}$$

where $\mu_u \geq 0$ is the base intensity for the u-th Hawkes process. The coefficient $a_{uu_i} \geq 0$ captures the mutually-exciting property between the u-th and u_i-th processes. Larger value of a_{uu_i} shows that events in the u_i-th dimension are more likely to trigger an event in u-th dimension in future. More intuitively, an event on one point process can cause an impulse response on other processes, which increases the

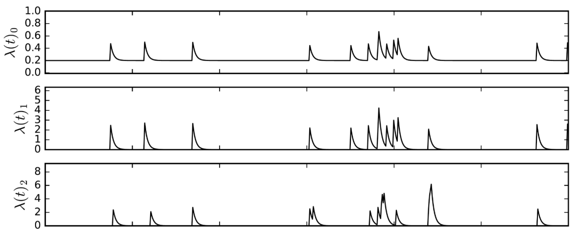


Figure 6.2: Illustration of the Hawkes Process. Events Induce Impulse on Other Processes and Cause Child Events. Background Event in e_0 Induces Impulse on Responses on Processes e_1 and e_2.

probability of an event occurring above the processes' background rates. We reiterate that in this study each URL is attributed to an event, i.e., if the URL u_1 triggers the URL u_2 (i.e., $u_1 \rightarrow u_2$), then $a_{u_2 u_1} \geq 0$

In Figure 6.2, we depict a multivariate example of three different streams of events, e_0, e_1 and e_2. As illustrated, e_0 is caused by the background rate $\lambda(t)_0$ and has an influence on itself and e_1. On the other hand, e_1 is caused by $\lambda(t)_1$ and has an influence on e_2. Simply put, a background event in e_0 induces impulse on responses on processes e_1 and e_2. Accordingly, the caused child event in e_1 leads to another child event in e_2.

We consider an infectivity matrix $\boldsymbol{A} = [a_{uu_i}] \in \mathbb{R}^{U \times U}$ which collects the self-triggering coefficients between Hawkes processes and measures the influence across events of different types Luo *et al.* (2015). Here, $U = 7$ is the number of processes (i.e., platforms) in our work. Each entry in this matrix indicates the strength of influence each platform has on other platforms. Our ultimate goal in this study is to estimate the infectivity matrix as it reflects the estimated influence of each platform on others. Next, we will provide the methodology that we follow to estimate the influence of the URLs on each other.

Methodology

We aim to assess the influence of the PSM accounts in our dataset via their posted URLs. We consider the URLs posted by two groups of users: (1) PSM accounts and (2) normal users. For both groups, we fit a Hawkes model with $K = 7$ point processes each for the seven categories of social media and news outlets discussed earlier. In each of the Hawkes models, every process is able to influence all the others including itself, which allows us to estimate the strength of connections between each of the seven categories for both groups of users, in terms of how likely an event (i.e., the posted URL) can cause subsequent events in each of the groups.

We use the ADM4 algorithm presented by Zhou *et al.* (2013) and follow the methodology presented by Zannettou *et al.* (2017) for fitting the Haweks processes for both PSM and normal users. ADM4 Zhou *et al.* (2013) is an efficient optimization that estimates the parameters $\boldsymbol{A}$ and $\boldsymbol{\mu}$ by maximizing the regularized log-likelihood $\mathcal{L}(\boldsymbol{A}, \boldsymbol{\mu})$:

$$\min_{\boldsymbol{A} \geq 0, \boldsymbol{\mu} \geq 0} -\mathcal{L}(\boldsymbol{A}, \boldsymbol{\mu}) + \lambda_1 ||\boldsymbol{A}||_* + \lambda_2 ||\boldsymbol{A}||_1 \tag{6.5}$$

where $\mathcal{L}(\boldsymbol{A}, \boldsymbol{\mu})$ can be obtained by substituting $\lambda_u(t)$ from Equation 6.4 into Equation 6.3. Also, $||\boldsymbol{A}||_*$ is the nuclear norm of matrix $\boldsymbol{A}$, and is defined as the sum of its singular value.

We consider two different sets of URLs posted by the PSM accounts and normal users by selecting URLs that have at least one event in Twitter (i.e., posted by a user). For each group, we construct a matrix $\boldsymbol{W} \in \mathbb{N}^{T \times U}$ with $U = 7$, whose entries are sequences of events (i.e., posted URLs) observed during a time period T. We note that each sequence of events is of the form $\mathcal{S} = \{(t_i, u_i)\}_{i=1}^{n_i}$ where n_i is the number of the events occurring at the u_i-th dimension (i.e., URLs posted containing one of the 7 platforms).

6.1.3 Experimental Results

Here, we conduct experiments to gauge the effectiveness of Hawkes process for moderling influence of PSMs.

Settings

In this work, we adopt the ADM4 algorithm Zhou *et al.* (2013) which implements parametric inference for Hawkes processes with an exponential kernel and a mix of Lasso and nuclear regularization. We initialize infectivity matrix $\boldsymbol{A}$, base intensities $\boldsymbol{\mu}$ and decays $\boldsymbol{\beta} \in \mathbb{R}$ randomly.

We further set the number of nodes $U = 7$ to reflect the 7 platforms used in this study. Level of penalization is set to $C = 1000$, and the ratio of Lasso-Nuclear regularization mixing parameter is set to 0.5. Finally, maximum number of iterations for solving the optimization is set to 50 and the tolerance of solving algorithm is set to $1e - 5$.

Results

We estimate infectivity matrix for both PSM and normal users by fitting the Hawkes model described earlier. In our study, this matrix characterizes the strength of the connections between the platforms and news sources. More specifically, each weight value represents the connection strength from one platform to another. In other words, each entry in this matrix can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first Zannettou *et al.* (2017). In Figure 6.3, we depict the estimated weights for all paired URLs for both PSM and normal users. Looking at the weights in both of the plots, we realize that greater weights belong to processes that have impact on Twitter, i.e. "$\rightarrow Twitter$". This implies that both of the groups in our Twitter dataset often post URLs that ultimately have greater impact on Twitter.

Overall, we observe the followings:

- URLs referencing all platforms and posted by the PSMs and regular users, mostly trigger URLs that contain the Twitter domain.
- Among all platforms studied here, URLs shared from Facebook.com and alternative news media contributed the most to the dissemination of malicious information from PSM accounts. In other words, they had largest impact on making a message viral and causing the subsequent events.
- Posts that were tweeted by the PSM accounts and contained URLs from Facebook.com, demonstrated more influence on the subsequent retweets containing URLs from Youtube.com, in contrary to the other way around. This means that ultimately tweets with URLs from Facebook will likely end up inducing external impulse on Youtube.com. In contrast, URLs posted by the normal users have nearly the same impact on the subsequent events regardless of the social media or news outlet used.

The above mentioned observations demonstrate the effectiveness of leveraging Hawkes process to quantify the impact of URLs posted by PSMs and regular users on the dissemination of content on Twitter. The observations we make here show that PSM accounts and regular users behave differently in terms of the URLs they post on Twitter, in that they have different tastes while disseminating URL links. Accordingly their impact on the subsequent events significantly differ from each other. Next, we leverage various characteristics of URLs shared by users into a holistic feature-driven approach for detecting PSM accounts.

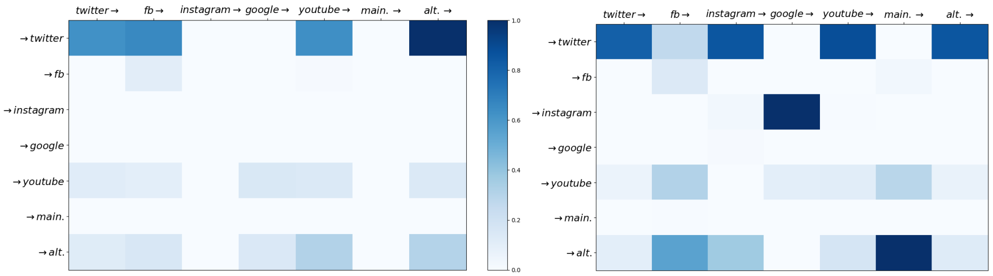


Figure 6.3: From Left to Right: Estimated Infectivty Matrices for all Paired URLs for PSMs and Normal Users. Among all URLs, those Shared from Facebook.com and Alternative News Media had the Largest Impact on Dissemination of Malicious Messages.

6.2 Feature-Driven Approach for Detecting PSM Accounts

Recent years have witnessed a surge of manipulation of public opinion and political events by different media outlets and malicious social media actors referred to as "Pathogenic Social Media" (PSM) accounts Alvari *et al.* (2018). The manipulation of opinion can take many forms from fake news Shao *et al.* (2017) to more subtle ones such as reinforcing specific aspects of text over others Baron (2006). It has been observed that media aggressively exert bias in the way they report the news to sway their reader's knowledge. On the other hand, PSM accounts are responsible for "agenda setting" and massive spread of misinformation Alvari *et al.* (2019b). Understanding misinformation from account-level perspective is thus a pressing problem.

Present Work. In this work, we aim to present an automatic feature-driven approach for detecting PSM accounts in social media. Inspired by the literature, we set out to assess PSMs from four broad perspectives: (1) causal and profile-related information, (2) source-related information (e.g., information linked via URLs) and (3) content-related information (e.g., tweets characteristics). For the causal and profile-related information, we investigate malicious signals using 1) causality analysis (i.e., if user is frequently a cause of viral cascades) Alvari *et al.* (2018) and 2) profile characteristics (e.g., number of followers, etc.) Kudugunta and Ferrara (2018) aspects of view. For the source-related information, we explore various properties that characterize the type of information being linked to URLs (e.g., URL address, content of the associated website, etc.) Baly *et al.* (2018); Phuong *et al.* (2014); Entman (1993); Morstatter *et al.* (2018); Kincaid *et al.* (1975). Finally, for the content-related information, we examine attributes from tweets (e.g., number of hashtags, certain hashtags, etc.) posted by users Kudugunta and Ferrara (2018). This work describes the results of research conducted by Arizona State University's Global Security Initiative and Center for Strategic Communication. Research support funding was provided by the US State Department Global Engagement Center.

Our corpus comprises three different real-world Twitter datasets, from Sweden, Latvia and United Kingdom (UK). These countries were selected to cover a range of population size and political history (former Soviet republic, neutral, founding member of NATO). In this study, we pose the following research questions and seek answers for them:

RQ1: *Does incorporating information from user activities and profile characteristics help in identifying PSM accounts in social media?*

RQ2: *What attributes could be exploited from URLs shared by users to determine whether or not they are PSMs?*

RQ3: *Could deploying tweet-level information enhance the performance of the PSM detection approach?*

To answer **RQ1**, we first investigate different profile characteristics that could indicate suspicious behavior. Next, We also examine whether or not users who make inauthentic information go viral, are more likely to be among PSM users. By exploring **RQ2**, we figure out which characteristics of URLs and their associated websites are useful in detecting PSM users in social media. By investigating **RQ3**, we aim to examine if adding a few content-related information on tweet-level could come in handy while identifying PSMs. Our answers to the above questions lead to a feature-driven approach that uses as little as three groups of user, source and content-related attributes to detect PSM accounts.

Key Ideas and Highlights. To summarize, this work makes the following main contributions:

- We present a feature-driven approach for detecting PSM accounts in social media. More specifically, we assess maliciousness from causal and profile-level, source-level and content-level aspects. Our casaul and profile-related information include signals in causal users (i.e., if user is frequently a cause of viral cascades) along with their profile characteristics (e.g., number of followers, etc.). For the source-related information, we explore different characteristics in URLs that users share and their associated websites (e.g., underlying themes, complexity of content, etc.). For the content-related information, we examine attributes from tweets (e.g., number of hashtags, certain hashtags, etc.) posted by users.

- We conduct a suite of experiments on three real-world Twitter datasets from different countries, using several classifiers. Using all of the attributes, we achieve average F1 scores of 0.81, 0.76 and 0.74 for Sweden, Latvian and U.K. datasets, respectively. Our observations suggest the effectiveness of the proposed method in identifying PSM accounts who are more likely to manipulate public opinion in social media.

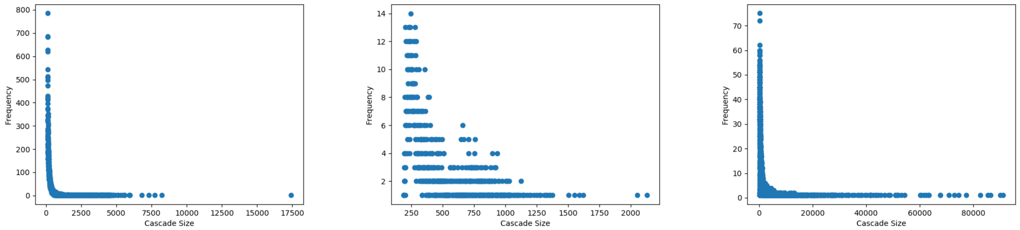


Figure 6.4: From Left to Right: Frequency Plots of Cascade Size for Sweden, Latvia and UK datasets.

6.3 Experimental Data

We collect three real-world Twitter datasets with different number of users and tweets/retweets from three countries, Sweden, Latvia and United Kingdom (UK). These countries were selected to cover a range of population size and political history (former Soviet republic, neutral, founding member of NATO). Description of the data is demonstrated in the Table 6.6. We use subsets of datasets from Nov 2017 to Nov 2018. Each dataset has different fields including user ID, retweet ID, hashtags, content, posting time as well as user profile information such as Twitter handles, number of followers/followees, description, location, protected, verified, etc. The tweets were collected using a predefined set of keywords and hashtags, and if they were geo-tagged in the country or user profile includes the country. We use subsets of the datasets with different number of cascades of different sizes and duration.

In our datasets, users may or may not have participated in viral cascades. We chose to use threshold $\theta = 20$ and take different number of viral cascades for each dataset with at least 20 tweets/retweets. We depict frequency plots of different cascade size for all datasets in Figure 6.4. For brevity, we only depict cascades size greater than 100 tweets/retweets.

Table 6.6: Description of the Datasets Used in this Work.

Dataset	**# Tweets/Retweets**	**# Labeled Users**		**# Viral Cascades**	**# URLs**
		Suspended	**Active**		
Sweden	780,250	16,010	48,030	12,174	160,702
Latvia	323,305	10,862	32,586	1,957	76,032
UK	254,915	4,553	13,659	21,429	41,332

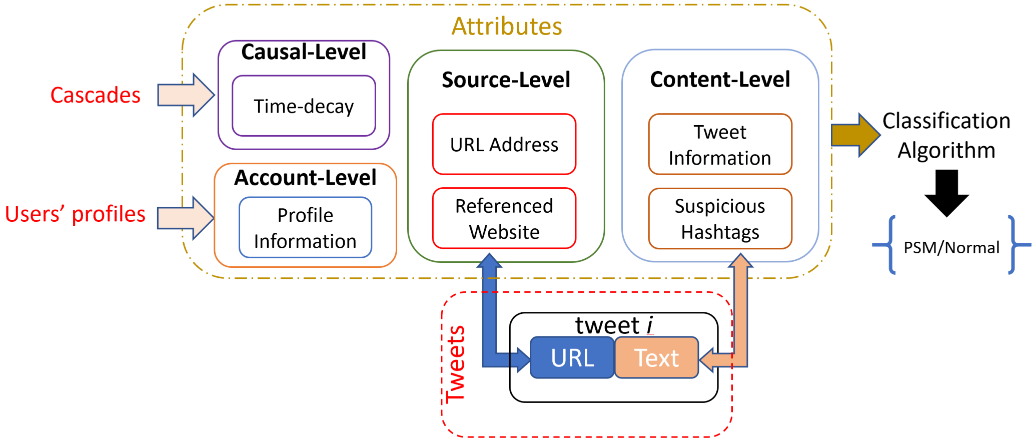


Figure 6.5: The Proposed Framework for Identifying PSM Users. It Incorporates Four Groups of Attributes Into a Classification Algorithm.

6.4 Identifying PSM Users

In this work, we take a machine learning approach (Figure 6.5) to answer the research questions posed earlier in the Introduction. More specifically, we incorporate different sets of malicious behavior indicators on causal-level, account-level, source-level and content-level to detect PSM users. In what follows, we describe each group of the attributes that will be ultimately utilized in a supervised setting to detect PSMs in social media.

6.4.1 Causal and Account-Level Attributes

We first set out to answer **RQ1** and understand attributes on the causal and account level that could be exploited in order to identify PSMs in social media.

Malicious Signals in Causal Users

Research has shown that user activity metrics are causally linked to viral cascades to the extent that malicious users who make harmful messages go viral are those with higher causality scores Alvari *et al.* (2018). Accordingly, we set out to investigate if incorporating causality scores in the form of attributes in a machine learning approach, can help identify users with higher malicious behavior in social media. More specifically, We leverage the causal inference introduced in Alvari *et al.* (2018) to compute a vector of causality attributes for each user in our dataset. Later, these causal-based attributes will be incorporated to our final vector of attributes that will be fed into a classifier. The causal inference takes as input cascades of tweets/retweets built from the dataset. We follow the convention of Goyal *et al.* (2010) and assume an *action log* $\mathcal{A}$ of the form *Actions(User,Action,Time)*, which contains tuples (i, a_i, t_i) indicating that user i has performed action a_i at time t_i. For ease of exposition, we slightly abuse the notation and use the tuple (i, m, t) to indicate that user i has posted (tweeted/retweeted) message m at time t. For a given message m we define a *cascade* of actions as $\mathcal{A}_m = \{(i, m', t) \in \mathcal{A} | m' = m\}$. User i is called m-participant if there exists t_i such that $(i, m, t_i) \in \mathcal{A}$. Users who have adopted a message in the early stage of its life span are called *key users* Alvari et al. (2018).

In this work we adopt the notion of *prima facie causes* which is at the core of Suppes' theory of probabilistic causation Suppes (1970) and utilize the causality metrics that are built on this theory. According to this theory, *a certain event to be* *recognized as a cause, must occur before the effect and must lead to an increase of the likelihood of observing the effect.* Accordingly, prima facie causal users for a given viral cascade, are key users who help make the cascade go viral. Finally, according to Alvari *et al.* (2018), we define 4 causal-based attributes for each user and add them to the final representative feature vector for the given user.

Malicious Signals in Profile Characteristics

Having defined our causality-based attributes, we now describe our next set of user-based features. Specifically, for each user, we collect account-level features and add them to the final feature vector for that user. We follow the work of Kudugunta and Ferrara (2018) and compute the following 10 features from users' profiles: *Statuses Count, Followers Count, Friends Count, Favorites Count, Listed Count, Default Profile, Geo Enables, Profile Uses Background Image, Verified, Protected.* Prior research has shown promising results using this small set of features Ferrara *et al.* (2016) with far less number of features than the established bot detection approach, namely, Botometer which uses over 1,500 features. Accordingly, we extend the final feature vector representation of each user by adding these 10 features.

6.4.2 Source-Level Attributes

Here, we seek an answer to **RQ2** and examine malicious behavior from the source-level perspective. Previous research has demonstrated the differences between normal and PSM users in terms of their shared URLs Alvari and Shakarian (2019) and their impact on creating subsequent events in future. We thus follow the same procedure described in the first part of this chapter and compute the infectivity matrices for the Sweden, Latvia and UK datasets. The matrices are depicted in Fig 6.6. Similarly, we observe clear distinctions between PSM and normal users' behaviors in terms of

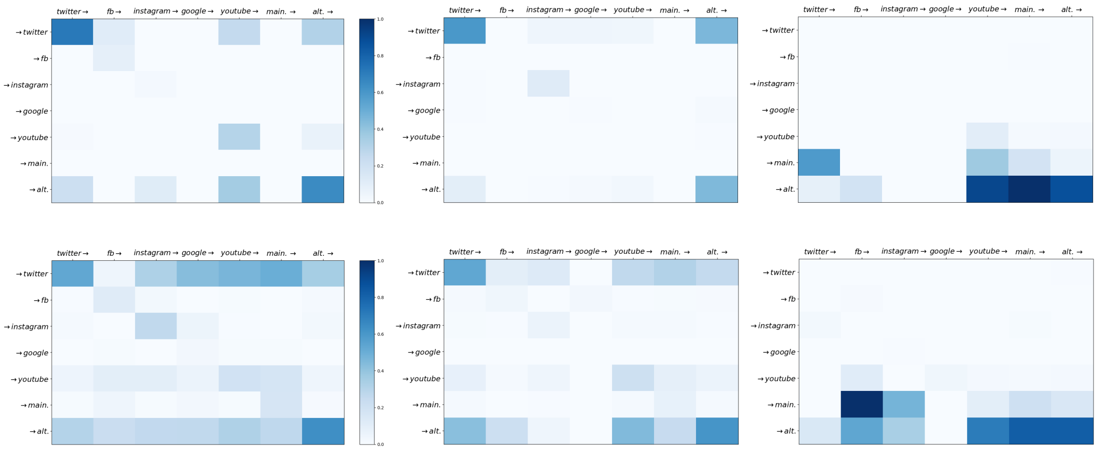


Figure 6.6: (Top) From Left to Right: Estimated Infectivity Matrices for PSM Accounts in Sweden, Latvia and UK Datasets. (Bottom) From Left to Right: Estimated Infectivity Matrices for Normal Users in Sweden, Latvia and UK Datasets.

their shared URLs. Specifically, URLs shared by PSM accounts more likely trigger subsequent events (i.e., future adoptions of URLs) when coming from alternative news sources. This is in contrast to the URLs shared by normal users which either trigger subsequent events on mainstream news outlets or social media platforms. Following our observations, we now take URLs posted by users as source-related information that could be used in our PSM user detection approach. Specifically, we set out to understand several characteristics of each URL from two broad perspectives: (1) URL address and (2) content collected from the website it has referenced.

URL Address

Far-right and pro-Russian URLs Here, we examine if the given URL refers to any of the following far-right websites: *https://voiceofeurope.com/*, *https://newsvoice.se/*, *https://nyadagbladet.se/*, *https://www.friatider.se/* or the pro-russian website *https://ok.ru/*. We further note that each user may have posted multiple URLs posted in our data.

To account for that, we compute the average of these attribute values for each user. Ultimately, this list leads to a vector of 5 values for each URL shared by each user in our dataset. We leave examining other malicious websites to future work.

Domain Extensions Previous research on assessing news articles credibility suggests looking at their URLs Baly *et al.* (2018) to examine if they contain features such as whether a website contains the *http* or *https* prefixes, or *.gov*, *.co* and *.com* domain extensions. Likewise, we investigate if the URLs in our dataset contain any of these 5 features by counting the number of times each URL triggers one of these attributes and taking the average if user has shared multiple of such URLs. This additional attribute vector will be added to the final attribute vector for each user.

Referenced Website Content

Topics We further investigate whether or not incorporating the underlying topics or themes learned from the text of the websites, could help us to build a more accurate approach to identify malicious activity. More specifically, we first set out to extract the content from each URL shared by users. To learn the topics, We follow the procedure described in Phuong *et al.* (2014) and train Latent Drichlet Allocation (LDA) Blei *et al.* (2003) on the crawled contents of the websites associated with each URL in the training set. This way, we obtain a fine-grained categorization of the URLs by identifying their most representative topics as opposed to a coarser-grained approach that uses predefined categories (e.g., sports, etc.). Using LDA also allows for uncovering the hidden thematic structure in the training data. Furthermore, we rely on the document-topic distribution given by the LDA (here each document is seen as a mixture of topics) to distinguish normal users from highly biased users. After training LDA, We treat each new document and measure their association with each of the K topics discovered by LDA. We empirically found $K = 25$ to work well in our dataset. Thus, each document is now treated as a vector of 25 probabilistic values provided by LDA's document-topic distribution- this feature space will be added to the final set of the features built so far. Finally, note that for users with more than one URL, we take the average of different probabilistic feature vectors.

Has Quote Social science research has shown that news agencies seek to make a piece of information more noticeable, meaningful, and memorable to the audience Entman (1993). This increases the chance of shifting believes and perceptions. One way to increase salience of a piece of information is emphasizing it by selecting particular facts, concepts and quotes that match the targeted goals Entman (1993); Scheufele and Tewksbury (2006); DellaVigna and Kaplan (2007). We thus check the existence of quotes within the referenced website content as an indicator of malicious behavior– this results in a single binary feature. Each user may post more than one URL. To account for this, We take the average values of this feature for each user. We observe that the PSM users' mean scores for this feature are 0.04 (Sweden), 0.05 (Latvia) and 0.04 (UK). Normal users have mean scores of 0.05 (Sweden), 0.05 (Latvia), and 0.03 (UK). We also deploy two-tailed two-sample t-test with the null hypothesis that value of this feature is not significantly different between normal and PSM accounts. Table. 6.7 summarizes the p-values for this test with significance level $\alpha = 0.01$. Results show that the null hypothesis could not be rejected. However, we still include this feature to see whether or not it helps in identifying PSMs in practice.

Complexity Research has shown that complexity of the given text could be different for malicious and normal users Morstatter *et al.* (2018). We thus use complexity feature to see whether or not it aids the classifier in finding users who create and share malicious content. We follow the same approach as in Morstatter *et al.* (2018) and approximate the complexity of reference website content as follows:

$$\text{complexity} = \frac{\text{number of unique part-of-speech tags}}{\text{number of words in the text}} \tag{6.6}$$

The higher this score is, the more complex the given context is. Surprisingly, our initial analysis show that mean of complexity score of website content by PSMs are 0.53 (Sweden), 0.54 (Latvia) and 0.51 (UK) while mean of complexity score of website contents shared by normal users are 0.46 (Sweden), 0.51 (Latvia), and 0.48 (UK). This shows contents shared by PSMs have higher complexity than those shared by normal users. We also deploy one-tail two-sample t-test with the null hypothesis that content of URLs shared by normal are more complex than those shared by PSMs. Table. 6.7 summarizes the p-values showing that the null hypothesis was rejected at significance level $\alpha = 0.01$. This indicates that content of websites referenced by PSM users are more complex than those shared by normal users.

Readability According to Horne and Adali (2017), readability of a given context can affect engagement of the individuals with the given piece of information. Therefore, readability of the referenced website content is another important feature which could be useful in distinguishing PSMs and normal users. We hypothesize that PSM users may share information with higher readability to increase the chance of transferring the concept and creating malicious content. We use Flesch-Kincaid reading-ease test Kincaid *et al.* (1975) on the text of the provided URLs. The mean readability scores are 61.16 (Sweden), 62.98 (Latvia), 59.08 (UK) for PSMs and 55.44 (Sweden), 56.79 (Latvia), 55.35 (UK) for other normal users. The higher the score is, the more readable the text is. We also deploy one-tail two-sample t-test with the null hypothesis that content of URLs shared by normal users are more readable than those shared

Table 6.7: Results of p-Values at Significance Level $\alpha = 0.01$. The Null Hypotheses for Complexity and Readability Tests are Refuted.

Feature	**Sweden**	**Latvia**	**UK**
Has Quote	0.29	0.36	0.32
Complexity	4.95e-50	3.23e-07	6.12e-08
Readability	5.56e-27	4.2e-03	1.9e-14

by PSMs. Table. 6.7 summarizes the p-values indicating that the null hypothesis was rejected at significance level $\alpha = 0.01$.

These results show that the content of URLs shared by PSM accounts are more complex yet more readable than those shared by normal users. Therefore, these two features, complexity and readability, could be a good indicator to distinguish between normal and PSMs.

Unigrams/Bigrams We use TF-IDF weighting for extracted word-level unigrams and bigrams. This feature gives us both importance of a term in the given context (i.e., term frequency) and term's importance considering the whole corpus. We remove stop words and select top 20 frequent unigrams/bigrams as the final set of features for this group. Using TF-IDF weighting helps to identify piece of information that is focusing on aspects not emphasized by others. For brevity, we only demonstrate top bigrams in Table 6.8.

Domain Expertise The presence of signal words (e.g., specific frames or keywords) could be indicator of existence of malicious behavior in the text. In this work, we hired human coders and trained them based on our codebook [4] in order to provide

[4] A codebook is survey research approach to provide a guide for framing categories and coding responses to the the categories definitions.

Table 6.8: Top Selected Bigrams for each Country.

Data	**Bigrams**
Sweden	asylum seeker, birthright citizenship, court justice, European commission, European Union (EU), European parliament, kill people, migrant caravan, national security, Russian military, school shooting, sexually assault, united nations, white supremacist, police officer
Latvia	Baltic exchange, Baltic security, battlefield revolution, cyber security, depository Estonia, Estonia Latvia, European parliament, European commission, European Union (EU), human rights, Latvian government, nasdaq Baltic, national security, Saeima election, Vladimir Putin
UK	court appeal, cosmic diplomacy, defence police, depression anxiety, diplomacy ambiguity, European Union (EU), human rights, Jewish community, police officer, police federation, political party, rebel medium, sexually liberate, support group, would attacker

signal words that can help identify suspicious behavior. We use the following framing categories: *Anti-immigrant, Crime rampant, Government, Anti-EU/NATO, Russia-ally, Crimea, Discrimination, Fascism.* For each country and each category, we have a list of corresponding keywords. We have illustrated examples of the keywords used in this study in Table 6.9.

6.4.3 Content-Level Attributes

In this section, we aim to understand **RQ3** by incorporating a few more attributes from the content-level information that could be used to enhance the performance of the PSM user detection. For the content-level information, we only rely on the tweets posted by each user in our dataset.

Table 6.9: Examples of the Keywords Used in this Study.

Data	**Keywords**
Sweden	no-go zones, violence overwhelmed, police negligence, Nato obsolete, bilateral cooperation, blighted areas, increase reported rapes, close police station, EU hypocrisy, anti-immigrant, fatal shootings, badly Sweden, Nato airstrikes
Latvia	Brussels silent, norms international law, bureaucrats, lack trust EU, based universal principles, Russia borders, anti Nato, purely political, European bureaucrats, silence Brussels Washington, rampant, harsh statements concerning, values Brussels silent
UK	Brexit, Theresa May, stop Brexit, hard Brexit, post Brexit, leave, referendum, Brexitshambles

Malicious Signals in Tweet-Level Information

We use the following 6 attributes extracted from each tweet Kudugunta and Ferrara (2018): *retweet count, reply count, favorite count, number of hashtags, number of URLs, number of mentions*. If the user has tweeted more than once, we take the average of these features.

Malicious Signals in Suspicious Hashtags

We further investigate if the given tweet aims to push propaganda using any of the following suspicious hashtags identified by our human coders. For Sweden, we use *#Swedistan, #Swexit, #sd (far right group), #SoldiersofOdin, #NOGOZones*. For Latvia, we use *#RussiaCountryFake, #BrexitChaos, #BrexitVote, #Soviet, #RussiaAttacksUkraine*. For UK, we use *#StopBrexit, #BrexitBetrayal, #StopBrexitSaveBritain, #StandUp4Brexit, #LeaveEU*. Similar to the previous attributes, for the users who have posted more than one tweet with these hashtags, we compute the average of the corresponding values. We leave examining other suspicious hashtags to future work.

6.4.4 Feature-Driven Approach

Having described the attributes (Table 6.10) used in this work, we now feed them into a supervised classification algorithm to detect PSM users (Figure 6.5). In more details, we feed the profile information and tweets into the different components of the proposed approach. For the causal and account-related information, we require both of the profile characteristics and tweets. We need tweets to build viral cascades and finally compute causality scores for different users. Each cascade contains tuples (i, m, t) indicating that user i has posted (tweeted/retweeted) the corresponding message m at time t. Given the cascades, causality features are computed for each user i based on her activity log in our dataset. For the source-level information, we only need to extract URLs from tweets. These URLs are either directly used to compute attributes or to collect the content from the websites to which they have referenced. For the content-related information, we only need tweets in order to compute the content-level attributes. Finally, for each user, we fuse all attributes into a feature vector representation and feed them into a classifier.

6.5 Experiments

In this section, we conduct experiments on three real-world Twitter datasets to gauge the effectiveness of the proposed approach. In particular, we compare the results of several classifiers and baseline methods. Note for all methods, we only report results when their best settings are used.

- **Ensemble Classifiers**

Table 6.10: Different Groups of Features Used in this Work.

	Feature	**Definition**	**# Feat.**
Causal	Time-Decay	Attributes computed using causality based metrics	4
Account	Profile-based	*Statuses Count, Followers Count, Friends Count, Favorites Count, Listed Count, Default Profile, Geo Enables, Profile Uses Background Image, Verified, Protected*	10
Source	Websites	Presence of far-right and pro-Russian websites	5
	Domains	Existence of *http* or *https* prefixes, or *.gov*, *.co* and *.com* domain extensions	5
	Topics	Features computed by comparing the listing against the learned topic distribution	25
	Has Quote	Single binary feature that shows whether the content of shared URLs contains quote or not.	1
	Complexity	Complexity of content of shared URLs by users.	1
	Readability	Readability of content of shared URLs by users.	1
	Unigram	TF-IDF scores of highly frequent word-level unigrams extracted from content of URLs shared by users.	20
	Bigram	TF-IDF scores of highly frequent word-level bigrams extracted from content of URLs shared by users.	20
	Expertise	Presence of signal keywords provided by our coders	8
Content	Tweet-based	*retweet count, reply count, favorite count, number of hashtags, number of URLs, number of mentions*	6
	Hashtags	Presence of suspicious hashtags	5

  - **Gradient Boosting Decision Tree (GBDT)** We train a Gradient Boosting Decision Tree classifier using the described features. We set the number of estimators as 200. Learning rate was set to the default value of 0.1.
  - **Random Forest (RF)** We train a Random Forest classifier using the features described. We use 200 estimators and entropy as the criterion.
  - **AdaBoost** We train an AdaBoost classifier using the described features. The number of estimators was set to 200 and we also set the learning rate to 0.01.

- **Discriminative Classifiers**
  - **Logistic Regression (LR)** We train a Logistic Regression using $l2$ penalty. We also set the parameter $C = 1$ (the inverse of regularization strength) and tolerance for stopping criteria to 0.01.
  - **Decision Tree (DT)** We train a Decision Tree classifier using the features. We did not tune any specific parameter.
  - **Support Vector Machines (SVM)** We use a linear SVM using the attributes described in the previous section. We set the tolerance for stopping criteria to 0.001 and the penalty parameter $C = 1$.

- **Generative Classifiers**
  - **Naive Bayes (NB)** We train a Multinomial Naive Bayes which has shown promising results for text classification problems Manning *et al.* (2008). We did not tune any specific parameter for this classifier

- **Baselines**

- **Long Short-Term Memory (LSTM) Kudugunta and Ferrara (2018)** The word-level LSTM approach here is similar to the deep neural network models used for sequential word predictions. We adapt the neural network to a sequence classification problem where the inputs are the vector of words in each tweet and the output is the predicted label of the tweet. We first use the word2vec Mikolov *et al.* (2013) embeddings which are trained jointly with the classification model. We use a single LSTM layer of 50 units on the textual content, followed by the loss layer which computes the cross entropy loss used to optimize the model.N

- **Account-Level (AL) + Random Forest Kudugunta and Ferrara (2018)** This approach uses the following features of the user profiles: *Statuses Count, Followers Count, Friends Count, Favorites Count, Listed Count, Default Profile, Geo Enables, Profile Uses Background Image, Verified, Protected.* We chose this method over Botometer Varol *et al.* (2017a) as it achieved comparable results with far less number of features (Varol *et al.* (2017a) uses over 1,500 features)(see also Ferrara *et al.* (2016)). According to Kudugunta and Ferrara (2018), we report the best results when Random Forest (RF) is used.

- **Tweet-Level (TL) + Random Forest Kudugunta and Ferrara (2018).** Similar to the previous baseline, this method uses only a handful of features extracted from tweets: *retweet count, reply count, favorite count, number of hashtags, number of URLs, number of mentions.* Likewise, we use RF as the classification algorithm.

Table 6.11: Performance Comparison on Different Datasets using all Features.

Classifier	**Sweden**		**Latvia**		**UK**	
	F1-macro	**F1-score**	**F1-macro**	**F1-score**	**F1-macro**	**F1-score**
GBDT	**0.80**	**0.81**	**0.76**	**0.76**	**0.73**	**0.74**
RF	0.79	0.79	0.75	0.75	0.70	0.71
AdaBoost	0.78	0.79	0.73	0.74	0.69	0.70
LR	0.75	0.75	0.74	0.74	0.71	0.72
DT	0.69	0.69	0.71	0.71	0.69	0.69
SVM	0.73	0.74	0.73	0.70	0.72	0.70
NB	0.71	0.71	0.65	0.67	0.66	0.67
LSTM	0.60	0.62	0.58	0.65	0.36	0.43
AL (RF)	0.64	0.64	0.63	0.64	0.64	0.65
TL (RF)	0.50	0.51	0.50	0.51	0.49	0.50

6.5.1 Results and Discussion

All experiments were implemented in Python 2.7x and run on a machine equipped with an Intel(R) Xeon(R) CPU of 3.50 GHz with 200 GB of RAM running Linux. We use tenfold cross-validation follows. We first divide the entire set of training instances into 10 different sets of equal sizes. Each time, we hold one set out for validation. This procedure is performed for all approaches and all datasets for the sake of fair comparison. Finally, we report the average of 10 different runs, using F1-macro and F1-score (only for PSM users) evaluation metrics and all features in Table 6.11.

Performance Evaluation

For any approach that requires special tuning of parameters, we conducted grid search to choose the best set of parameters. Also, for LSTM, we preprocess the individual tweets in line with the steps mentioned in Soliman *et al.* (2017). We use word vectors of dimensions 100 and deploy the skip-gram technique for obtaining the word vectors where the input is the target word, while the outputs are the words surrounding the target words. To model the tweet content in a manner that uses it to predict whether an account is biased or not, we used LSTM models Hochreiter and Schmidhuber (1997). For the LSTM architecture, we use the first 20 words in the tokenized text of each tweet and use padding in situations where the number of tokens in a tweet are less than 20. We use 30 units in the LSTM architecture (many to one). The output of the LSTM layer is fed to a dense layer of 32 units with ReLU activations. We add dropout regularization following this layer to avoid overfitting and the output is then fed to a dense layer which outputs the category of the tweets.

Observations. Overall, we make the following observations:

- In general, results from different classifiers compared to the baselines demonstrate the effectiveness of the described attributes in identifying PSM users in social media. Thus, the answers to the research questions **RQ1**–**RQ3** are all positive, i.e., we could exploit attributes from user activities and profile characteristics, source and content-related information for identifying PSM users in social media. More specifically, for **RQ1**, we investigate different profile characteristics that could indicate suspicious behavior. We also examine whether or not users who make inauthentic information go viral, are more likely to be among PSM users. By answering **RQ2**, we figure out which characteristics of URLs and their associated websites are useful in detecting PSM users in social media. By investigating **RQ3**, we examine if adding a few content-related information on tweet-level could come in handy while identifying PSMs. Our answers to the above questions lead to a feature-driven approach that uses as little as three groups of user, source and content-related attributes to detect PSM accounts.

- Ensemble classifiers using the described features, outperform all other classifiers and baselines. Amongst the ensemble classifiers, Gradient Boosting Decision Trees classifier achieves the best results in terms of both F1-macro and F1-score metrics.

- Amongst the discriminative classifiers, linear Support Vector Machines classifier marginally beats Logistic Regression. Decision Tree classifier achieves the worst results in this category.

- Overall, Decision Tree and Naive Bayes classifiers achieve the worst performance among all classifiers.

- For LSTM, we achieve slightly poor performance than the logistic regression classifier. One reason behind the poor performance of the classifier is the lack of trained word embeddings suited to our dataset. Also, the poor performance might suggest that the sequential nature of the texts might not be very helpful for the task of PSM users detection.

- Overall, results on Sweden data demonstrate better performances achieved using the attributes. One reason behind this might be the size of data and higher number of PSMs in Sweden data compared to others. This could also indicate that PSMs in Latvia and UK data are more sophisticated.

Feature Importance Analysis

We further conduct feature import analysis to investigate what feature group contributes the most to the performance of the proposed approach. More specifically, we use GBDT and perform different 10-fold cross validations using each feature group. We report the F1-score results in Table 6.12. According to our observations, we conclude that the most significant and less significant feature groups are *source-related* and *content-related* attributes, respectively. We also perform feature ablation test by taking out a single feature group at a time from the rest. We observe that eliminating content-related attributes has the least impact on the performance, while taking out source-related attributes deteriorates the performance drastically. One final note though is, despite the effectiveness of the attributes from the user-level information, they may not be always available or we may not always know the suspicious sources beforehand for the task at hand. This further demonstrates the effectiveness of the causal-related features extracted from users' activities for identifying PSM users and thus confirms the observations in Chapters 4 and 5.

6.6 Conclusion

In the first part of this chapter, we provided analyses on a real-world ISIS TWitter data to demonstrate differences between PSM accounts and normal users. In particular, we leverage a statistical technique known as Hawkes Process for modeling the differences between users while disseminating content on the Web. We use URLs posted by two groups of users, PSMs and normal users, on major social media and mainstream and alternative news outlets. Overall, our findings indicate that the URLs posted by the PSM accounts have the largest impact if contained either Facebook.com or alternative news media. In contrast, their counterparts, i.e., normal

Table 6.12: Feature Importance on Different Datasets.

Feature	Sweden	Latvia	UK
Causal	0.64	0.62	0.61
Account	0.62	0.61	0.59
Content	0.45	0.43	0.40
Source	0.73	0.70	0.68
All \ Source	0.71	0.65	0.63
All \ Causal	0.73	0.67	0.62
All \ Account	0.76	0.70	0.69
All \ Content	0.79	0.73	0.72
All	0.81	0.76	0.74

users, often post URLs that have nearly the same impact on the Web, no matter what social media or news outlet they use.

In the second part of this chapter, we present an automatic feature-driven approach for identifying PSM accounts in pro-Russian social media. In particular, we assess the malicious behavior from four broad perspectives: (1) causal, (2) account, (3) source and (4) content-related information. For the first two groups, we investigate malicious signals using 1) causality analysis (i.e., if user is frequently a cause of viral cascades) and 2) profile characteristics (e.g., number of followers, etc.) aspects of view. For the source-related information, we explore various properties that characterize the type of information being linked to URLs (e.g., URL address, content of the associated website, etc.). Finally, for the content-related information, we examine attributes from tweets (e.g., number of hashtags, certain hashtags, etc.).

Chapter 7

CONCLUSION AND FUTURE WORK

7.1 Summary

Online platforms such as online social networks, microblogging websites and other Web platforms, have become widespread tools exploited by various malicious actors that orchestrate large-scale and societal-significant threats ranging from online human trafficking to misinformation spread. To better understand the behavior and impact of the malicious actors and counter their activity, we need certain capabilities to reduce their threats. Due to the large volume of information published online and because of the limited manpower, the burden usually falls to algorithms that are designed to automatically identifying these bad actors. However, this is a subtle task facing online platforms due to several challenges: (1) malicious users have strong incentives to disguise themselves as normal users (e.g., intentional misspellings, camouflaging, etc.), (2) malicious users are high likely to be key users in making harmful messages go viral and thus need to be detected at their early life span to stop their threats from reaching a vast audience, and (3) available data for training automatic approaches for detecting malicious users, are usually either highly imbalanced (i.e., higher number of normal users than malicious users) or comprise insufficient labeled data. This dissertation investigates the propagation of online malicious information from two broad perspectives: (1) *content* posted by malicious users and (2) *malicious information cascades* formed by malicious users and by resharing mechanisms in social media. For the former, the problem of online human trafficking and potential countermeasures to combat them are studied. We present non-parametric and semi-supervised learning algorithms for detecting online human trafficking. For the latter, we study and understand "Pathogenic Social Media" (PSM) accounts who are likely to be key users in making malicious campaigns. Various machine learning-based algorithms are then presented to detect PSM accounts.

In sum, this dissertation makes the following contributions in identification of human trafficking and PSM accounts as two forms of malicious activities:

- We use the user-generated content posted on Backpage and present semi-supervised algorithms to identify high likely human trafficking-related posts. The models were trained on both of the *labeled* and *unlabeled* data from Backpage and the results were further verified by our law enforcement experts.
- We investigate the extent to which various forms of user-generated data could contribute to identifying PSM accounts on Twitter. In particular, we examine (1) resharing activities and user profile information (user-level), (2) URLs and contents of their associated websites posted by users (source-level), and (3) tweets' textual characteristics (content-level). We present causality-based, semi-supervised causality-based and feature-driven approaches for detecting PSM accounts on Twitter.
    - For causality-based algorithms, we leverage information from resharing activities in the form of cascade structure on Twitter and present *time-decay* causal metrics for early identification of PSMs, based on the Suppes' probabilistic causal theorem Suppes (1970). We further investigate the role of community structure in early detection of PSMs by demonstrating that users within a community establish stronger causal relationships compared to the rest.

  - For semi-supervised causality-based approaches, we frame the problem of detecting PSM accounts with far less number of PSMs than the normal users, as an optimization problem and present a Laplacian semi-supervised causal inference SemiPsm for solving it. The unlabeled data are utilized via manifold regularization. Manifold regularization used in the resultant optimization formulation is built upon causality-based features created on a notion of *Suppes' theory of probabilistic causation.*
  - For feature-driven approach, we further fuse different attributes of user activities and profiles, source and content levels into one holistic approach. More specifically, we assess the extent to which causal and account levels, source-level and content-level attributes contribute to identification of PSM accounts. Our causal and account attributes investigate signals in causal users along with their profile information. For the source-level attributes, we explore different characteristics in URLs content that users share (e.g., underlying themes, complexity of text, etc.). For the content-level attributes, we examine attributes from tweets posted by users. Our observations suggest the effectiveness of the proposed method in identifying PSM accounts in real-world Twitter data.

7.2 Future Work

This dissertation studies the propagation of malicious information online from various perspectives, but only touch upon the tip of the iceberg of this fertile research area. Below, we present some of promising research directions that require further explorations:

- **Detecting Human Trafficking**: In this dissertation, we only leverage textual information (e.g., title, description, timestamp, poster's age, etc.) from user-generated data on Backpage. In future, we plan to utilize other available yet prominent forms of user data such as images and videos. Another potential future direction is to leverage data from social media such as Instagram and Twitter to present more sophisticated algorithms for detecting human trafficking. This way, we could identify orchestrated human trafficking related crimes and their associated rings. Furthermore, we plan to replicate the study by integrating more interesting features especially those supported by the criminology literature. Also, since hand-labeling unlabeled examples is expensive, an interesting research direction would be to deploy active learning to enable iterative supervised learning to actively query the user for labels. We also note that real-world data is often more imbalanced compared to our data, and the reason is that number of negative samples usually outweigh positive ones. We would thus like to apply the proposed framework on a more realistic dataset which contains much less suspicious posts than normal posts.

- **Identifying Pathogenic Social Media Accounts**: This dissertation utilizes several aspects of user-generated data including causality-based attributes, profile information and contents from URLs and tweets for detecting PSM accounts in real-world related Twitter datasets. Although Chapter 4 presents causality-based algorithms for early detection of PSM accounts, to support more real-time detection criteria, we plan to incorporate time-related attributes such as those extracted from point process, time-series and LSTM. Our future plans also include investigating other forms of causality inferences such as Granger causality Didelez (2008) and other regularization terms to seek if we can further improve the classification performance. Another direction for future work would be to present more accurate approaches to reduce false positives which have high costs for social media. Also we would like to present methods that can distinguish between different types of PSMs.

BIBLIOGRAPHY

"Trafficking victims protection act of 2000", URL https://www.state.gov/j/tip/laws/61124.htm (2000).

"UNODC on human trafficking and migrant smuggling", URL https://www.unodc.org/unodc/en/human-trafficking/ (2011).

"Trafficking in persons report", URL https://www.state.gov/j/tip/rls/tiprpt/2015/ (2015).

"Average height to weight chart - babies to teenagers", URL https://www.disabled-world.com/calculators-charts/height-weight-teens.php (2017).

Alvari, H., G. Beigi, S. Sarkar, S. W. Ruston, S. R. Corman and P. Shakarian, "A feature-driven approach for identifying pathogenic social media accounts", IEEE Conference on Data Intelligence and Security (2020).

Alvari, H., A. Hajibagheri, G. Sukthankar and K. Lakkaraju, "Identifying community structures in dynamic networks", Social Network Analysis and Mining (SNAM) (2016a).

Alvari, H., S. Hashemi and A. Hamzeh, "Detecting overlapping communities in social networks by game theory and structural equivalence concept", in "International Conference on Artificial Intelligence and Computational Intelligence", pp. 620–630 (Springer, 2011).

Alvari, H., S. Sarkar and P. Shakarian, "Detection of violent extremists in social media", IEEE Conference on Data Intelligence and Security (2019a).

Alvari, H., E. Shaabani, S. Sarkar, G. Beigi and P. Shakarian, "Less is more: Semi-supervised causal inference for detecting pathogenic users in social media", in "Companion Proceedings of The 2019 World Wide Web Conference", pp. 154–161 (ACM, 2019b).

Alvari, H., E. Shaabani and P. Shakarian, "Early identification of pathogenic social media accounts", IEEE Intelligent and Security Informatics (2018).

Alvari, H. and P. Shakarian, "Causal inference for early detection of pathogenic social media accounts", CoRR **abs/1806.09787** (2018).

Alvari, H. and P. Shakarian, "Hawkes process for understanding the influence of pathogenic social media accounts", in "2019 2nd International Conference on Data Intelligence and Security (ICDIS)", pp. 36–42 (2019).

Alvari, H., P. Shakarian and J. Snyder, "A non-parametric learning approach to identify online human trafficking", Intelligence and Security Informatics (ISI), 2016 IEEE Conference on pp. 133—-138 (2016b).

Alvari, H., P. Shakarian and J. K. Snyder, “Semi-supervised learning for detecting human trafficking”, Security Informatics **6**, 1, 1 (2017).

Backstrom, L. and J. Leskovec, “Supervised random walks: predicting and recommending links in social networks”, in “Proceedings of the fourth ACM international conference on Web search and data mining”, pp. 635–644 (ACM, 2011).

Bacry, E., T. Jaisson and J.-F. Muzy, “Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics”, Quantitative Finance **16**, 8, 1179–1201 (2016).

Badjatiya, P., S. Gupta, M. Gupta and V. Varma, “Deep learning for hate speech detection in tweets”, in “Proceedings of WWW”, (2017).

Baly, R., G. Karadzhov, D. Alexandrov, J. Glass and P. Nakov, “Predicting factuality of reporting and bias of news media sources”, arXiv preprint arXiv:1810.01765 (2018).

Baron, D. P., “Persistent media bias”, Journal of Public Economics **90**, 1-2, 1–36 (2006).

Beigi, G., R. Guo, A. Nou, Y. Zhang and H. Liu, “Protecting user privacy: An approach for untraceable web browsing history and unambiguous user profiles”, in “Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining”, pp. 213–221 (ACM, 2019a).

Beigi, G., M. Jalili, H. Alvari and G. Sukthankar, “Leveraging community detection for accurate trust prediction”, in “In ASE International Conference on Social Computing, Palo Alto, CA”, (May 2014).

Beigi, G. and H. Liu, “Privacy in social media: Identification, mitigation and applications”, arXiv preprint arXiv:1808.02191 (2018a).

Beigi, G. and H. Liu, “Similar but different: Exploiting users’ congruity for recommendation systems”, in “International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction”, (Springer, 2018b).

Beigi, G., A. Mosallanezhad, R. Guo, H. Alvari, A. Nou and H. Liu, “Privacy-aware recommendation with private-attribute protection using adversarial learning”, in “Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining”, (ACM, 2020).

Beigi, G., S. Ranganath and H. Liu, “Signed link prediction with sparse data: The role of personality information”, in “Companion Proceedings of The 2019 World Wide Web Conference”, pp. 1270–1278 (ACM, 2019b).

Beigi, G., K. Shu, R. Guo, S. Wang and H. Liu, “Privacy preserving text representation learning”, in “Proceedings of the 30th ACM Conference on Hypertext and Social Media”, pp. 275–276 (2019c).

Beigi, G., K. Shu, Y. Zhang and H. Liu, “Securing social media user data-an adversarial approach”, Proceedings of the 29th on Hypertext and Social Media pp. 156–173 (2018).

Beigi, G., J. Tang and H. Liu, “Signed link analysis in social media networks”, in “International AAAI Conference on Web and Social Media (ICWSM)”, (2016a).

Beigi, G., J. Tang, S. Wang and H. Liu, “Exploiting emotional information for trust/distrust prediction”, in “Proceedings of the 2016 SIAM International Conference on Data Mining (ICDM)”, (2016b).

Belkin, M., P. Niyogi and V. Sindhwani, “On manifold regularization.”, in “AISTATS”, (Citeseer, 2005).

Belkin, M., P. Niyogi and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples”, Journal of machine learning research **7**, Nov, 2399–2434 (2006).

Bengio, Y., O. Delalleau and N. Le Roux, *In semi-supervised learning* (MIT Press, 2006).

Benigni, M. C., K. Joseph and K. M. Carley, “Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter”, PloS one (2017).

Blei, D. M., A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation”, the Journal of machine Learning research **3**, 993–1022 (2003).

Blondel, V. D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre, “Fast unfolding of communities in large networks”, Journal of Statistical Mechanics: Theory and Experiment (2008).

Blum, A. and T. Mitchell, “Combining labeled and unlabeled data with co-training”, in “Proceedings of the eleventh annual conference on Computational learning theory”, pp. 92–100 (ACM, 1998).

Broniatowski, D. A., A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn and M. Dredze, “Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate”, American journal of public health **108**, 10, 1378–1384 (2018).

Cao, Q., X. Yang, J. Yu and C. Palow, “Uncovering large groups of active malicious accounts in online social networks”, in “CCS”, (2014).

Chen, C., K. Wu, S. Venkatesh and R. K. Bharadwaj, “The best answers? think twice: online detection of commercial campaigns in the CQA forums”, in “ASONAM”, (2013).

Chen, C., K. Wu, S. Venkatesh and X. Zhang, “Battling the internet water army: Detection of hidden paid posters”, CoRR **abs/1111.4297**, URL `http://arxiv.org/abs/1111.4297` (2011).

Chu, Z., S. Gianvecchio, H. Wang and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?", IEEE Transactions on Dependable and Secure Computing **9**, 6, 811–824 (2012).

Cortes, C. and V. Vapnik, "Support-vector networks", Machine learning **20**, 3, 273–297 (1995).

Cui, P., S. Jin, L. Yu, F. Wang, W. Zhu and S. Yang, "Cascading outbreak prediction in networks: A data-driven approach", in "KDD", (2013).

Daley, D. J. and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure* (Springer Science & Business Media, 2007).

Davis, C. A., O. Varol, E. Ferrara, A. Flammini and F. Menczer, "Botornot: A system to evaluate social bots", (International World Wide Web Conferences Steering Committee, 2016).

DellaVigna, S. and E. Kaplan, "The fox news effect: Media bias and voting", The Quarterly Journal of Economics **122**, 3, 1187–1234 (2007).

Desplaces, C., "Police run 'Prostitution' sting; 19 men arrested, charged in Fourth East Dallas operation.", (Dallas Morning News, 21 Nov. 1992, 34A. Web. 24 Apr. 2012).

Dickinson Goodman, J. and M. Holmes, "Can we use RSS to catch rapists", (2011).

Didelez, V., "Graphical models for marked point processes based on local independence", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 1, 245–264 (2008).

Dubrawski, A., K. Miller, M. Barnes, B. Boecking and E. Kennedy, "Leveraging publicly available data to discern patterns of human-trafficking activity", Journal of Human Trafficking **1**, 1, 65–85 (2015).

Entman, R. M., "Framing: Toward clarification of a fractured paradigm", Journal of communication **43**, 4, 51–58 (1993).

Ferrara, E., "Disinformation and social bot operations in the run up to the 2017 french presidential election", (2017).

Ferrara, E., O. Varol, C. Davis, F. Menczer and A. Flammini, "The rise of social bots", Communications of the ACM **59**, 7, 96–104 (2016).

Fu, H. C.-Y., Yu-Hsiang and C.-T. Sun, "Identifying super-spreader nodes in complex networks", Mathematical Problems in Engineering (2015).

Gomez-Rodriguez, M., J. Leskovec and B. Schölkopf, "Modeling information propagation with survival theory", in "International Conference on Machine Learning", pp. 666–674 (2013).

Goyal, A., F. Bonchi and L. V. Lakshmanan, “Learning influence probabilities in social networks”, in “WSDM”, (2010).

Green, T. and F. Spezzano, “Spam users identification in wikipedia via editing behavior”, ICWSM (2017).

Grigor’yan, A., “Heat kernels on weighted manifolds and applications”, Cont. Math **398**, 93–191 (2006).

Gupta, A., P. Kumaraguru, C. Castillo and P. Meier, *TweetCred: Real-Time Credibility Assessment of Content on Twitter* (Springer International Publishing, 2014).

Gupta, A., H. Lamba and P. Kumaraguru, “$1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter”, in “2013 APWG eCrime Researchers Summit”, (2013).

Hawkes, A. G., “Spectra of some self-exciting and mutually exciting point processes”, Biometrika **58**, 1, 83–90, URL `http://www.jstor.org/stable/2334319` (1971).

Hetter, K., “Fighting sex trafficking in hotels, one room at a time”, URL `http://www.cnn.com/2012/02/29/travel/hotel-sex-trafficking/` (2012).

Hochreiter, S. and J. Schmidhuber, “Long short-term memory”, Neural computation **9**, 8, 1735–1780 (1997).

Hooi, B., H. A. Song, A. Beutel, N. Shah, K. Shin and C. Faloutsos, “Fraudar: Bounding graph fraud in the face of camouflage”, in “Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining”, pp. 895–904 (ACM, 2016).

Horne, B. D. and S. Adali, “This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news”, in “Eleventh International AAAI Conference on Web and Social Media”, (2017).

Hughes, D. M., “The use of new communications and information technologies for sexual exploitation of women and children”, Hastings Women’s Law Journal **13**, 1, 129–148 (2002).

Hughes, D. M. *et al.*, “The demand for victims of sex trafficking”, Women’s Studies Program, University of Rhode Island (2005).

Kanaris, I., K. Kanaris and E. Stamatatos, “Spam detection using character n-grams.”, in “SETN”, vol. 3955 of *Lecture Notes in Computer Science*, pp. 95–104 (Springer, 2006).

Kempe, D., J. Kleinberg and E. Tardos, “Maximizing the spread of influence through a social network”, in “KDD”, (2003).

Kennedy, E., “Predictive patterns of sex trafficking online”, (Dietrich College Honors Theses, 2012).

Khader, M., *Combating Violent Extremism and Radicalization in the Digital Era*, Advances in Religious and Cultural Studies (IGI Global, 2016).

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel", (1975).

Klausen, J., C. Marks and T. Zaman, "Finding online extremists in social networks", CoRR **abs/1610.06242** (2016).

Kleinberg, S., "A logic for causal inference in time series with discrete and continuous variables", in "IJCAI", (2011).

Kleinberg, S. and B. Mishra, "The temporal logic of causal structures", CoRR **abs/1205.2634** (2012).

Konishi, T., T. Iwata, K. Hayashi and K.-I. Kawarabayashi, "Identifying key observers to find popular information in advance", in "IJCAI", (2016).

Kudugunta, S. and E. Ferrara, "Deep neural networks for bot detection", arXiv preprint arXiv:1802.04289 (2018).

Lancichinetti, A., F. Radicchi, J. J. Ramasco and S. Fortunato, "Finding statistically significant communities in networks", PloS one **6**, 4, e18961 (2011).

Latonero, M., "Human trafficking online: The role of social networking sites and online classifieds", Available at SSRN 2045851 (2011).

Li, M. and P. M. Vitányi, *An introduction to kolmogorov complexity and its applications* (Springer Publishing Company, Incorporated, 2008), 3 edn.

Lloyd, R., "An open letter to jim Buckmaster", (2012).

Luo, D., H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang and W. Zhang, "Multi-task multi-dimensional hawkes processes for modeling event sequences", in "Twenty-Fourth International Joint Conference on Artificial Intelligence", (2015).

Manning, C., R. PRABHAKAR and S. HINRICH, "Introduction to information retrieval, volume 1 cambridge university press", Cambridge, UK (2008).

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", in "Advances in neural information processing systems", pp. 3111–3119 (2013).

Miller, K., E. Kennedy and A. Dubrawski, "Do public events affect sex trafficking activity?", ArXiv e-prints URL `https://arxiv.org/abs/1602.05048` (2016).

Mitchell, T. M., "Learning from labeled and unlabeled data", Machine learning **10**, 701 (2006).

Mitchell, T. M. *et al.*, "Machine learning", (1997).

Morstatter, F., L. Wu, U. Yavanoglu, S. R. Corman and H. Liu, “Identifying framing bias in online news”, ACM Transactions on Social Computing **1**, 2, 5 (2018).

Nagpal, C., K. Miller, B. Boecking and A. Dubrawski, “An entity resolution approach to isolate instances of human trafficking online”, ArXiv e-prints URL `https://arxiv.org/abs/1509.06659` (2015).

Nicholas D., K., “How pimps use the web to sell girls.”, URL `http://www.nytimes.com/2012/01/26/opinion/how-pimps-use-the-web-to-sell-girls.html` (2012).

Pearl, J., *Causality: Models, Reasoning and Inference* (Cambridge University Press, New York, NY, USA, 2009), 2nd edn.

Pei, S., L. Muchnik, J. S. A. Jr., Z. Zheng and H. A. Makse, “Searching for superspreaders of information in real-world social media.”, CoRR (2014).

Phuong, T. M. *et al.*, “Gender prediction using browsing history”, in “Knowledge and Systems Engineering”, pp. 271–283 (Springer, 2014).

Roe--Sepowitz, D., J. Gallagher, K. Bracy, L. Cantelme, A. Bayless, J. Larkin, A. Reese and L. Allbee, “Exploring the impact of the super bowl on sex trafficking”, (2015).

Sarkar, S., H. Alvari and P. Shakarian, “Leveraging motifs to model the temporal dynamics of diffusion networks”, in “Companion Proceedings of The 2019 World Wide Web Conference”, pp. 1079–1086 (ACM, 2019).

Scanlon, J. R. and M. S. Gerber, “Automatic detection of cyber-recruitment by violent extremists”, Security Informatics **3**, 1, 5 (2014).

Scanlon, J. R. and M. S. Gerber, “Forecasting violent extremist cyber recruitment”, IEEE Transactions on Information Forensics and Security **10**, 11, 2461–2470 (2015).

Scheufele, D. A. and D. Tewksbury, “Framing, agenda setting, and priming: The evolution of three media effects models”, Journal of communication **57**, 1, 9–20 (2006).

Shaabani, E., R. Guo and P. Shakarian, “Detecting pathogenic social media accounts without content or network structure”, in “2018 1st International Conference on Data Intelligence and Security (ICDIS)”, pp. 57–64 (IEEE, 2018).

Shaabani, E., A. Sadeghi-Mobarakeh, H. Alvari and P. Shakarian, “An end-to-end framework to identify pathogenic social media accounts on twitter”, IEEE Conference on Data Intelligence and Security (2019).

Shannon, C. E., “A mathematical theory of communication”, ACM SIGMOBILE Mobile Computing and Communications Review **5**, 1, 3–55 (2001).

Shao, C., G. L. Ciampaglia, O. Varol, A. Flammini and F. Menczer, “The spread of fake news by social bots”, arXiv preprint arXiv:1707.07592 (2017).

Soliman, A. B., K. Eissa and S. R. El-Beltagy, “Aravec: A set of arabic word embedding models for use in arabic nlp”, Procedia Computer Science **117**, 256–265 (2017).

Stanton, A., A. Thart, A. Jain, P. Vyas, A. Chatterjee and P. Shakarian, “Mining for causal relationships: A data-driven study of the islamic state”, CoRR (2015).

Subrahmanian, V. S., A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini and F. Menczer, “The darpa twitter bot challenge”, (2016).

Suppes, P., “A probabilistic theory of causality”, (1970).

Szekely, P. A., C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, D. Stallard, S. S. Karunamoorthy, R. Bojanapalli, S. Minton, B. Amanatullah, T. Hughes, M. Tamayo, D. Flynt, R. Artiss, S.-F. Chang, T. Chen, G. Hiebel and L. Ferreira, “Building and using a knowledge graph to combat human trafficking.”, in “International Semantic Web Conference (2)”, vol. 9367 of *Lecture Notes in Computer Science*, pp. 205–221 (Springer, 2015).

Thomas, K., C. Grier, D. Song and V. Paxson, “Suspended accounts in retrospect: an analysis of twitter spam”, in “Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference”, pp. 243–258 (ACM, 2011).

Tran, L., M. Farajtabar, L. Song and H. Zha, “Netcodec: Community detection from individual activities”, in “Proceedings of the 2015 SIAM International Conference on Data Mining”, pp. 91–99 (SIAM, 2015).

van der Maaten, L. and G. Hinton, “Visualizing high-dimensional data using t-SNE”, Journal of Machine Learning Research, vol. 9 pp. 2579–2605 (2008).

Varol, O., E. Ferrara, C. A. Davis, F. Menczer and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization”, ICWSM (2017a).

Varol, O., E. Ferrara, F. Menczer and A. Flammini, “Early detection of promoted campaigns on social media”, EPJ Data Science (2017b).

Wang, K., Y. Xiao and Z. Xiao, “Detection of internet water army in social network”, (2014).

Weng, L., F. Menczer and Y.-Y. Ahn, “Predicting successful memes using network and community structure.”, in “ICWSM”, (2014).

Xia, Z., C. Liu, N. Z. Gong, Q. Li, Y. Cui and D. Song, “Characterizing and detecting malicious accounts in privacy-centric mobile social networks: A case study”, in “Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining”, pp. 2012–2022 (ACM, 2019).

Xiao, S., M. Farajtabar, X. Ye, J. Yan, L. Song and H. Zha, “Wasserstein learning of deep generative point process models”, in “Advances in Neural Information Processing Systems”, pp. 3247–3257 (2017).

Xu, H., M. Farajtabar and H. Zha, “Learning granger causality for hawkes processes”, in “International Conference on Machine Learning”, pp. 1717–1726 (2016).

Zannettou, S., T. Caulfield, E. De Cristofaro, N. Kourtelris, I. Leontiadis, M. Sirivianos, G. Stringhini and J. Blackburn, “The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources”, in “Proceedings of the 2017 Internet Measurement Conference”, pp. 405–417 (ACM, 2017).

Zhang, X., J. Zhu, Q. Wang and H. Zhao, “Identifying influential nodes in complex networks with community structure”, Know.-Based Syst. **42** (2013).

Zhou, D., O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, “Learning with local and global consistency”, in “Advances in neural information processing systems”, pp. 321–328 (2004a).

Zhou, D., O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, “Learning with local and global consistency”, in “Advances in Neural Information Processing Systems 16”, pp. 321–328 (MIT Press, 2004b).

Zhou, K., H. Zha and L. Song, “Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes”, in “Artificial Intelligence and Statistics”, pp. 641–649 (2013).

Zhu, K. and L. Ying, “Information source detection in the sir model: A sample-path-based approach”, IEEE/ACM Trans. Netw. **24**, 1 (2016).

BIOGRAPHICAL SKETCH

Hamidreza Alvari has been a PhD student in Computer Science since January 2013. His research has focused on social computing, social network analysis, user behavioral modeling, and misinformation detection. He has authored or co-authored over 30 papers including a book chapter in major data mining and machine learning international venues, such as Cambridge University Pres, The Web (former WWW), WSDM, SDM, CIKM, SNAM, ASONAM, ISI, and SBP. Hamidreza's research has received over 300 citations globally and has led to practical solutions for real-world problems of societal significance such as malicious user detection in social media, detecting online human trafficking and locating missing persons. Hamidreza has worked on several projects funded by agencies such as NSF, Department of State, and the U.S. Office of Naval Research. His work has been featured multiple times in the local and major news outlets including Wall Street Journal. He is also a co-inventor on several provisional patents. Hamidreza has also reviewed articles on many occasions, which speaks to the amount of trust imbued in him by the editorial boards of major journals and conference proceedings including but not limited to KDD, IJCAI, RecSys, CSCW, IEEE Transactions on Neural Networks and Learning Systems, and AAMAS. He has served on the Program Committee for RecSys conference in 2018 and 2019 and for SBP in 2019. He also interned as a data scientist at American Express, Phoenix, in 2015. Hamidreza obtained his master's degree in Artificial Intelligence from Shiraz University, Iran, in 2012 and his bachelor's degree in Software Engineering from Shahid Chamran University of Ahvaz, Iran, in 2009.