COMBATING USER MISBEHAVIOR ON SOCIAL MEDIA

A Dissertation

by

CHENG CAO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	James Caverlee
Committee Members,	Jianer Chen
	Richard Furuta
	Randy Kluver
Head of Department,	Dilma Da Silva

December 2017

Major Subject: Computer Science

ABSTRACT

Social media encourages user participation and facilitates user's self-expression like never before. While enriching user behavior in a spectrum of means, many social media platforms have become breeding grounds for user misbehavior. In this dissertation we focus on understanding and combating three specific threads of user misbehaviors that widely exist on social media — spamming, manipulation, and distortion.

First, we address the challenge of detecting spam links. Rather than rely on traditional blacklist-based or content-based methods, we examine the behavioral factors of both who is posting the link and who is clicking on the link. The core intuition is that these behavioral signals may be more difficult to manipulate than traditional signals. We find that this purely behavioral approach can achieve good performance for robust behavior-based spam link detection.

Next, we deal with uncovering manipulated behavior of link sharing. We propose a four-phase approach to model, identify, characterize, and classify organic and organized groups who engage in link sharing. The key motivating insight is that group-level behavioral signals can distinguish manipulated user groups. We find that levels of organized behavior vary by link type and that the proposed approach achieves good performance measured by commonly-used metrics.

Finally, we investigate a particular distortion behavior: making bullshit (BS) statements on social media. We explore the factors impacting the perception of BS and what leads users to ultimately perceive and call a post BS. We begin by preparing a crowdsourced collection of real social media posts that have been called BS. We then build a classification model that can determine what posts are more likely to be called BS. Our experiments suggest our classifier has the potential of leveraging linguistic cues for detecting social media posts that are likely to be called BS.

We complement these three studies with a cross-cutting investigation of learning user topical profiles, which can shed light into what subjects each user is associated with, which can benefit the understanding of the connection between user and misbehavior. Concretely, we propose a unified model for learning user topical profiles that simultaneously considers multiple footprints and we show how these footprints can be embedded in a generalized optimization framework.

Through extensive experiments on millions of real social media posts, we find our proposed models can effectively combat user misbehavior on social media.

DEDICATION

For my wife, Xi

ACKNOWLEDGMENTS

I owe my gratitude to all who made this dissertation possible.

I consider myself extremely lucky to have Dr. James Caverlee as my advisor. I always admire his visions in finding interesting and important research problems. He keeps guiding me to take the most innovative perspectives for solving the right problems or applications. He has the best presentation skills I have ever seen. Fortunately, I have learned a lot from him through our countless iterations of editing slides and practicing talks. He teaches me how to make the slides organized with a clear storyline. He advises me how to write a well-organized paper. In daily life, besides research, it is not an exaggeration to say that my advisor is more like my friend. He is such a wonderful person that I am pretty sure I have some uniquely great experience with my advisor that other graduate students in the world do not have. We walk to Starbucks to grab a cup of coffee together. We talk about yesterday's football games and this week's new movies. Sometimes he shares his life lessons with me. I will forever be grateful to his mentorship and inspiration on me.

I would like to thank the rest of my dissertation committee — Dr. Richard Furuta, Dr. Jianer Chen, and Dr. Randy Kluver — for their invaluable feedbacks on my dissertation and serving on my committee.

I owe special thanks to my good friends, Dr. Xia Hu and Dr. Kyumin Lee. My collaborations with them have been one of the most enriching experience in my pursuing of PhD. I learned a lot from their working styles.

My journey of PhD study has been vibrant and exciting thanks to all my lab-mates: Majid Alfifi, Hancheng Ge, Habeeb Hooshmand, Parisa Kaghazgaran, Haokai Lu, Wei Niu, Henry Qiu, Haiping Xue, Yin Zhang, and Xing Zhao.

My parents are my foundation. It was my mother who encouraged me to study abroad when I was a junior undergraduate student in China. Her suggestion changed the trajectory of my life. I want to thank my whole family for their ever-lasting love and support.

I am deeply indebted to my wife Xi. She has endured much because of me. She has devoted herself to me and our child. This dissertation is for her, the love of my life.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor James Caverlee, Richard Furuta, and Jianer Chen of the Department of Computer Science and Engineering and Professor Randy Kluver of the Department of Communication.

All work conducted for this dissertation was completed by the student independently.

Funding Sources

No outside funding sources were used in this study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Motivation

The cornerstone of social media is user-generated behavioral activity. For example, users power these systems by sharing content, commenting, messaging, befriending, and engaging in among many other behaviors. This behavioral activity implicitly reveals user preferences, interests, and relationships (e.g., Youtube users vote “like” or “dislike” for a video; Twitter users retweet or mention other’s tweets), which can play an important role in a variety of applications. For example, for social media service providers, by understanding how their users behave in the system, they can improve their system’s personalization module to enhance user experience. For online marketers, they can exploit user’s preference and interaction patterns to spread their content quickly and widely [1]. For Internet service providers, learning traffic patterns on social media websites can guide traffic optimization in their infrastructures [2].

While facilitating user’s self-expression and information spread like never before [3, 4], many social media platforms have also become major breeding grounds for *user misbehavior*. These misbehaviors lead to a big volume of misinformation on social media such as spam [5, 6], fraud [7, 8], and rumors [9, 10], which may result in several levels of damage to user experience and even society. For example, spammers on Facebook send victims unsolicited requests and messages many of which include links to commercial ads, phishing websites, or malware [6, 11, 12]. Fake accounts on Yelp purposely post deceptive reviews to mislead potential customers due to profit or fame [13, 14]. It has been observed that abusive behaviors on Twitter have considerable influence on the outcome of political campaigns [15, 16, 17]. And rumors that are deliberately spread during mass emergencies and disasters (e.g., Hurricane Sandy in 2012 and Boston Marathon bombing in 2013), can cause anxiety, panic, and insecurity to the whole society [8, 9, 18].

In this dissertation we focus on *understanding* and *combating* three specific threads of user misbehavior that widely exist on social media — spamming, manipulation, and distortion (see Figure 1.1). And toward combating these misbehaviors, we investigate one specific important application for each misbehavior as follows:

- **Misbehavior 1: Spamming**. First, we address the problem of detecting spam links (URLs). Link sharing is a core attraction of many existing social media systems like Twitter and Facebook. Recent studies find that around 25% of all status messages in these systems contain links [19, 20], amounting to millions of links shared per day. With this opportunity comes challenges, however, from malicious users who share spam links to promote ads, phishing, malware, and other low-quality content. Those spamming behaviors ultimately degrade the quality of information available in these systems. Several recent efforts have identified the problem of spam links on the Web [21, 22, 23], but it has not been fully explored particularly on social media.

- **Misbehavior 2: Manipulation**. Next, we take a step further from individual spamming behavior, and are interested in uncovering manipulated behavior of coordination in link sharing. While some link sharing is organic, other sharing is strategically organized with a common (perhaps, nefarious) purpose, such as campaign-like advertising and other adversarial propaganda. These manipulated campaigns conduct fraudulent activities, which can wreck havoc on business, politics, and social security [24, 25, 26, 27]. To purify and improve the information quality on social media, it becomes imperative that the service providers can detect those manipulated behaviors of link sharing.

- **Misbehavior 3: Distortion**. Finally, we investigate one concrete distortion behavior that widely exists on social media — making bullshit (BS) statements. We follow the concept of BS by philosopher Harry Frankfurt: BS is a statement that does not address facts, but rather distorts what the BS-er is up to [28, 29]. The current ecosystem of online social media has made it trivial to spread distorted information without accountability, accelerating the production of BS. Some BS statement on social media can increase stress and fear, which often leads to real-world violence [30]. Moreover, it has been observed that BS has reached issues like politics and advertising where it can actually cause severe problems for BS-receivers [17, 31].

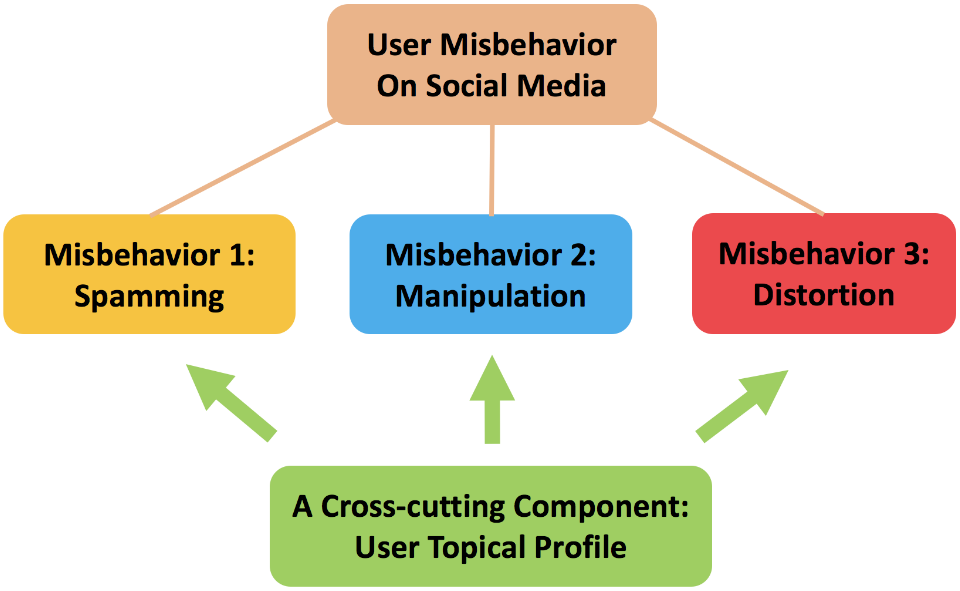

User Misbehavior On Social Media
Misbehavior 1: Spamming
Misbehavior 2: Manipulation
Misbehavior 3: Distortion
A Cross-cutting Component: User Topical Profile

Figure 1.1: An overview of all problems studied in this dissertation

1.2 Research Challenges

While investigating these three types of user behavior is important, there are significant research gaps toward modeling and solving them efficiently and effectively. Here, we identify several main research challenges:

- **Defining Misbehavior**: Detecting user misbehavior on social media has not been fully explored so that many relevant problems are not clearly formulated. For example, when studying manipulation in link sharing, the definition of organized behavioral pattern is unsettled, e.g., how to mathematically define organic behavior and organized behavior in the context of link sharing? This issue becomes even more obvious in the problem of BS detection. How to adapt the concept of bullshit in philosophy or linguistics and properly define BS within the scope of social media? How to formulate a model of automatic BS detection in social media? All the problem formulations need to be resolved before solutions are figured out.

- **Distinguishing Misbehavior from Legitimate Behavior**: Considering the massive noise on social media, it becomes extremely challenging to clearly distinguish the trails of user misbehavior, even for human judgment. For example, to evaluate the distortion in a post, we need to find out whether the poster cares if his post is true or not. Yet, it is unrealistic to fully mine a user's intent of distortion — we will never truly know what a user thinks when he posts. In the problem of detecting manipulated link sharing, the difference between the two extremes — organic and organized — is often not a simple distinction. Those "good intriguers" try hard to disguise themselves, which makes effectively differentiating them a tricky job.

- **Uncovering Behavioral Signals**: Behavioral signals have historically been difficult to collect. Many online social media systems provide restricted (or even no) research access (like public API) to posts published on them, such as Facebook and Instagram. Even for those systems that do provide a sample of its posts (like Twitter), it is still hard to collect fine-grained behavioral signals. For instance, in the problem of spam link detection, we do not know how those links posted on social media are actually received by the users via clicks. As a result, much insight into behavioral patterns of link sharing has been limited to proprietary and non-repeatable studies.

1.3 Contributions of This Dissertation

Keeping those research challenges in mind, in this dissertation we aim for developing novel computational models toward *discovering*, *formulating*, *modeling*, and *solving* those problems introduced in Section 1.1. In particular, we first turn in this section to describe our contributions toward each of those three misbehavior-related applications. Then, we introduce a fourth problem we are going to study, serving as a *cross-cutting* component that supports better understandings of the other three tasks shown in Figure 1.1.

1.3.1 Detecting Spam Links via Behavioral Analysis

We investigate the potential of behavioral analysis for uncovering which links are spam and which are not. By behavioral signals, we are interested both in the aggregate behavior of who is posting these links in social systems and who is clicking on these links once they have been posted. These behavioral signals offer the potential of rich contextual evidence about each link that goes beyond traditional spam detection methods that rely on blacklists, the content of the link, its in-links, or other link-related metadata.

Concretely, we propose and evaluate fifteen click and posting-based behavioral features, including: for postings — how often the link is posted, the frequency dispersion of when the link is posted (e.g., is it posted only on a single day in a burst? or is it diffusely posted over a long period?), and the social network of the posters themselves; and for clicks — we model the click dynamics of each link (e.g., does it rapidly rise in popularity? are there multiple spikes in popularity?) and consider several click-related statistics about each link, including the total number of clicks accumulated and the average clicks per day that a link was actually clicked. We accommodate these behavioral signals into a classification model to automatically detect spam links.

Through extensive experimental study over a dataset of 7 million Bitly-shortened links posted to Twitter, we find that these behavioral signals provide overlapping but fundamentally different perspectives on links. Through this purely behavioral approach for spam link detection, we can achieve high precision (0.86), recall (0.86), and area-under-the-curve (0.92). Compared to many existing methods that focus on either the content of social media posts or the destination page — which may be easily manipulated by spammers to evade detection — this behavior-based approach suggests the potential of leveraging these newly-available behavioral cues for robust, on-going spam detection.

1.3.2 Revealing Organized Link Sharing Behavior

In the context of manipulated link sharing on social media, we investigate a method to automatically (i) identify user groups in terms of similar link sharing behaviors; and (ii) differentiate strategically organized and genuinely organic user groups, through the development of a link sharing behavior based model. The key insight is that the publicly available link sharing information can help model users with similar behaviors of link sharing, and that some group-level behavioral signals can help characterize whether the behaviors of a group of users are organic or organized.

Concretely, we propose a four-step approach. We first formulate the behavior of link sharing based on its three key factors: user, link, and the posting activity. Based upon such a model, we design a similarity measurement of user behaviors in link sharing. Then, given the pairwise similarity function, we build a user graph model from which we identify user groups each of which contains users with similar link sharing behaviors. Next, on the group level, we characterize the organic and organized user groups based on the link posting behaviors of their members. Finally, we embed those characteristics into a classification framework to systematically distinguish organic and organized groups of users.

We test our approach on four different classification algorithms and in most cases it performs well in terms of precision, recall, F-measure, and ROC area. Random Forest algorithm works best with 0.921 ROC. Our experimental analysis demonstrated the capability of our approach for (i) understanding users with similar link sharing behaviors; and (ii) distinguishing the level of manipulated user misbehaviors in link sharing.

1.3.3 Identifying BS on Social Media

We seek for footprints of human judgments left by other social media users to help locate potential BS posts, and we focus on the replying behavior as the evidence of a "*BS call*". BS can originate from anywhere (i.e. many users post BS; no user posts *only* BS), which makes it very difficult to identify. The volume of potential BS and other noise on social media make this problem even trickier. We choose replying behavior because replies can help us easily track *who* and *what* have been called BS. Also, replies are often topically motivated, meaning repliers care about the content of the post as opposed to other factors (e.g. author of the post) [32].

Instead of diving into the extremely nuanced task of detecting "actual" BS, we build a model that can automatically determine what social media posts are likely to be *called* BS. There is a gap between *BS-called* posts and *actual* BS posts — we have observed many real examples showing identifying BS is a challenging task, even for humans. We focus on the audience's perception of BS. We do this by mining how the audience perceives poster's intent of BS-ing through signals from the post itself. Our work can serve as a stepping-stone to the ultimate goal of BS detection. For instance, our results can be used to filter posts that are unlikely to be BS from the vast social media stream. This could provide a more narrow search-space to develop a true BS identification method.

In concrete, we first prepare a curated crowd-sourced collection of BS-called tweets gathered over a sequence of 100 consecutive days. Next, we identify four factors within a post that influence a reader's perception: *attitude*, *sentiment*, *sincerity*, and *content*. This characterization leads to a classification model to differentiate between posts that are more likely to be called BS and posts that are less likely to be called BS. Finally, we conduct a series of experiments showing our model is capable of leveraging linguistic cues for identifying posts are likely to be called BS, suggesting the great potential of a preliminary BS auto-filter on social media.

1.3.4 A Cross-cutting Component: Learning User Topical Profile

Each of those three applications mentioned above has its own scope in the context of user misbehavior, but they all connect to the user who is the *cross-cutting* subject. This triggers an important question — who are those users? Answering this question can shed light into what subjects each user is associated with, which can serve as an important cross-cutting component toward better understanding the mutual relationship between users and the misbehaviors they commit. For example, we can analyze a user's "morality" on social media based on his behaviors; and judging a user "good" or "bad" can guide further investigations on this user's behaviors on social media.

Therefore, we study a fourth task of learning social media user's *topical profile* — topical interests (i.e., what she likes) or expertise (i.e., what she is known for). On the one hand, those users who are learned to have "problematic" topical profiles are suspicious targets of misbehaviors. On the other hand, those who associate with "high-profiles" (like experts with enormous reputation in certain domain) tend to keep distance from noxious behavior. For instance, if we find a group of Twitter accounts whose topical profiles are highly similar and all associate with spam, it becomes wary that these accounts are likely to conduct manipulated spamming behaviors. If a Facebook user posts extensively on topics that are distant from the domain of his real expertise (e.g., a sports journalist makes voluminous claims on politics), it hints that this user is probably BS-ing. Besides, in general, the problem of learning a user's topical profile also has its own important applications in many domains. It can improve user experience and powering important applications like personalized web search [33], recommendation system [34, 35], expert mining [36], and community detection [37].

We propose to exploit heterogeneous footprints (e.g., tags, friends, interests, behavior) for intelligently learning user topical profiles. Based on a small set of explicit user tags, our goal is to extend this known set to the wider space of users who have no explicit tags. The key intuition is to identify "similar" users in terms of their topical profiles by exploiting their similarity in a *footprint space*. For instance, Twitter users who post similar hashtags may have similar interests, and YouTube users who upvote the same videos may have similar preferences. Such evidence of *homophily* has been widely studied in the sociological literature [38] and repeatedly observed in online social media, e.g., [39, 40, 41, 42, 43].

In a summary, we make four main contributions toward learning user topical profile on social media. First, we formulate the problem of learning user topical profiles in social media, with a focus on leveraging heterogeneous footprints. Second, we demonstrate how to model different footprints (e.g., like interests, social, and behavioral footprints) under this framework, and we present a unified 2-D factorization model in which we simultaneously consider all of these footprints. Third, we then extend this initial approach through a generalized model that integrates the pairwise relations across all potential footprints via a tensor-based model, which provides a more robust framework for user profile learning. Finally, through extensive experiments, we find the proposed model is capable of learning high-quality user topical profiles, and leads to a 10-15% improvement in precision and mean average error versus a state-of-the-art baseline. We find that behavioral footprints are the single strongest factor, but that intelligent integration of multiple footprints leads to the best overall performance.

1.4 Structure of This Dissertation

The remainder of this dissertation is organized as follows:

- Section 2 — we explore the problem of detecting spam links via behavioral analysis. We first give the problem statement and setup. Then we provide posting-based features and click-based features toward building a classification model of spam link detection. Finally we present extensive experiment results.

- Section 3 — we investigate the problem of revealing organized link sharing behavior. We present our four-stage approach step by step: (1) modeling the behavior of link sharing, (2) extracting user groups with similar behavioral patterns, (3) characterizing organic and organized groups based on group-level behavioral signals, and (4) embedding extracted features into a classification framework.

- Section 4 — we look into the problem identifying BS on social media. We first introduce our motivation and methodology of data collection. We then formulate the problem and explain four factors impacting the likelihood a post gets called BS. In the end we present experiment analysis on our dataset and a classification model detecting social media posts that are likely called BS.

- Section 5 — we propose a generalized framework for learning user topical profiles. After providing preliminaries including all notations and problem definition, we first identify multiple implicit footprints and demonstrate how to model them. We then introduce a matrix-factorization-based approach, before extending this version to a more general tensor-based approach.

- Section 6 — We conclude with a summary of contributions made by this dissertation research. We discuss several potential directions of future extension to the results presented in this dissertation.

2. COMBATING SPAMMING: DETECTING SPAM LINKS VIA BEHAVIORAL ANALYSIS[*]

In this section we explore the problem of detecting spam links via behavioral analysis, as a specific application toward tackling the first misbehavior — spamming. We begin with the introduction and related work, followed by problem setup and details of our solution. In the end we report a series of experiments designed to evaluate the quality of the proposed solution.

2.1 Introduction

Link sharing is one of the key functions in most existing social media systems. In the early days of Twitter in 2007, Java et al. already saw that about 13% of a collection of 1.3 million tweets included a link [44]. Recent studies have confirmed the ongoing popularity of link sharing on Twitter. In 2010, Boyd et al. found 22% of a sample of 720,000 tweets included a link [45]. And in 2011, Rodrigues et al. found that nearly a quarter of 1.7 billion tweets contained links [19].

With the popularity of link sharing comes challenges, however, from malicious user behaviors of spreading phishing, malware, and other spam content. Indeed, several recent efforts have identified the problem of spam links in social media [46, 22, 23, 47], which ultimately deteriorates the information quality in these systems. Defending social media systems from spam links is important for shielding unsuspecting social media users from these threats.

Our goal is to investigate the potential of *behavioral analysis* for uncovering which

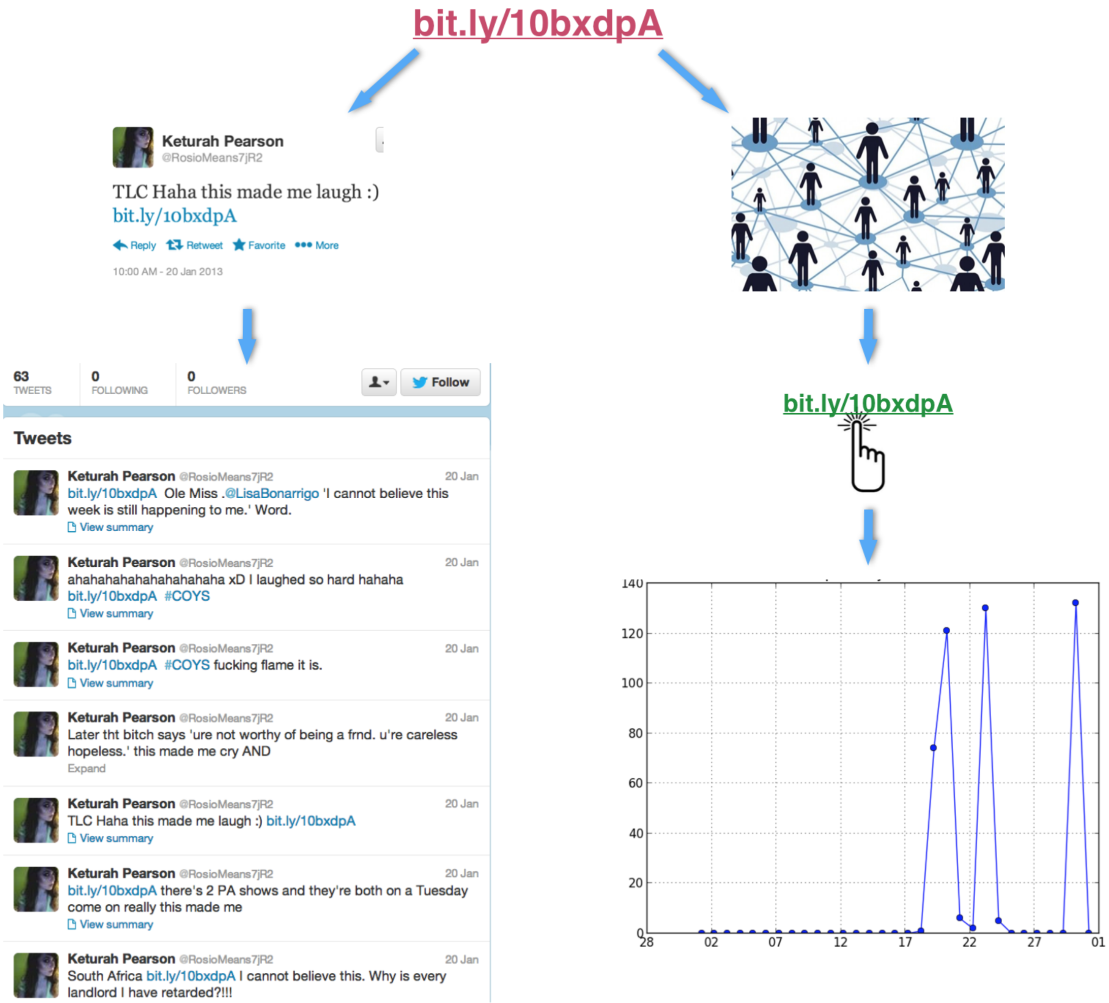
[*] Part of this section is reprinted with permission from "Detecting Spam URLs in Social Media via Behavioral Analysis" by Cao C., Caverlee J. (2015) In: Hanbury A., Kazai G., Rauber A., Fuhr N. (eds) Advances in Information Retrieval. ECIR 2015. Lecture Notes in Computer Science, vol 9022. Springer, Cham. Copyright 2015 by Springer.

links are spam and which are not. By behavioral signals, we are interested both in the aggregate behavior of who is *posting* these links on social media and who is *clicking* on these links once they have been posted. These behavioral signals offer the potential of rich contextual evidence about each link that goes beyond traditional detection methods that rely on blacklists, the content in the link, its in-links, or other link-related metadata.

Unfortunately, it has historically been difficult to investigate behavioral patterns of posts and clicks. First, many social media platforms provide restricted (or even no) access to posts, like Facebook. Second, even for those systems that do provide research access to a sample of its posts (like Twitter), it has been difficult to assess how these links are actually received by the users of the system via clicks. As a result, much insight into behavioral patterns of link sharing has been limited to proprietary and non-repeatable studies.

Hence, we begin a behavioral examination of spam link detection through two distinct perspectives (see Figure 2.1): (i) the first is via a study of how these links are posted through publicly-accessible Twitter data; (ii) the second is via a study of how these links are received by measuring their click patterns through the publicly-accessible Bitly click API. Concretely, we propose and evaluate fifteen *click-based* and *posting-based* behavioral features. For posting we are interested in how often a link is posted, the frequency dispersion of when the link is posted (e.g., is it posted only on a single day in a burst? or is it diffusely posted over a long period?), and the social network of the posters themselves. And for click, we model the click dynamics of each link (e.g., does it rapidly rise in popularity? are there multiple spikes in popularity?) and consider several click-related statistics about each link — including the total number of clicks accumulated and the average clicks per day that a link was actually clicked.

Through extensive experimental study over a dataset of 7 million Bitly-shortened links posted to Twitter, we find that these behavioral signals provide overlapping but fundamentally different perspectives on links. Through this purely behavioral approach for spam



Keturah Pearson @RosioMeans7jR2

TLC Haha this made me laugh :) bit.ly/10bxdpA

Keturah Pearson @RosioMeans7jR2 20 Jan
bit.ly/10bxdpA Ole Miss .@LisaBonarrigo 'I cannot believe this week is still happening to me.' Word.
View summary

Keturah Pearson @RosioMeans7jR2 20 Jan
ahahahahahahahahahaha xD I laughed so hard hahaha bit.ly/10bxdpA #COYS
View summary

Keturah Pearson @RosioMeans7jR2 20 Jan
bit.ly/10bxdpA #COYS fucking flame it is.
View summary

Keturah Pearson @RosioMeans7jR2 20 Jan
Later tht bitch says 'ure not worthy of being a frnd. u're careless hopeless.' this made me cry AND
Expand

Keturah Pearson @RosioMeans7jR2 20 Jan
TLC Haha this made me laugh :) bit.ly/10bxdpA
View summary

Keturah Pearson @RosioMeans7jR2 20 Jan
bit.ly/10bxdpA there's 2 PA shows and they're both on a Tuesday come on really this made me
View summary

Keturah Pearson @RosioMeans7jR2 20 Jan
South Africa bit.ly/10bxdpA I cannot believe this. Why is every landlord I have retarded?!!!

Figure 2.1: Studying spam link detection in social media from two perspectives: (i) Posting behavior (left); (ii) Click behavior (right)

link detection, we can achieve high precision (0.86), recall (0.86), and area-under-the-curve (0.92). Compared to many existing methods that focus on either the content of social media posts or the destination page – which may be easily manipulated by spammers to evade detection – this behavior-based approach suggests the potential of leveraging these newly-available behavioral cues for robust, on-going spam detection.

2.2 Related Work

Links (and in particular, shortened links) have been widely shared on social media systems in recent years. Antoniades et al. [46] conducted the first comprehensive analysis of short links in which they investigated usage-related properties such as life span. With the rising concern of short links as a way to conceal untrustworthy web destinations, there have been a series of studies focused on security concerns of these links, including: a study of phishing attacks through short links [22], geographical analysis of spam short links via usage logs [23], an examination of security and privacy risks introduced in shortening services [47], and a long-term observation of shortening services on security threats [48].

Separately, Twitter spam detection has been widely studied in recent years. In general, three types of approaches have been proposed: user profile based, content based, and network relation based. User profile based methods [49, 50, 51] build classifiers using features extracted from account profiles, e.g., profile longevity. Content-based features [52, 51] focus on the posting text. Network-based features [53, 54, 55] are those extracted from the social graph such as clustering coefficient. A couple of detection systems of suspicious links on Web have been developed. Some of these [56, 57, 58, 59] directly use URL's lexical features, redirecting patterns, and link's metadata such as IP and DNS information. Some [60, 61] consider features extracted from the HTML content of the landing page. Additionally, several dynamic spam link filtering systems have also been developed [62, 63, 64].

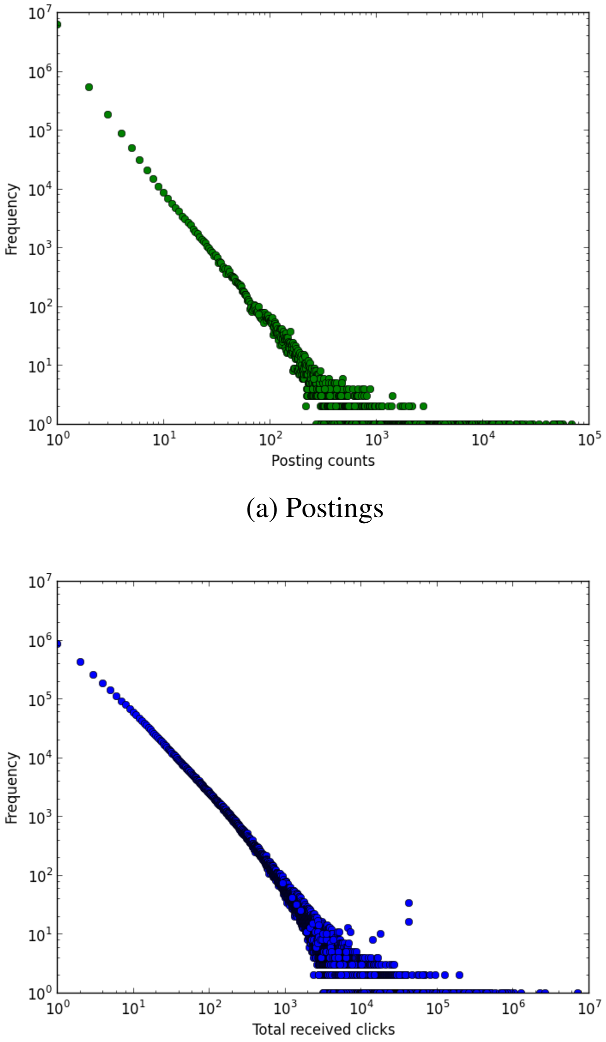
Several recent works have used clicks extracted from the Bitly API, typically to study the properties of known spam links. For example, Grier et al. [52] recovered clicking statistics of blacklisted Bitly links, with the aim of measuring the success of those spam links on Twitter. Maggi et al. [48] submitted malicious long links to the Bitly API in order to examine the performance in terms of spam pre-filtering. Chhabra et al. [22] shortened a set of known phishing long links and analyzed factors like the referrer and location. There recently has been some research on using proprietary server-side click log data to defend against some types of spam (e.g., [65, 66]). In contrast, our aim is to investigate how large-scale publicly-available click-based information may be used as behavioral signals in the context of spam link detection on social media.

2.3 Behavior-based Spam Link Detection

In this section, we investigate a series of behavioral-based features for determining whether a link shared in social media is spam or not. We view this problem as a binary classification problem. For both the posting-based and click-based perspectives, we are interested to explore questions like: What meaningful patterns can we extract from these publicly-available resources? Are posting or click-based features more helpful for spam link detection? And which specific features are most informative?

2.3.1 Problem Statement and Setup

Given a link v that has been shared on a social media platform, the *behavior-based spam link detection problem* is to predict whether v is a spam link through a classifier $c : v \rightarrow \{\text{spam, benign}\}$, based only on behavioral features. We consider two types of behavioral features associated with each link – a set of posting-related behavioral features F_p and a set of click-based behavioral features F_c. Such a behavior-based approach requires both a collection of links that have been shared, as well as the clicks associated with each link. Since many social media platforms (like Facebook) place fairly stringent limits on crawling, we targeted Bitly-shortened links.



(a) Postings

(b) Clicks

Figure 2.2: Distribution of postings and clicks for the sampled dataset.

Postings. Concretely, we first used the Twitter public streaming API to sample tweets during January 2013. We collected only tweets containing at least one Bitly link (that is, a complete link that had been shortened using the Bitly link shortening service). In total, we collected 13.7 million tweets containing 7.29 million unique Bitly-shortened links. We see in Figure 2.2a the typical "long tail" distribution: a few links have been posted upwards of

100,000 times, whereas most have been posted once or twice.

Clicks. We accessed the Bitly API to gather fine-grained click data about each of the 7.29 million links. For example, we can extract the number of clicks per time unit (e.g., minute, hour, day, month) and by country of origin. In total, we find that nearly all – 7.27 million out of 7.29 million – of the links have valid click information, and that 3.6 million (49.5%) of the links were clicked at least once during our study focus (January 2013). As in the case of postings, we find a "long tail" distribution in clicks, as seen in Figure 2.2b.

2.3.2 Posting-based Features

In the first perspective, we aim to study the links through the posting behaviors associated with them. For example, some links are posted by a single account and at a single time. Others may be posted frequently by a single account, or by many accounts. Similarly, links may be temporally "bursty" in their posting times are spread more evenly across time. Our goal in this section is to highlight several features that may describe each link based on its posting behavior.

Posting Count. The first feature of posting behavior is the total number of times a link has been posted on Twitter during our study window. Our intuition is that this count can provide an implicit signal of the topic of the link destination as well as the intent of the sharer: e.g., links that are posted only a few times may indicate more personal, or localized interest. We formulate this feature as *posting count*, denoted as $PostCount(u)$ given a short link u.

Posting Standard Deviation. A Weather Channel link and a CNN breaking news link may have a similar *posting count* on Twitter. However, the Weather Channel link may be posted every day of the month (linking to a routine daily forecast), whereas a breaking news link may be posted in a burst of activity in a single day. To capture this posting concentration, we consider the standard deviation of the days in which a link is posted.

Concretely, for each link u we have a list of days when u was posted. We refer to this list as u's *posting days*, denoted by $PostDays(u)$. We define the *posting standard deviation* of a link u as the standard deviation of all elements in $PostDays(u)$, denoted as $std(u)$. For example, if a link u was posted 10 times on January 22nd and not tweeted on any other day, we have $std(u) = 0$. On the contrary, a link u shared once per day will have a much larger $std(u)$.

Posting Intensity. The posting standard deviation gives insight into how concentrated a link has been posted, but it does not capture the total intensity of the posting. For example, two links both of which have only one single posting day will have the same posting standard deviation, even if one was posted thousands of times while the other appeared only once. To capture this difference, we introduce *posting intensity* to capture how intense the posting behaviors of a link are. Given a link u, we calculate u's "intensity score" via:

$$intensity(u) = \frac{PostCount(u)}{(std(u) * |set(PostDays(u))|) + 1}$$

where $|set(PostDays(u))|$ is the number of distinct posting days of u. For those links whose scores are the highest, they have high posting frequency, but also a low intensity of posting days. To illustrate, we find in our dataset that the link with the highest intensity score was posted nearly 30,000 times on a single day.

Posting User Network. The sharer's personal network and reputation have certain connection with what and why she posts. A typical example is the comparison between celebrities and spammers. Spammers whose networks commonly are sparse tend to post spam links to advertise, whereas a celebrity may not share such low-quality links. Thus, for each link we consider features capturing the poster's personal network. Here, we use the counts of followers and friends as simple proxies for user popularity, and take the *median* among all posters.

2.3.3 Click-based Features

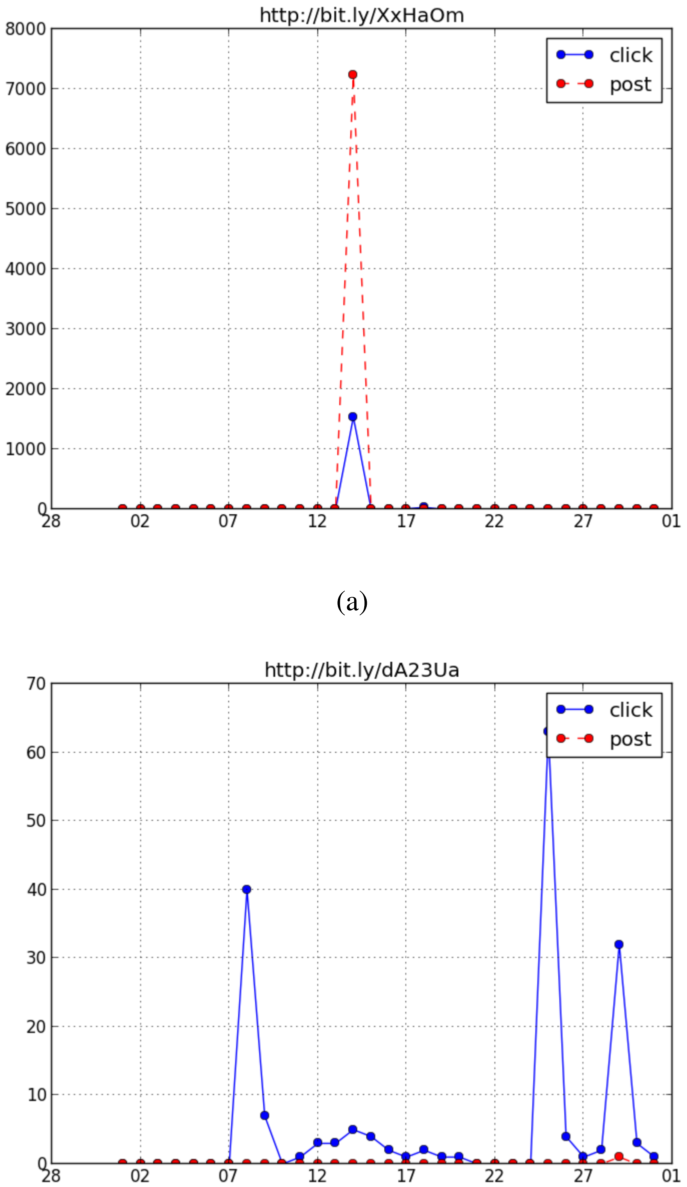
Now we turn our attention to how links are received in social media by considering the clicks that are associated with each link in our dataset. We consider two kinds of clicking patterns: *clicking timeline* features that consider the temporal series of daily received clicks, and *clicking statistics* features that capture overall statistics of the clicks.

For the first kind of clicking pattern, we have every short link's fine-grained daily clicking data – which we can plot as its *clicking timeline*. We adopt three features extracted from this clicking timeline curve:

Rises + Falls. The first question we are interested is: how to capture the overall shape of a link's clicks – do some go up continuously? Or do some periodically go up and down? To measure these changes, let n_i denote the number of clicks on the ith day. We define a *rise* if there exists an i such that $n_{i+1} - n_i > \alpha * n_i$ where α is a threshold. We set it to be 0.1, ensuring the change is non-trivial. Based on this criteria, we observe eight rises in Figure 2.3b (some are quite small). Similarly, let n_i denote the number of clicks on the ith day. We define a *fall* if there exists an i such that $n_i - n_{i-1} > \beta * n_{i-1}$ where β is a threshold value (set to 0.1 in our experiments). We observe eleven falls in Figure 2.3b.

Spikes + Troughs. In Figure 2.3b, we observe that while there are eight rises, there are only five spikes of interest. So rather than capturing consecutive monotonic changes (as in the rises and falls), we additionally measure the degree of fluctuation of a link through its *spikes* and *troughs*. That is, if there is an i such that $n_{i-1} < n_i > n_{i+1}$ we call it a *spike*. If there exists an i satisfying $n_{i-1} > n_i < n_{i+1}$, then we define it is a *trough*. To illustrate, Figure 2.3b has 5 spikes and 3 troughs.

Peak Difference. Naturally, there is a relationship between how and when a link is posted and the clicks the link receives. For example, Figure 2.3a illustrates a close relationship between posting and clicking for a link. In contrast, Figure 2.3b demonstrates a much looser



(a)

(b)

Figure 2.3: The click and post timelines for two links. In (a), post and click behaviors are tightly coupled. In (b), the relationship is more relaxed.

connection, indicating some external interest in the link beyond just its Twitter postings (in this case, the link refers to a university website which attracts attention from many sources beyond Bitly-shortened links on Twitter). To capture the extent to which posting behaviors influence clicks, we define the *peak difference*. For each link, we identify its *clicking peak* as the day it received the most clicks. Similarly, we identify its *posting peak* as the day it was posted the most. Note that a link may have more than one posting peak and clicking peak. Here we define the *peak difference* as the minimum difference between two peaks among all pairs. The range of peak difference is from 0 to 30. In this way, peak difference can represent the level of tightness between clicking and posting.

We augment these timeline-based features with several click statistics:

Total Clicks. The first statistic is the *total clicks* a link received in the period of study, which is a clear indicator of the popularity of a link.

Average Clicks. Given a link's total clicks and posting count, we can measure its *average clicks* per posting. By intuition more exposures bring more clicking traffic, but the average clicks is not necessarily large. Compared to total clicks, average clicks has a starker representation of popularity: many clicks via few postings suggest highly popular.

Clicking Days. Next, we measure the number of *clicking days* in which a link received clicks. This feature captures the consistency of attention on a link.

Max Clicks. *Max clicks* is the maximum daily clicks. Unlike total clicks, this statistic can distinguish links that receive a burst of attention.

Effective Average Clicks. For those links with great total clicks, we observe some have a large number of clicking days while some have only one clicking day but thousands of clicks. Since average clicks considers only the relationship between total clicks and posting count, here we introduce *effective average clicks* defined as $\frac{\text{total clicks}}{\text{clicking days}}$

Click Standard Deviation. We already have features representing the fluctuation of timelines, now we consider a feature for the fluctuation of daily clicks given that we have specific sequence of daily clicks. We can calculate the standard deviation of daily clicks, defined as *click standard deviation*. Note that we fix a month as our time window of study. So, for each short link we have a sequence of daily clicks and we can compute the standard deviation.

Mean Median Ratio. Finally, given 31 daily clicks of a link u, we can calculate its mean and median daily clicks, denoted as $mean(u)$ and $median(u)$ respectively. Now suppose we have a link obtaining thousands of clicks on a day but very few on other days. It may have a considerable mean value but a low median. To build a connection between mean and median, we define *mean median ratio* of u as the following: mean median ratio (u) $= \frac{mean(u)}{median(u)+1}$

2.4 Experiments

In this section, we report a series of experiments designed to investigate the capacity of these two behavioral perspectives – posting-based and click-based – on the effectiveness of spam link detection. Recall that our goal here is to examine the effectiveness of *behavioral signals alone* on spam detection. The core intuition is that these signals are more difficult to manipulate than signals such as the content of a social media post or the content of the underlying destination page. Of course, by integrating additional features such as those studied in previous works – e.g., lexical features of tweet texts, features of user profiles, and so forth – we could enhance the classification performance. Since these traditional features may be more easily degraded by spammers, it is important to examine the capability of a behavioral detector alone.

2.4.1 Experimental Setup

We consider two different sources of spam labels:

Spam Set 1: List Labeled. For the first set of spam labels, we use a communitymaintained link-category website *URLBlacklist* (`http://urlblacklist.com`) that provides a list of millions of domains and their corresponding high-level category (e.g., “News”, “Sports”). Among these high-level categories are two that are clearly malicious: “Malware” and “Phishing”, and so we assign all links in our dataset that belong to one of these two categories as *spam*. We assign all links that belong to the category “Whitelist” as *benign*. It is important to note that many links belong to potentially dangerous categories like “Adult”, “Ads”, “Porn”, and “Hacking”; for this list-based method we make the *conservative* assumption that all of these links belong to the *unknown* class. For all remaining links, we assume they are *unknown*. This labeling approach results in 8,851 spam links, 223 benign, and 1,009,238 unknown. Of these links, we identify all with at least 100 total clicks, resulting in 1,049 spam, 21 benign, and 60,012 unknown. To balance the datasets, we randomly select 1,028 links from the unknowns (but avoid those above-mentioned dangerous categories), and consider them as *benign*, leaving us with 1,049 spam and 1,049 benign links.

Spam Set 2: Manually Labeled. We augment the first spam set with this second collection. We randomly pick and manually label 500 short links, each of which has been posted at least 30 times along with at least 5 original tweets (i.e., not a retweet, nor a reply tweet). We label a link as “spam” if its landing page satisfies one of the following conditions: (1) The browser client (Google Chrome in our work) or Bitly warns visitors that the final page is potentially dangerous before redirecting; (2) The page is judged as a typical phishing site; (3) After several redirectings, the final page is judged to be a typical “spam page”; (4) Apparent crowdturfing websites such as what were introduced in [67]. Finally, we end up with 124 manually-labeled malicious links: 79 spam ones, 30 irrelevant ads ones, and 15 pornographic ones. We also collect 214 benign links: 85 news ones, 70 blog ones, 49 video-audio ones, and 10 celebrity-related ones.

Table 2.1: Evaluation results for the list-based dataset

Set of features	Precision	Recall	F-Measure	ROC area
All 15 features	0.742	0.737	0.736	0.802
Click-based only	0.647	0.647	0.647	0.705
Posting-based only	0.648	0.695	0.694	0.756
Clicking statistics only	0.622	0.622	0.622	0.679

For each dataset, we construct the five posting-based features and the ten click-based features for all of the links. Then, we adopt the Random Forest classification algorithm (which has shown strong results in a number of spam detection tasks, e.g., [68, 53, 51]), using 10-fold cross-validation. The output of the classifier is a label for each link, either *spam* or *benign*. We evaluate the quality of the classifier using several standard metrics: average precision, recall, F-Measure and area under the ROC curve, equally-weighted for both classes.

2.4.2 Experimental Results

Classification on the List-labeled Dataset. For the first dataset, we report the evaluation results in Table 2.1. We find that using all features – both posting-based and click-based – leads to a 0.74 precision, recall, and F-Measure, and a ROC area of 0.802. These results are quite compelling, in that with no access to the content of the tweet nor the underlying web destination, spam links may be identified with good success using only behavioral patterns.

Next, we ask whether posting-based features or click-based features provide more power in detecting spam links. We first exclude the five posting-based features and report the *Click-based only* result in the table. We see even in this case we find a nearly 0.65 precision, recall, and F-Measure. When we drop click-based features in favor of a *Posting-based only*, we see a similar result. These results show that individually the two

Table 2.2: Top-10 features for list-labeled dataset (Chi-square)

Rank	Features	Score	Category
1	Median friends	277.43	Posting
2	Average clicks	199.11	Clicking
3	Median followers	159.53	Posting
4	Effective average clicks	150.72	Clicking
5	Click standard deviation	141.62	Clicking
6	Mean median ratio	141.49	Clicking
7	Max clicks	137.38	Clicking
8	Total clicks	120.67	Clicking
9	Rise & Fall	78.18	Clicking
10	Score function	66.50	Posting

feature sets have reasonable distinguishing power, but that in combination the two reveal complementary views of links leading to even better classification success.

We additionally consider the very restricted case of *clicking statistics only* (recall that our click-based features include both clicking statistics and clicking timeline features). Using only the seven click statistics, we observe only a slight degradation in quality relative to all click-based features.

To provide more insights into the impact of each feature, we use the Chi-square filter to evaluate the importance of features to the classification result. The top 10 features are shown in Table 2.2. Median friends and average clicks are the most two important features. Generally speaking, click-based features tend to play more important roles than posting-based features. Recall that our list-labeled dataset are those links with abundant clicks received, but it is not guaranteed that they have adequate posting counts, which may explain the ranking. For instance, if most links, either malicious or benign, have only one or two posting days and posting counts is less than 5, their posting counts and posting standard deviations will tend to be similar.

Table 2.3: Evaluation results for the manually-labeled dataset

Set of features	Precision	Recall	F-Measure	ROC area
All 15 features	0.860	0.861	0.859	0.921
Click-based only	0.828	0.828	0.828	0.888
Posting-based only	0.839	0.84	0.837	0.904
Clicking statistics only	0.842	0.843	0.841	0.875

Classification on the Manually-labeled Dataset. We repeat our experimental setup over the second dataset and report the results here in Table 2.3. When we use the complete 15 features, the precision, recall, and F-Measure are all even higher than in the list-labeled dataset case, around 0.86, with a ROC area of around 0.92. These results are encouraging. We attribute the increase in performance relative to the first dataset to the more expansive labeling procedure for the second dataset. In the list-labeled dataset, we only considered extremely "bad" links since we considered only the "Malware" and "Phishing" categories. This conservative assumption may lead to many spam-like links lurking in the set of benign links. In contrast, the manually-labeled dataset considers more broadly the context of what makes a spam link.

Continuing our experiments, we again consider subsets of features in the classification experiment. Again, we find that using only a single feature type – either *Click patterns only* or *Posting patterns only* – leads to fairly strong classification performance. But in combination, the two provide complementary views on links that can be used for more successful spam link detection.

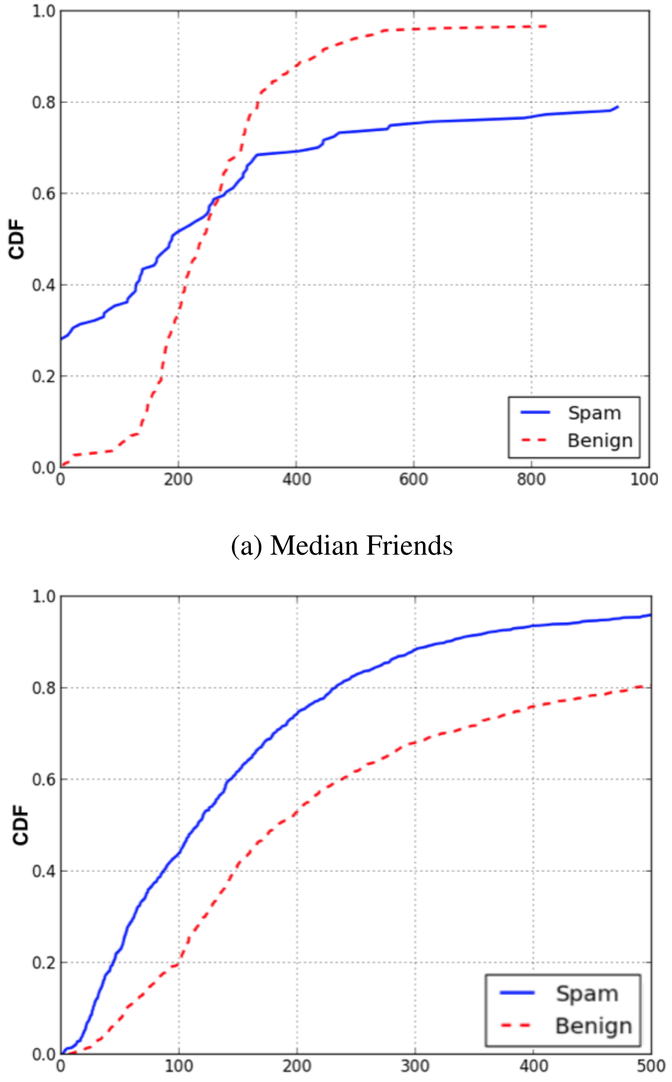
Again, we use Chi-square filter to rank features, as shown in Table 2.4. Interestingly, the ranking is quite different from what we found in Table 2.2, though again we observe a mix of both posting and click-based features. We attribute some of this difference to the click data's availableness in the manually-labeled dataset; most of the links in the

Table 2.4: Top-10 features for manually-labeled dataset (Chi-square)

Rank	Features	Score	Category
1	Average clicks	149.41	Clicking
2	Posting count	144.23	Posting
3	Median followers	123.24	Posting
4	Median friends	118.19	Posting
5	Score function	87.00	Posting
6	Posting standard deviation	63.66	Posting
7	Click standard deviation	59.27	Clicking
8	Max clicks	58.56	Clicking
9	Mean median ratio	54.17	Clicking
10	Clicking days	45.93	Clicking

manually-labeled dataset have abundant posting information and we can see that the posting behavior features play important roles in classification. On the contrary, most of the links in the manually-labeled dataset do not have very large clicking traffic to support clicking-based features. However, these two results – on the two disparate ground truth datasets – demonstrate the viability of integrating click-based features into spam link detection in social media, and the importance of integrating complementary perspectives (both posting-based and click-based) into such tasks.

To further illustrate the significance of click and posting-based features, we consider two of the top-ranked features in both datasets (recall Table 2.2 and Table 2.4): median friends and average clicks. We compare the distributions of these two strongly correlated features for all spam links and benign links, in Figure 2.4. For links in the list-based dataset, as in Figure 2.4a, around 20% spam links are posted by users with a median friends count of 0, and yet around 20% have a median friends count that exceeds 1,000. These two types of posters could correspond to newly-registered accounts (0 friend) and "high-quality" accounts like those in a for-pay campaign. In contrast, legitimate accounts



(a) Median Friends

(b) Average clicks

Figure 2.4: Example feature comparison for spam and benign links

who posted benign links have relatively “normal” distribution of median friends, that is, most have median friends less than 300 and almost none has a zero median. For links in manually-labeled dataset, as in Figure 2.4b, we see that spam links tend to have a lower average clicks. A potential reason is that malicious links require more exposure or other “abnormal means” to support consistent clicks, while legitimate links can survive longer due to their appealing contents. We find similar distributions for other click-based statistics, including the click standard deviation and the effective average clicks.

2.5 Summary

In summary, we investigate the potential of behavioral analysis aiding in uncovering spam links in social media. Purely by behavioral signals, we have considered two perspectives – (i) how links are posted through publicly-accessible Twitter data; and (ii) how links are received by measuring their click patterns through the publicly-accessible Bitly click API. The core intuition is that these signals are more difficult to manipulate than signals such as the content of a social media post or the content of the underlying destination page. Through an extensive experimental study over a dataset of 7 million Bitly-shortened links posted to Twitter, we find accuracy of up to 86% purely based on these behavioral signals. These results demonstrate the viability of integrating these publicly-available behavioral cues into spam link detection in social media.

3. COMBATING MANIPULATION: REVEALING ORGANIZED LINK SHARING BEHAVIOR*

In the previous section we explored the problem of detecting spam links via behavioral analysis, as a concrete application toward combating the misbehavior of spamming on social media. In this section, we turn to our second user misbehavior — manipulation on social media. In particular, we investigate the task of revealing organized link sharing behavior on social media and present a four-step approach, as follows.

3.1 Introduction

Link sharing is one of the most popular avenues to share information on Twitter. Users can enrich their inherently limited length postings by inserting a link pointing to an external resource such as a blog, video, or image. By doing so, many different viewpoints and additional context can be expressed through link sharing.

While some link sharing is legitimate (i.e., "*organic*"), other sharing behaviors are strategically manipulated (i.e., "*organized*") with a common purpose. And the boundary between these two extremes – organic and organized – is often not clear. Consider the following examples:

- Figure 3.1 shows three users who tweeted the same link `bit.ly/1dtous`, linking to a YouTube webpage related to the boy band One Direction. They all express their affections for the band in their tweets. It seems very likely they are fans of One Direction, which explains that they spontaneously posted the same link. This *common interest* in a subject related to a link leads to the coincidence of multiple

*Part of this section is reprinted with permission from "Organic or Organized?: Exploring URL Sharing Behavior" by Cao C., Caverlee J., Lee K., Ge H., and Chuang J. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15). ACM, New York, NY, USA, 513-522. Copyright 2015 by ACM.

organically posted links.

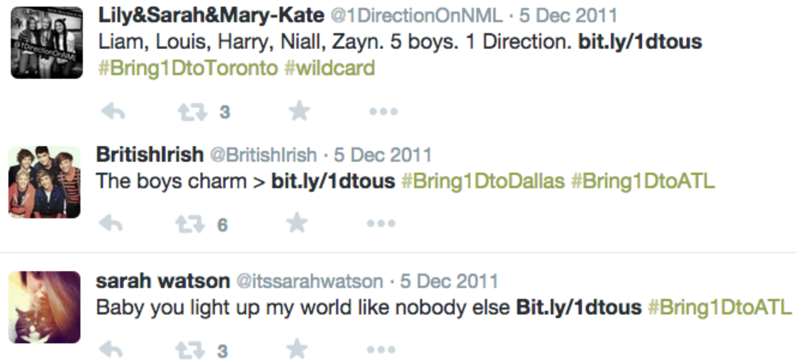
- Figure 3.2 shows four more users who posted the same One Direction YouTube link – `bit.ly/1dtous`. In this case, however, we can deduce that the users are participating in a voting campaign to attract the band to their hometown. The users are somewhat linked in this common desire.

- Continuing this theme, Figure 3.3 shows four additional users who have all tweeted a "vote" for Boston to attract One Direction. In this case, the voting behavior is suspicious: the tweets have highly similar text, and the latter three tweets were posted on the same day and the accounts names are quite similar. Were they organized to post the same link? Are these accounts controlled by the same person? Is the first account "innocent"?

- Finally, Figure 3.4 highlights three users who engage in a clear example of a somewhat sophisticated organized link spamming. Each user posts slightly different text and different appearing links, though ultimately all of the links redirect to the same destination link – an advertising webpage. This coordinated behavior of link sharing is fundamentally different from the first case of organic link sharing.[†]

These observations motivate us to investigate link sharing in social media. Our goal, *in the context of link sharing*, is to automatically (i) identify *user groups* in terms of similar link sharing behaviors; and (ii) differentiate strategically *organized* and genuinely *organic* user groups, through the development of a link-posting behavior based model. The key insight motivating our work is that the publicly available link-posting information can help model users with similar behaviors of link sharing, and that group-level behavioral signals can characterize a group of users as organic or organized. To purify and improve

[†] These accounts have been alive for more than two years, which suggests the official Twitter spam policy has limited impact on this type of manipulated link sharing behavior.

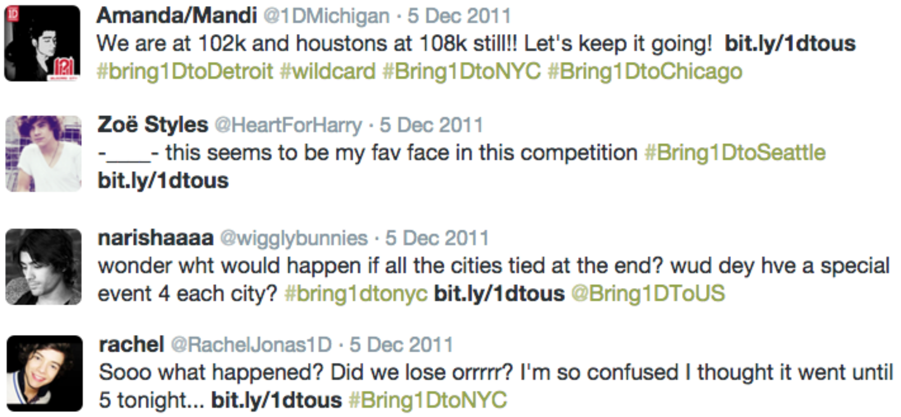


Lily&Sarah&Mary-Kate @1DirectionOnNML · 5 Dec 2011
Liam, Louis, Harry, Niall, Zayn. 5 boys. 1 Direction. **bit.ly/1dtous** #Bring1DtoToronto #wildcard

BritishIrish @BritishIrish · 5 Dec 2011
The boys charm > **bit.ly/1dtous** #Bring1DtoDallas #Bring1DtoATL

sarah watson @itssarahwatson · 5 Dec 2011
Baby you light up my world like nobody else **Bit.ly/1dtous** #Bring1DtoATL

Figure 3.1: One example of three users who have organically posted the same link: bit.ly/1dtous



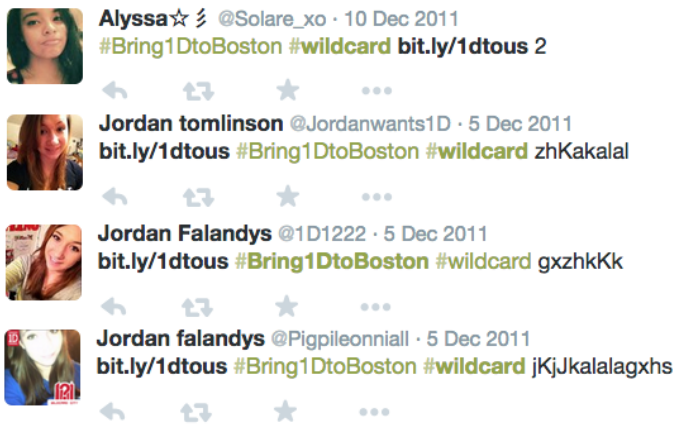
Amanda/Mandi @1DMichigan · 5 Dec 2011
We are at 102k and houstons at 108k still!! Let's keep it going! **bit.ly/1dtous** #bring1DtoDetroit #wildcard #Bring1DtoNYC #Bring1DtoChicago

Zoë Styles @HeartForHarry · 5 Dec 2011
-____- this seems to be my fav face in this competition #Bring1DtoSeattle **bit.ly/1dtous**

narishaaaa @wigglybunnies · 5 Dec 2011
wonder wht would happen if all the cities tied at the end? wud dey hve a special event 4 each city? #bring1dtonyc **bit.ly/1dtous** @Bring1DToUS

rachel @RachelJonas1D · 5 Dec 2011
Sooo what happened? Did we lose orrrrr? I'm so confused I thought it went until 5 tonight... **bit.ly/1dtous** #Bring1DtoNYC

Figure 3.2: Four users seemingly post link for a voting campaign



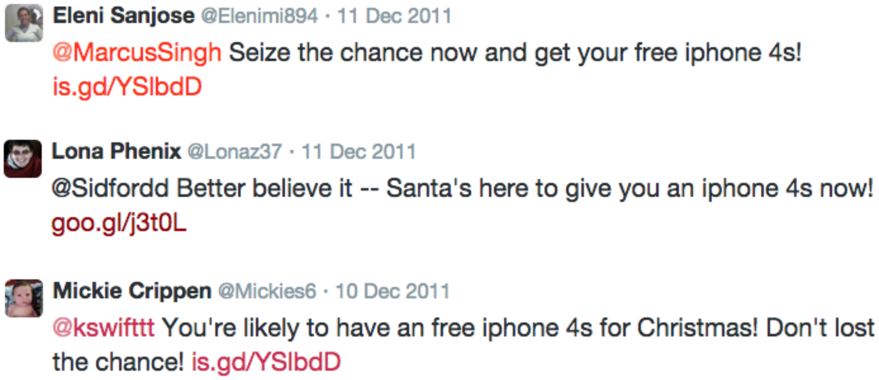
Alyssa☆ϟ @Solare_xo · 10 Dec 2011
#Bring1DtoBoston #**wildcard** **bit.ly/1dtous** 2

Jordan tomlinson @Jordanwants1D · 5 Dec 2011
bit.ly/1dtous #Bring1DtoBoston #**wildcard** zhKakalal

Jordan Falandys @1D1222 · 5 Dec 2011
bit.ly/1dtous #**Bring1DtoBoston** #wildcard gxzhkKk

Jordan falandys @Pigpileonniall · 5 Dec 2011
bit.ly/1dtous #Bring1DtoBoston #**wildcard** jKjJkalalagxhs

Figure 3.3: Four users suspiciously cooperate to post the same link


Eleni Sanjose @Elenimi894 · 11 Dec 2011
@MarcusSingh Seize the chance now and get your free iphone 4s! is.gd/YSlbdD

Lona Phenix @Lonaz37 · 11 Dec 2011
@Sidfordd Better believe it -- Santa's here to give you an iphone 4s now! goo.gl/j3t0L

Mickie Crippen @Mickies6 · 10 Dec 2011
@kswifttt You're likely to have an free iphone 4s for Christmas! Don't lost the chance! is.gd/YSlbdD

Figure 3.4: Three users coordinate to post the same advertising link

the information quality on their platform, it becomes imperative that the service provider can detect those organized behaviors of link sharing, such as campaign-like advertising, spamming, and other adversarial propagandas.

Given a group of organized users, on the one hand, we expect those users – no matter whether managed by a command-and-control structure or not – post links toward a common goal. We do not argue the goal has to be malicious, like the example in Figure 3.3. On the other hand, we focus on *group-level* behavioral evidence that can reflect their coordination. For instance, the users in Figure 3.2 were seemingly participating in the same "voting campaign", but they actually have different goals (the targeting cities) and their tweet content are quite distant. Even the users in Figure 3.3 seem suspicious, we need more evidence and should design a systematic framework for detection.

Concretely, we propose a four-step approach. We first formulate link sharing based on its three key factors: user, link, and the posting activity. Based upon such a model, we design a similarity measurement of user behaviors in link sharing. Then, given the pairwise similarity function, we build a user graph model from which we identify user groups each of which contains users with similar link sharing behaviors. Next, on the group level, we characterize the organic and organized user groups based on the link sharing behaviors of their members. Finally, we embed those characteristics into a classification framework to systematically distinguish organic and organized groups of users.

3.2 Related Work

Recent studies have started to investigate link sharing in social media, with different goals. One thread is about sharing intention, i.e., why people share links in social media. Suh et al. found that embedding links is one of the most important factors for increasing the *retweetability* of a tweet [69]. Smith et al. found that Twitter users add links to their tweets when discussing controversial topics, toward spreading information instead of conversing [70].

Another thread focuses on what people spread through links in social media, e.g., [15, 71, 72, 73]. These efforts have mainly focused on grouping similar messages or grouping users, such that the links provide additional context that may reflect the interests of the people posting these links.

Recently, several efforts have mentioned the dark side of link sharing in social media. Stringhini et al. examined spam campaigns on Twitter that posted messages with links pointing to the same site [74]. Grier et al. defined a spam campaign as the set of Twitter accounts that spammed at least one blacklisted link in common [75], and Gao et al. did similar things on Facebook [76]. Ghosh et al. studied the *link farming* on Twitter and found many participants are sybil accounts. Among those top link farmers, 79% have links pointing to their external webpages [7]. More recently, Nikiforakis et al. explored the ecosystem of ad-based URL shortening services, and the vulnerabilities made possible by these services [77].

Since 2006, there have been many social network-based sybil defense methods proposed such as SybilGuard [78], SybilLimit [79], and SybilInfer [80]. Viswanath et al. pointed out most of those approaches are essentially *graph partitioning* algorithms [81].

They all made certain assumptions of the social network topology, used ground-truth information of trusted users, ranked all users, and determined who are sybils based on some cutoff. Rather than exclusively focus on spammers or sybils, our interest is to reveal groups of strategically organized users who engage in link sharing with a purpose: some of the groups will post spam, but many others spread propaganda, aggressively promote products or services, and generally engage in coordinated manipulation. Unlike spamming or subverting reputation, the users we consider can be genuine and legitimate, as in Figure 3.3. Additionally, our problem is more general in the sense that our approach can detect those "link-posting based" sybils attacks or spam campaigns. We explicitly model and identify groups of users who have similar behaviors of link sharing, and differentiate the organized and organic groups via a group-level classification framework.

3.3 Methodology

In this section, we propose our approach to explore similar user behaviors in link sharing. In this context, our objective is to (i) formulate and collect user groups; and (ii) differentiate the organic and organized user groups. To tackle this problem, we formulate the concept of user group in the context of link sharing, and focus exclusively on behavioral signals. We are interested in answering the following questions: How do we model user in link sharing? How do we define and find user group? And how can we distinguish between organic and organized user group?

Toward answering these questions, our approach is structured in four steps. The main intuition is that the users from an organic group coincidently share similar interests on certain subjects so that they have similar behaviors of link sharing. On the contrary, organized groups consist of users who plot to post links with a certain goal in common so that their behaviors conform to a notion of *cooperation* or *coordination*.

- First, we model the user in link sharing, and design a similarity measurement to quantify similar behaviors between users.

- Second, we construct a user graph where nodes are users who have posted links and then identify user groups from the graph.
- Then, we extract group-level features to characterize organized user groups and organic user groups.
- Finally, we build a classifier for distinguishing organized and organic user groups.

We tackle each of these steps in turn, as follows.

3.3.1 Modeling the Behavior of Link Sharing

Link sharing on Twitter is fundamentally different from other popular activities such as *tweeting*, *re-tweeting*, and *following*. To systematically investigate users with similar behaviors of link sharing, we model the three factors in link sharing: the user, the link, and the action of sharing (posting). In the meanwhile, we need to pay attention on the following three issues. First, we should consider all links every user has ever posted on Twitter so that our model is general. Second, we should design a measurement that can quantify user similarity in link sharing. Third, it should take the link posting behavior into account. Such a measurement should be more specific towards link sharing than other traditional user similarity based on profiles, social neighborhood, tweet content, and so forth. Therefore, we aim to model users engaged in link sharing with two facets in mind: (i) which links a user has ever posted; and (ii) how (s)he posted them.

3.3.1.1 User, link, and posting

Our key idea is that, in terms of the posting behaviors, a user can be characterized by all the links he has posted and how often he has posted them. For instance, a user who likes sports tends to share more links linking to sports websites than links of politics websites. A

Japanese user is prone to post more links of Japanese websites than English websites. Two users who may have a realistic social relationship can have distinct tastes and preferences in link sharing, but their posting behaviors (especially the posting frequency of different links) may reflect such a scenario. It is even more prominent for strategically organized users. Those who have the same intention of link sharing can have quite similar posting behaviors, e.g., intensively posting certain link but barely mention others.

Formally, suppose we have a set of m users $\mathcal{U} = \{u_1, u_2, ..., u_m\}$. If u_i, in total, posted k different links $v_1, v_2, \ldots, v_k$, we define such an associated *link set* of u_i as $\mathbf{p}_{u_i} = \{v_1, v_2, \ldots, v_k\}$. By aggregating all users, we get a link set $\mathcal{V} = \{v_1, v_2, ..., v_n\} = \bigcup_{u_i \in \mathcal{U}} \mathbf{p}_{u_i}$. Therefore, each pair (u_i, v_j) can be seen as an action of posting a link.

We introduce the function $f(u_i, v_j)$ that quantifies such a posting of v_j by u_i. Here, we give a straightforward definition $f(u_i, v_j) = PostingCount(u_i, v_j)$ which represents the concrete *posting count* of every pair (u_i, v_j). Given a user $u_i \in \mathcal{U}$, we can represent u_i by an n-dimension vector $\mathbf{u_i} = (f(u_i, v_1), f(u_i, v_2), \ldots, f(u_i, v_n))$.

3.3.1.2 User similarity in link sharing

We have two considerations to design a user similarity measure in link sharing. On the one hand, if both users have posted overlapping links, the more intersections they have, the closer they are. On the other hand, we want to take the posting count into account. If two users posted the same link many times, we reward the similarity between them. If not, we penalize them if they have quite distant posting counts for the same link they have posted. Thus, we propose a measurement of user similarity defined as:

$$sim(u_i, u_j) = \sum_{k=1}^{n} \frac{ln(min(f(u_i, v_k), f(u_j, v_k)) + 1)}{|f(u_i, v_k) - f(u_j, v_k)| + 1} \tag{3.1}$$

We sum over all links to favor those pairs having posted many links in common. We pick the smaller posting count among two users ($min(f(u_i, v_k), f(u_j, v_k))$) as the pair-level scale of the posting count, and take the logarithm considering it can be quite large. We put the difference of posting counts as the denominator. We ensure a link can contribute towards the similarity only if it was posted at least once by both users, yet it's different from traditional "cosine-like" measures as we explicitly emphasize the penalization.

3.3.2 Identifying User Groups

Given the definition of user similarity, the next question is how to find those users having similar behaviors of link sharing. Given a set of users $\mathcal{U} = \{u_1, u_2, ..., u_m\}$, the task of *user group identification* is to find a collection of user groups $\mathcal{C} = \{\mathbf{c_0}, \mathbf{c_1}, \ldots, \mathbf{c_k}\}$ where $\forall \mathbf{c_i}, \mathbf{c_j} \in \mathcal{C}$, $\mathbf{c_i}, \mathbf{c_j} \subset \mathcal{U}$ and $\mathbf{c_i} \cap \mathbf{c_j} = \emptyset$.

The user similarity function can only locally measure the pairwise connection. Therefore, if we want to globally consider all possible users, adopting a graph structure is a natural choice. Extensive existing algorithms can partition a graph into *connected components*, which can fit our concept of user group here. In general, a *user graph* $G = (V, E)$ can be defined by $V = \mathcal{U}$ and $E = \{(u_i, u_j) | \forall u_i, u_j \in \mathcal{U}, weight(u_i, u_j) = sim(u_i, u_j)\}$. However, if we consider all possible user pairs, the resulting graph can be huge: it will have only one connected component where most connections are weak. Hence, we need a specific version of a user graph upon which we can extract our interested user groups.

3.3.2.1 The kNN user graph

If we simply set a global threshold to filter out low-weight edges, we may lose important information. For example, the users from an organized group unnecessarily (and usually do not) post many links of popular websites — they have their own targets to spread. As a result, those users may be excluded from the graph due to the lack of overlapping links with others, and finally we may obtain only a few big components in which people share popular websites.

Since we are interested in those organized users, our graph model should be able to grasp those “abnormal connections”. Organized users do not often share various links, whereas the nature of coordination leads to their locally firm neighborhoods within the group. To retain such “conspiracies” as much as possible, we require a model that emphasizes the mutually steady neighborhood. Thus, we adapt the model of *k nearest neighbor (kNN)* graph.

The kNN user graph connects u_i and u_j only if u_i is among the k-nearest neighbors of u_j. Such a restriction retains only those k strongest neighborhoods, with less emphasis on the edge weight. A formal definition of the kNN user graph $G = (V, E)$ is $V = \mathcal{U}$ and $E = \{(u_i, u_j) \,|\, \forall u_i \in knn(u_j), \forall u_j \in knn(u_i)\}$. Here, $knn(u_i)$ is a function defined as $\{u_j | \forall u_j \in neib(u_i), sim(u_i, u_j) \in max_k(\{sim(u_i, u_j) | \forall u_j \in neib(u_i)\})\}$, where $neib(u_i)$ is the set of u_i's neighbors and $max_k(\mathcal{S})$ returns the k largest elements given $\mathcal{S}$. Now, since any node can have k neighbors, a group of users who post unusual links can still form a big component.

3.3.2.2 Extracting user groups

It is a non-trivial task to extract a collection of user groups from the user graph. The concept of *connected component* in *graph theory* naturally matches our concept of user group, but we need two more considerations. First, we require that every group should have a compact size. Hence, we discard those small connected components (e.g., fewer than five members), and decompose those large components into smaller ones. Second, we hope the partition algorithm to be general, i.e., having been proved effective and efficient on many different types of graphs. Thus, we choose the well-known *Louvain method* [82], which is one of the most widely-used algorithms in *community detection* [83].

The Louvain method is a greedy algorithm to maximize the *modularity* of a graph structure. It starts by locally optimizing modularity for small communities, and then iteratively repeats to aggregate until reaching a maximum of modularity. Though the modularity optimization problem is known as NP-hard, the Louvain algorithm can run in $O(n \log n)$ in most practical cases so that it already has many mature implementations. Its maximum modularity ensures it conservatively segments big components only when it does contain multiple modularities.

3.3.3 Characterization: Organized vs. Organic

Given a collection of user components (groups) that we have extracted, our ultimate goal is to systematically distinguish organized and organic user groups. But before that, we need to characterize these two types of groups. Suppose we have a group of users who have similar behaviors of link sharing, we want to feature to what degree they have organized behaviors when posting links.

First, we should ensure our features build upon the group level. Our biggest interest is the collective user behavior, which is fundamentally different from individual spamming or sybil behavior. Second, our features should target the behavioral signals of posting. We have seen in Figure 3.4 how easily the posters manipulated their tweet content to disguise themselves. Moreover, they unnecessarily follow each other as long as they post the same links. So, our insight here is those traditional features based on either text content or network structure become vulnerable for the organized user group in our context. On the contrary, what they cannot cover are the links they posted (they hope more exposures after all), the time-stamps they posted (they have to leave the records anyway), and their own profiles (they use public profiles so that others can see their postings). Therefore, in this section, we introduce nine group-level features that cover three posting-related aspects: *posted links*, *posting time*, and *poster profile*.

3.3.3.1 Posted link based features

Our motivation is that, intuitively, the users in an organized group usually have a clear goal of promoting certain links. Thus, they have a relatively narrow selection of links they post, and each link gets high-volume postings. Instead, an organic group usually posts a variety of links each of which has reasonable amounts of exposure. Therefore, we come up with two group-level features that capture the diversity of the posted links.

Suppose we have a user group $\mathbf{c} = \{u_1, u_2, \ldots, u_k\}$, and we know the set of links posted by every member, i.e., $\mathbf{p_{u_i}}$. Thus, we can extend it to the group level, as well as the posting count $f(u_i, v_j)$. Both can be formulated as the following:

$$\mathbf{p_{c_i}} = \cup_{u_j \in \mathbf{c_i}} \mathbf{p_{u_j}} \tag{3.2}$$

$$f(\mathbf{c_i}, v_j) = \sum_{u_i \in \mathbf{c_i}} f(u_i, v_j) \tag{3.3}$$

Based upon these definitions, we provide the following two *posted link* based features:

- Average Posting Count. We can calculate the *average posting count* per link by the ratio of $|\mathbf{p_c}|$ and $\sum_{v_j \in \mathbf{p_c}} f(\mathbf{c}, v_j)$. By our motivation explained above, we expect organized groups have higher values of such feature than organic ones.

- Link Posting Entropy. Entropy is an important measure of *uncertainty* in *information theory*. Here, in our case, to describe the diversity of posted links in a group, we extend it to *link posting entropy*, computed as:

$$H(\mathbf{c}) = -\sum_{v_j \in \mathbf{p_c}} \frac{f(\mathbf{c}, v_j)}{\sum_{v_j \in \mathbf{p_c}} f(c, v_j)} log \frac{f(\mathbf{c}, v_j)}{\sum_{v_j \in \mathbf{p_c}} f(c, v_j)} \tag{3.4}$$

  And based on the same idea, we suppose that organic groups have larger link posting entropy than organized ones.

3.3.3.2 Posting time based features

One of the most important posting behavioral signals is the *posting timestamp* of every tweet. The poster has no access to tamper with such information, making it a potentially robust feature. For each poster, we can collect a *posting time series* so we can compute all the *posting intervals*, defined as the temporal differences between consecutive posting timestamps. Our motivation here is: with the goal of promoting or advocating, the users from an organized group usually post tweets in a similar frequency, and the intervals tend to be short as they are eager to rapidly increase the exposures of their links. Therefore, we propose two group-level posting interval related features, measuring both the quantity and the deviation.

- Posting Interval Median. For every user, we can always quantify the posting interval by some temporal unit. Then we get the *individual posting interval median* by taking the median among all intervals. We use median rather than mean mainly because it is more robust to outliers. Moreover, since we care more about the group level, we take a further median over all members in a group. As mentioned, we expect organized groups have shorter interval medians than organic ones.

- Posting Interval Deviation. As said, our inference is that the organized accounts have similar posting manners, or even are bots manipulated by the same person. Thus, we can compute the group-level *posting interval deviation* given the individual interval median. Since an organic group is likely to post more randomly, we expect organic groups have larger deviations.

3.3.3.3 Poster profile based features

Compared to an organized group, our main motivation here is that the users in an organic group have more various demographics. Given a group of organized accounts, if their goal is improper (e.g., advertising, spamming), they are mostly managed (like sybils) or hired (like for-pay turkers) by the same agent so they have close demographics. Even if their goal is relatively legitimate (e.g., propaganda, voting), their conspiracy attributes their enthusiastic interests on some common subjects, which can be reflected on their profiles to certain degree, too.

The profile information is one of the best publicly-available resources we can utilize to infer the diversity of demographics. We reify it into three aspects: total number of posted tweets, account registration date, and followers (friends) count. We calculate the group-level deviation of the total number of tweets, of the followers counts, and measure the average interval of the registration time.

- Tweets Counts Deviation. We count the total number of tweets an account has posted ever since the beginning, and take the deviation among all accounts in a group. The larger deviation means the more variety, so we expect an organic group has a larger *tweets counts deviation* than an organized one.

- Followers Counts Deviation. We record the count of followers for every user in a group. The count can be dynamic over time so we take the median. Then we take the deviation among all members. We believe the count of followers is more reliable than of friends simply because it is more difficult to fabricate. Again, we expect organized groups have lower deviations.

- Registration Interval Median. It is similar to how we computed the *posting interval median*. We look into the registration date of each user so that we have a *registration time series* for every group, and then we take the median. Our idea is that the registration interval is one of the most direct evidences for those fabricated or hired accounts. We prefer median to mean because the latter one is too sensitive to outliers. We expect organized groups have smaller *registration interval medians* than organic groups, similar to the reason for the posting interval median.

Then, we come up with two more features related to the registration date. We define the *user lifespan* by counting the temporal span, in terms of days, from the day an account registered to his latest posting date in our dataset. The motivation here is that the user lifespan is one of the most important user profile information and it should be quite random for users from an organic group. Instead, the organized users are created for a certain link-promotion mission, so they all tend to have short lifespans. Another idea is that Twitter may already have detected the suspicious behaviors from organized users and suspended them, leading to short lifespans too. Hence, we provide two features to characterize the group-level user lifespan via the quantity and the deviation.

- Poster Lifespan Median. We take the median of all accounts' lifespans in a group. As said, we suppose an organized group generally has a shorter *poster lifespan median*.

- Poster Lifespan Deviation. Every member from a group has a lifespan so we can calculate the group-level deviation. We expect organic ones have larger *lifespan deviations*.

3.3.4 Classification: Organized vs. Organic

Recall that our ultimate goal is to automatically discern organized and organic user groups, in the context of link sharing. Given the features in the previous section, it becomes quite natural that we embed them into a classification framework. To choose appropriate classification algorithms, we should guarantee: (i) the algorithm has been widely used and maturely implemented; and (ii) we need to test on multiple algorithms.

We choose 4 well-known classification algorithms in our framework: *Random Forest* [84], *Naive-Bayes Decision Tree* [85], *Sequential Minimal Optimization* [86], and *Additive*

Logistic Regression [87]. We notate them *RandomForest*, *NBTree*, *SMO*, and *LogitBoost*, respectively. NBTree is a *decision tree learning* algorithm in which the tree leaves are naive Bayes classifiers. SMO algorithm is used for training the support vector classifier and has been implemented in many existing SVM libraries like LIBSVM [88]. LogitBoost can be seen as a variant of AdaBoost [89] that adapts *logistic regression* techniques. All of them have mature implemented packages. With different theoretical foundations, these four candidates can well-serve the testing algorithms in our experiment.

3.4 Experiments

We present our experimental studies in this section. We first introduce our data. Then, we describe how we identified all the user groups we formulated. Third, we provide details how we collected the ground truth. Finally, we show our analysis results towards distinguishing organized versus organic user groups.

3.4.1 Data

We deployed a tweet crawler via the official Twitter Streaming API from October 2011 to October 2013. Since our main interest is link sharing, we only collect tweets that contain links. The API provides a 1% sample of all published tweets, but our 24-month uninterrupted crawling gives us 1.6 billion *"raw URLs"* posted by 136 million accounts. Raw URLs are those URLs in their original format when posted, without any further processing (e.g., resolution) after being crawled. Due to either irregular typing or the URL shortening service, many raw URLs become inaccessible or actually link to the same webpage. For instance, we have both `WWw.TwiTtEr.com` and `tWITTER.com/` in our dataset, as well as `bit.ly/1a8jUOr` and `bitly.com` both of which direct to the same destination.

To address such an issue of *URL variants*, we need to resolve for all URLs. Since resolving billions of URLs can be expensive, we focus on URLs that appeared at least 50 times. We resolve through standard HTTP requests and record the landing long links. The summary of our dataset is shown in Table 3.1: 47 million accounts have shared 1.6 million raw URLs within 445 million tweets. 82% raw URLs get resolved to nearly 1.2 million distinct long links. 869 thousand accounts have generated at least 50 postings.

Table 3.1: Dataset Summary

# Raw URLs (resolved)	# tweets	# long links	# domain names	# Posters (with $\geq$ 50 postings)
1,617,234 (1,327,729)	445 million	1,199,930	166,107	47,658,839 (869,571)

Given a resolved long link, we ignore all its *post-parameters* as a reasonable approximation to the link. We call the remaining part *domain name* and we obtain 166,000 unique ones. Compared to a complete link, we believe the domain name has better interpretability because it can conceptually represent a “website”. For users, we exclude those who just occasionally share links, i.e., less than 50 postings. To model the user, we decide to also use link’s domain name as the dimension. One reason is that using original long link may result in extremely sparse user vector given enormous dimensions. Another reason is a user group can be better interpreted if each member corresponds to some “website” instead of a long HTML page link.

3.4.2 Collecting User Groups

To construct the kNN user graph, a non-trivial issue is how to choose an appropriate k. We adopt the idea in [90] to pick k roughly equal to $ln(|\mathcal{U}|)$. Thus, 869,571 users give us k of 14. In the end, we obtain a user graph containing 216,523 nodes and 3,862,116 edges. This graph contains 12,251 connected components most of which are small: only 2,150 components have no less than 5 nodes and just 36 components are bigger than 100. We filter out those tiny components smaller than 5, and exhaustively decompose (if possible) large ones bigger than 100 to ensure maximal modularity. Eventually, we identify 2,775 groups, together including 192,719 users. Among those 2,775 groups, we find around 40% groups are smaller than 10, and nearly 90% groups are bounded by 100 users. The largest one has 14,080 users.

How can we find a way to see whether our identified user groups are “meaningful”? And how to interpret and measure it here? Naturally, the most direct evidence is our groups maintain closer manners of link sharing behaviors than “a random user group”, yet we need a way to measure it. The entropy of posted links in a group is one of the most typical properties that capture the similar link sharing behavior, as explained in Section 3.3.3. Thus, first, we simulate a collection of user groups with the exact same sizes via randomly picking users from our dataset. Then, we compare the distributions of the link posting entropy for our collected groups and the simulated ones. The result is in Figure 3.5, and we clearly see the difference. Around 20% of our groups have a zero entropy, i.e., all the members posted the same link all the time, and the median is about 1.5. On the contrary, none of the simulated groups have a 0 entropy, and the median is around 4 while 80% of our groups are below it.

Besides the natural evidence from those link-posting related features, we hope to see more different evidence showing our collected groups are meaningful. Here, we adopt the group-level *language usage entropy* to measure certain “homophily” among all members in a group. The intuition is that if a set of users have similar selections of links (embedded in their tweets), then their language choice in their tweets should be similar. Conversely, two users who have distinct language backgrounds would hardly overlap many links. Therefore, for each group, we aggregate all the published tweets (with or without links, having valid language usage information), count the usage frequency, and calculate the entropy. To compare, again we simulate the user groups with the exactly same sizes. We compare their distributions in Figure 3.6, and find the contrast is apparent.

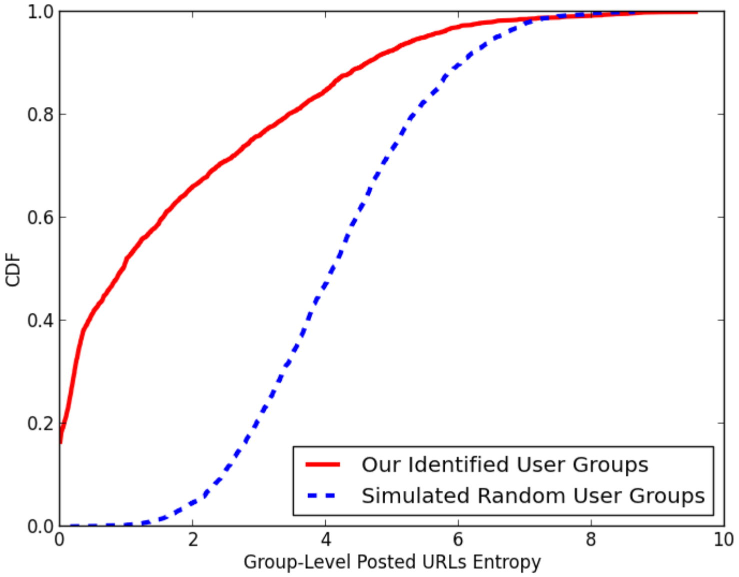


Figure 3.5: The link posting entropy CDF of our collected user groups, compared with a collection of simulated random groups

About 30% groups in our collection whose language entropy is 0, i.e., all members always use the same language writing tweets. 90% of our groups have entropies less than 2.0. In contrast, in the simulated collection, almost none have 0 entropy, and the median is between 1.5 to 2.0 where 80% of our groups are below it. This comparison shows the users in our identified groups have much similar language usages. Recall that our method of computing user groups has not used any user information on language usage, which demonstrates the potential of extending our approach into general problems related to user similarity.

3.4.3 Ground Truth

To test our proposed features in Section 3.3.3 on characterizing organized and organic groups, we need a set of groups with known labels of either organized or organic. Since there is no such existing ground truth, we randomly pick 1,000 of our identified groups,

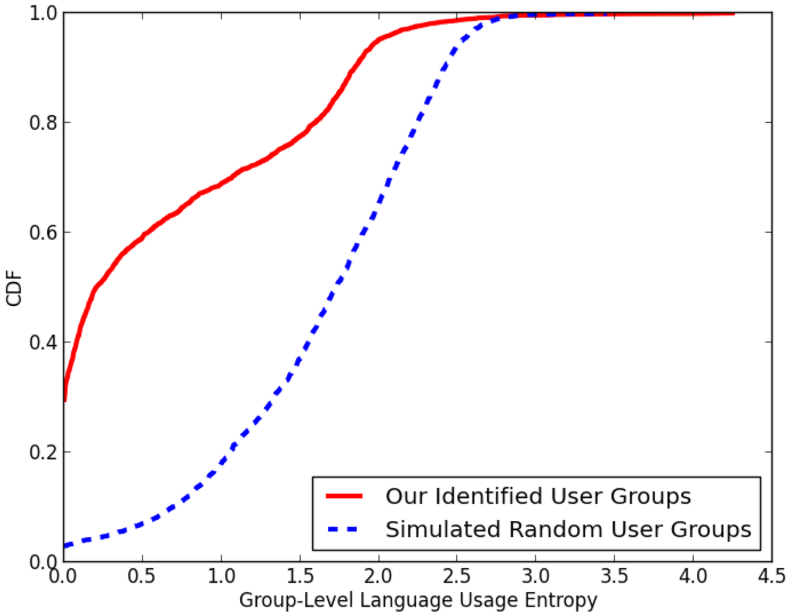


Figure 3.6: The language usage entropy CDF of our collected user groups, compared with a collection of simulated random groups

and manually check each of them as follows.

3.4.3.1 Manual labeling setup

Our labeling for each group is a two-tier task: (i) categorizing the content of the links the members posted and inferring the purpose that they posted links; (2) and rating to what degree we think their behaviors were organized or organic.

The rating task directly connects to our interest here, but the categorization task can help us interpreting the similar user behaviors. Our focus is to collectively seek for group-level evidence that reflects the coordinated behaviors of posting links, not just individually inspecting each account. In particular, we examine the accounts in each group through looking into their: (1) tweets (e.g., all the posted links, landing webpages, and tweets that contain links); (2) account profile (e.g., self-introduction, name, avatar, language, geo location); and (3) posting pattern (e.g., the URL shortening pattern, textual pattern in tweets, posting timeline pattern). If a group is too large, we randomly pick at least 5 users with accessible information to judge. Judges make decisions without knowledge of the proposed features in Section 3.3.3.

To measure to what extent a group is coordinating can be subtle. Our ratings will be the scores between 1 to 5, the larger the more suspicious towards an organized group. Then, we transform our scores to the label of either organized or organic. It is an organized group if its score is above 3, and organic if below 3. If its score is 3, we inspect it again. By such a 5-scale rating, we avoid curtly labeling a group is organized or organic.

We assign 3 annotators the exactly same set of 1,000 user groups, and separately ask them to manually label each group as organized or organic. Since the number of our annotators is odd, if the individual decisions of our 3 annotators have a clear "major voting", we take it as the final result. The rating usually has an obvious favor (below or above 3), and we accept a category if it has been mentioned by at least two annotators. Otherwise, a fourth annotator adds a label and a rating. We finalize a category or rating only when it has two or more endorsements, and count a user group has valid labels only if it finally obtains at least one category or rating.

3.4.3.2 Categorizing a user group

As said, our categorization aims to infer the common intent of a group of users who have similar link sharing behaviors. The idea is that organized groups have a much clearer goal in common than organic groups, even though the purpose can be either improper (e.g., advertising, spamming) or legitimate (e.g., self-promoting, preaching). In practice, we do find sometimes speculating the intention is not obvious, so we change to summarize the content of posted links and tweets in this case. In the meanwhile, we find usually categorizing a group is more difficult than rating it, especially if the members post quite various subjects or many non-sense words. Finally, among all the 1,000 randomly picked groups we have judged, 773 ones have been labeled by at least one of the 12 categories in Table 3.2.

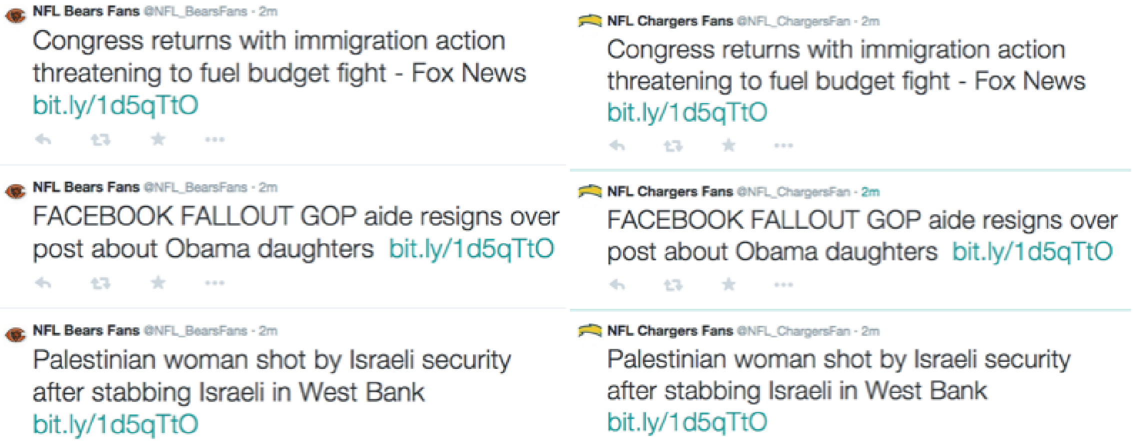
Table 3.2: The distribution (percentage) of twelve categories that we have labeled for our user groups

advertising	spamming	app-auto-generated	entertainment	social-media	adult
318 (41%)	195 (25%)	171 (22%)	66 (8%)	33 (4%)	22 (3%)
news	blog	follow-back	public-information	propaganda	voting
20 (3%)	10 (1%)	8 (1%)	6 (0.8%)	6 (0.8%)	4 (0.5%)

The category of *advertising* emphasizes the intention of posting links. It includes everything about absorbing viewer's attention on the link, such as marketing, funds-raising, or even self-promotion of personal websites or uploads. We have seen instances that all members in a group shared links connecting to the exactly same news articles or blogs. We still think such a group is suspicious to advertising. Moreover, the irrelevance between the content of tweets and the linked webpages can also indicate the poster is advertising, and one example is shown in Figure 3.7.

We make the assumption that Twitter's official suspension indicates *spamming* activities to some extent. In addition, if the browser or the shortening service warns when we click into a link, we believe it is a spam link. Other than that, we label spamming if the links reaching to some typical spam webpage of "phishing" or "malware". For instance, we categorize a group full of users like the left-hand two in Figure 3.8 into spamming, because their posted links point to the phishing website shown in the rightmost subfigure.

App-auto-generated is a special category. All users in such a group posted tweets (containing links) automatically generated by some external app or service. Usually it happens when the poster is unconscious or at least not intentionally doing so, e.g., using



NFL Bears Fans @NFL_BearsFans · 2m
Congress returns with immigration action threatening to fuel budget fight - Fox News bit.ly/1d5qTtO

NFL Chargers Fans @NFL_ChargersFan · 2m
Congress returns with immigration action threatening to fuel budget fight - Fox News bit.ly/1d5qTtO

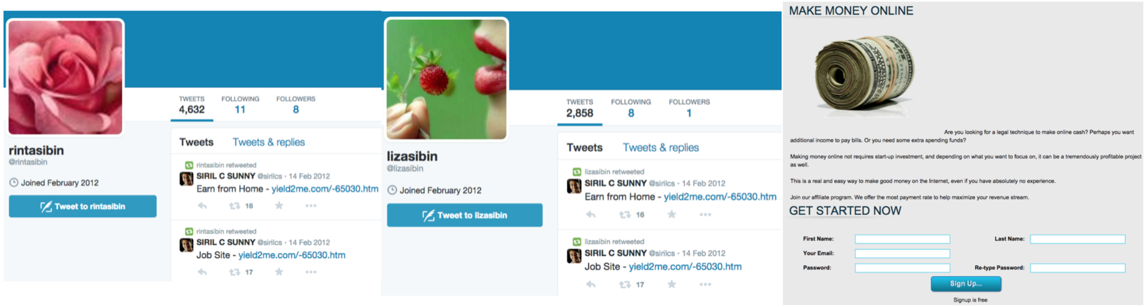
NFL Bears Fans @NFL_BearsFans · 2m
FACEBOOK FALLOUT GOP aide resigns over post about Obama daughters bit.ly/1d5qTtO

NFL Chargers Fans @NFL_ChargersFan · 2m
FACEBOOK FALLOUT GOP aide resigns over post about Obama daughters bit.ly/1d5qTtO

NFL Bears Fans @NFL_BearsFans · 2m
Palestinian woman shot by Israeli security after stabbing Israeli in West Bank bit.ly/1d5qTtO

NFL Chargers Fans @NFL_ChargersFan · 2m
Palestinian woman shot by Israeli security after stabbing Israeli in West Bank bit.ly/1d5qTtO

Figure 3.7: Example users whose group is organized



rintasibin
@rintasibin
Joined February 2012
Tweet to rintasibin

TWEETS	FOLLOWING	FOLLOWERS
4,632	11	8

Tweets Tweets & replies

rintasibin retweeted
SIRIL C SUNNY @sirilcs · 14 Feb 2012
Earn from Home - yield2me.com/-65030.htm
16

rintasibin retweeted
SIRIL C SUNNY @sirilcs · 14 Feb 2012
Job Site - yield2me.com/-65030.htm
17

lizasibin
@lizasibin
Joined February 2012
Tweet to lizasibin

TWEETS	FOLLOWING	FOLLOWERS
2,858	8	1

Tweets Tweets & replies

lizasibin retweeted
SIRIL C SUNNY @sirilcs · 14 Feb 2012
Earn from Home - yield2me.com/-65030.htm
16

lizasibin retweeted
SIRIL C SUNNY @sirilcs · 14 Feb 2012
Job Site - yield2me.com/-65030.htm
17

MAKE MONEY ONLINE

Are you looking for a legal technique to make online cash? Perhaps you want additional income to pay bills. Or you need some extra spending funds?

Making money online not requires start-up investment, and depending on what you want to focus on, it can be a tremendously profitable project as well.

This is a real and easy way to make good money on the Internet, even if you have absolutely no experience.

Join our affiliate program. We offer the most payment rate to help maximize your revenue stream.

GET STARTED NOW

First Name:
Last Name:
Your Email:
Password:
Re-type Password:

Sign Up...

Signup is free

Figure 3.8: Example users whose group we categorize into "spam" and think is organized

own Twitter account to log in some mobile game. For *entertainment*, we find groups of fans who posted links about their supported artists like we have seen in Figure 3.1, 3.2, 3.3. Sometimes they can be *voting* (like in Figure 3.2 and 3.3), but many are just bonded by fan's passion. Other entertainment groups often talked about music, video game or anime. *News* are mostly about politics and technology. *Propaganda* usually relates to politics and religion.

Social-media and *blog* stress the source website. These websites can host various webpages, and typical examples are `youtube.com`, `instagram.com`, and `blogspot.`

com, etc. This also explained we found many such groups. The groups of *public-information* posted links of public information like traffic or weather. The posters usually are accounts managed by regional institutions.

Among all these 12 categories, some are capturing the purpose of posting (e.g., advertising, spamming, propaganda), and some are summarizing the content of posting (e.g., entertainment, social-media, news). A user group can have more than one category, e.g., mixed with spam, adult, and advertising. In total, from Table 3.2 we can see advertising and spamming together take up 66% appearances of all categories. This observation reflects our idea that the kNN graph model, emphasizing the locally pairwise connection, can potentially retain many abnormal behaviors.

Another interesting finding is the category of app-auto-generated surprisingly occupied 22%. Its substantive existence leads to a new thread for future work: the *user-unconscious* posting behavior in social media, caused by some external application. What security issues could it raise? Do most posts contain links? If yes, can our study of link sharing be an entry point?

3.4.3.3 Rating a user group

For rating, what has been posted matters less, and we care more on who have posted and how. As mentioned, we look over the profiles of all accounts from the same group. We seek for similar patterns between them in self-introduction, name, avatar picture, etc. We rate above 3 if we see highly overlapped textual patterns in their tweets such as the posting date, hashtag, language and URL shortening. Figure 3.7 includes two accounts with identical posting timeline, tweets content, and similar names. Their group is the instance that we rate 5. It becomes even more dubious when the members always retweet each other's tweets. A typical "rate-5" example is in Figure 3.8. The users, who have extremely close posting timelines and names, always retweeted from the other account.

Our rating focuses on the collective behavioral evidences of coordination, independent with the posting purpose. For example, we find most groups of entertainment are users spontaneously talk about the subjects they like, but it becomes suspicious when the URLs link to the same music video page on Youtube. Most groups of public-information have fairly legitimate goal of posting, but naturally many accounts are centrally managed so they still have many common patterns.

If all members in a group have been suspended (and we label it as a spamming group), we cannot access their accounts so we conservatively leave them unrated. For those groups of app-auto-generated or social-media, we mainly judge their postings are whether advertising-oriented or unconscious, and we usually find most of them fall into the latter scenarios. The groups of news, social-media or blog often share popular news, blog article, and online forum. We rate them low except we find they advertised their own-related (person or institution) webpages.

3.4.4 Experiments: Organized vs. Organic

3.4.4.1 Analyzing our labeling

As introduced, we finalize the labeling result for each user group via adopting the major vote over our 3 annotators. We find 986 groups that have been labeled at least one of the categorization and rating results, and 602 ones that received both. The fourth annotator agreed on 871 (88.3%) and 520 (86.4%) ones, respectively. Finally, we obtain 815 groups with ratings where 325 (40%) are organized and 490 (60%) are organic.

Those 602 groups with both information of category and rating give us the opportunity to understand the following questions: Which subjects our collected groups mostly talked about? What kinds of content the organized groups and organic groups usually posted? Are they different? Can we see a relation between the levels of organized behavior and the types of link?

We aggregate the average rating each group received by the group category, and plot the CDFs of our ratings for all 12 categories in Figure 3.9. We can see the distributions take on clear gradients. If we look at the median, 12 categories are evenly divided by the rate of 3. The groups of spamming, adult, and follow-back are most likely as organized, followed by advertising, public-information, and propaganda whose distributions are quite close. Then, blog, voting, entertainment and news have similar distributions, most of whose groups are organic. Social-media and app-auto-generated groups are the least likely organized (recall how we rate them). These observations reveal the connection between the level of organized behavior and the inherent content of different categories. For spamming, adult, follow-back, advertising, they essentially include those improper ingredients that motivate more organized behaviors. On the contrary, the category of app-auto-generated is user-unconscious, and blog, news, entertainment and social-media naturally contain more legitimate activities, so their groups tend to be more organic.

3.4.4.2 Classifying organized and organic groups

We are ready to build the classifier. In terms of posting behaviors in link sharing, can our proposed group-level features distinguish the organized and organic groups? Which features work best? Which classification algorithm performs best?

To address the issue of class imbalance, we give a hybrid solution that combines undersampling of the majority class and oversampling the rare one. In the end, we have 406 organic and 406 organized groups, and we normalize the values of each feature to the interval of 0 to 1. To evaluate, we do 10-fold cross validation and focus on precision, recall, F-measure, and ROC area. We first consider the two classes are equally important and take the average. The results are in Figure 3.10.

The performance achieves around 0.8 no matter which method or measure we choose. RandomForest works best with high F-measure (0.836) and ROC area (0.921). All the

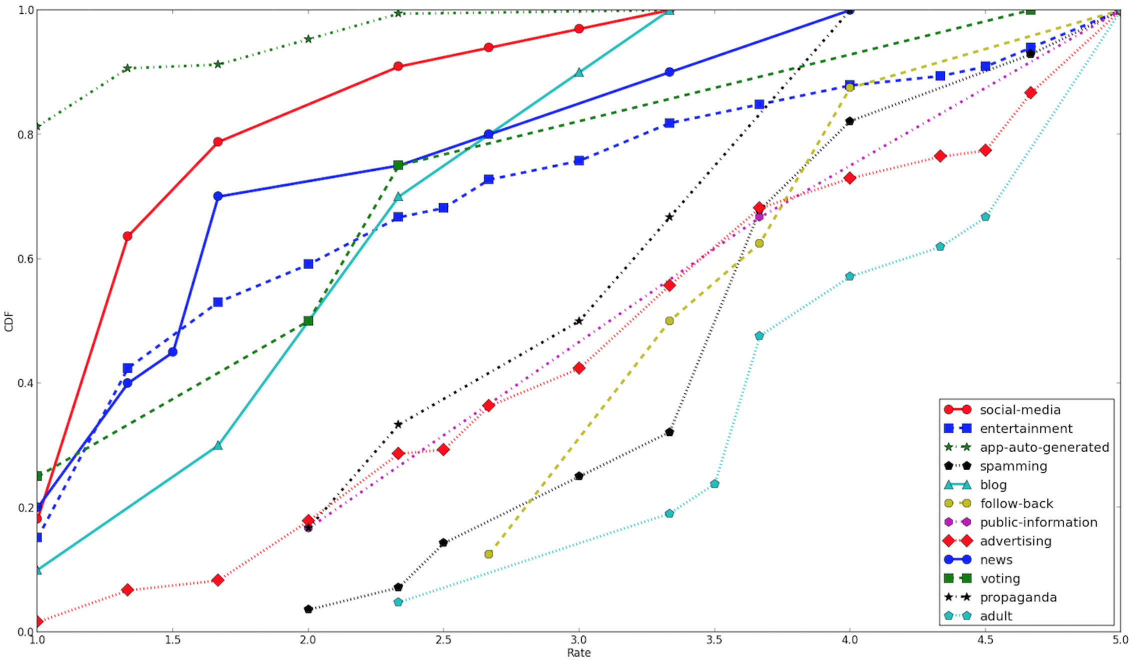


Figure 3.9: The CDFs of our ratings for all 12 types of group categories

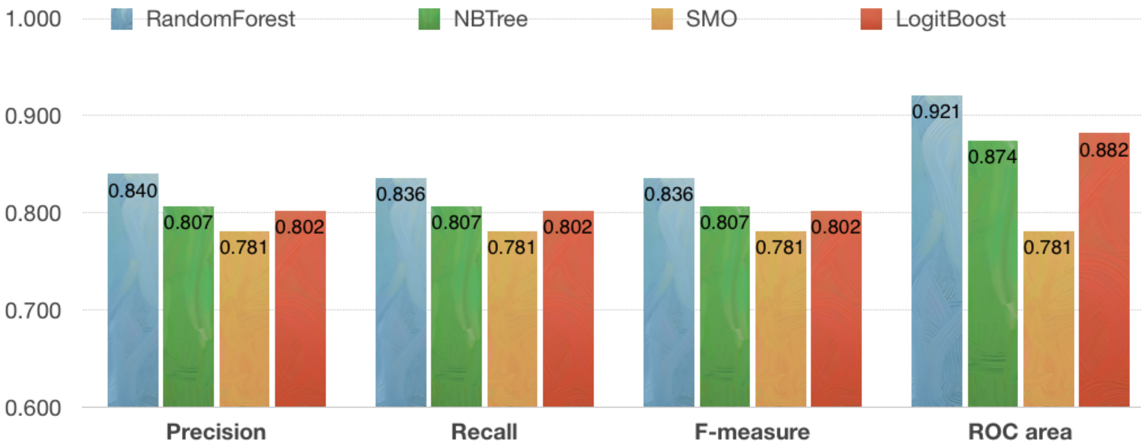


Figure 3.10: Evaluation results by four classification methods

results suggest the potential of our approach for distinguishing organized and organic user groups, purely based on behavioral signals in link sharing.

In reality, detecting organized user groups becomes more important. We especially want to find out organized groups as many as possible, corresponding to the measure of recall. Thus, we further show the recall result for the class of organized in Figure 3.11. We see those two decision tree based algorithms outperform the other two, and even better than their own averaged results in Figure 3.10. This observation hints us we can prioritize decision tree based algorithms when we solve such a problem in practice.

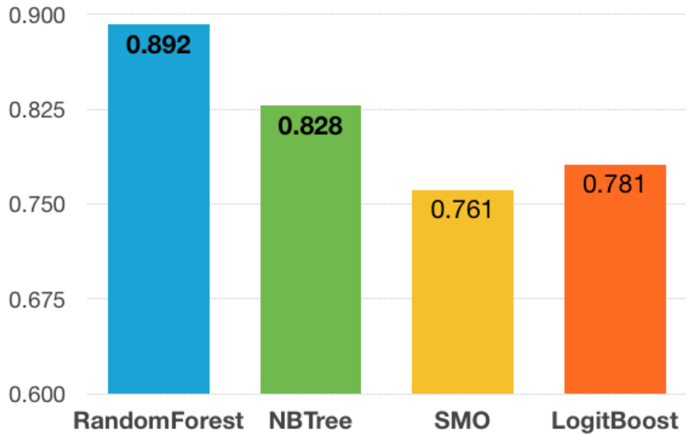


Figure 3.11: The recall results for the class of organized group by four classification methods

Another investigation is the feature impact. We select two popular measures — *Chi-squared* and *Information gain* — to evaluate each feature with respect to the class. The full ranking is in Table 3.3. We see the rankings by chi-square and information gain are almost identical. We have 3 interesting discoveries. First, the two link based features always rank in the top 3. This suggests the link-related property is the most reliable aspect when we study link sharing. Second, if we have two features from the same type (e.g., lifespan deviation and median), the deviation feature always has more influence than the median one. One possible explanation is that the median value is too sensitive compared to the dispersion value. This tells us those features derived from a relative measure can be more robust than those from an absolute measure. Third, the feature of poster lifespan deviation performs the best. It favors our intuition in Section 3.3.3: User lifespan is one of the strongest signals to capture the diversity of poster demographics.

Table 3.3: The rankings of the feature impact measured by Chi-Squared and Info Gain

Chi-Squared	Info Gain
Poster Lifespan Deviation	Poster Lifespan Deviation
Average Posting Count	Average Posting Count
Link Posting Entropy	Link Posting Entropy
Registration Interval Median	Registration Interval Median
Poster Lifespan Median	Tweets Counts Deviation
Tweets Counts Deviation	Poster Lifespan Median
Posting Interval Deviation	Posting Interval Deviation
Posting Interval Median	Posting Interval Median
Followers Counts Deviation	Followers Counts Deviation

We conduct two more informative comparisons of organic versus organized groups. One is about the number of distinct link domains that a group has posted. Our idea is that organized groups tend to have a much tighter selection of links to post than organic groups, mainly due to their specific common goal of posting. In Figure 3.12, we find the significant gap between two classes. More than 20% of organic groups have mentioned at least 100 different link domains in their tweets, and yet the median for organized groups is merely no more then 10. 20% of organized groups have posted only one kind of link domain all along.

The final exploration comes back to Twitter itself. We would like to see how the Twitter's official monitoring reacts for the two types of similar user behaviors in link sharing. So, we calculate the *group-level suspension percentage* in each of our groups, i.e., how many accounts in a group have been suspended. Recall that we have excluded those

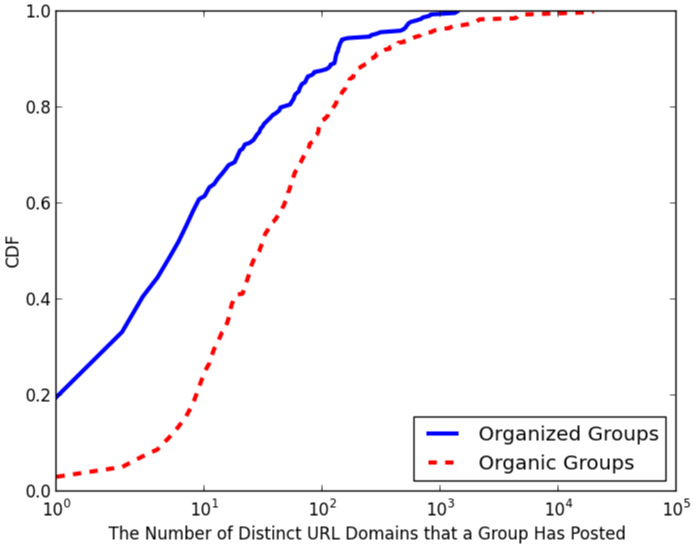


Figure 3.12: Organized vs. organic: number of link domain names

groups whose members were all suspended. We have two interesting observations from Figure 3.13. On the one hand, the distributions of our two types of groups are quite discerning. We find 80% of organic groups have less than 10% suspension percentage, where 80% of organized groups are more than it in contrast. Twitter's suspension is for individuals, yet it still reflects on our user groups formulated through link sharing. On the other hand, we find the official suspension still has limited impact on the organized posting behavior: the median is just around 30%. These findings suggest the complementary potential of our investigations on organized user behavior in link sharing.

3.5 Summary

In summary, we are interested in exploring users with manipulated behaviors when they share links on social media. While some users organically share common interests on certain websites, some are organized to aggressively promote the same links towards a common goal. This motivates us to tackle the problem of distinguishing organized and

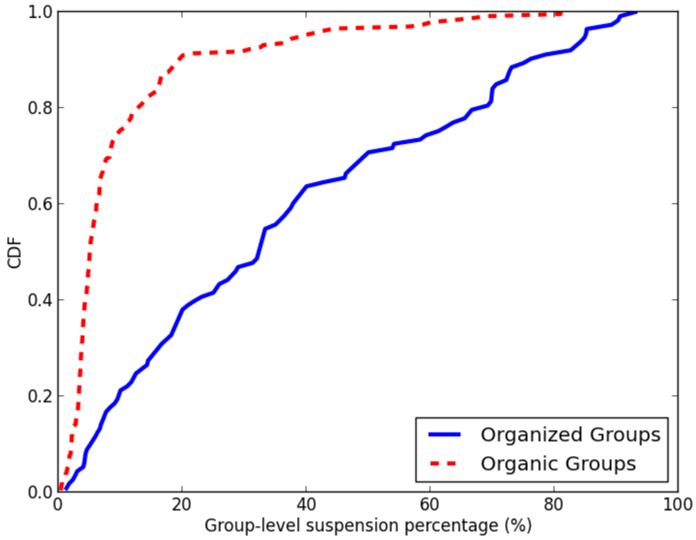


Figure 3.13: Organized vs. organic: group-level suspension ratio

organic users in the context of link sharing on social media. Focusing on the behavioral signals of posting links, we propose a four-step approach to model, identify, characterize, and classify those two types of user groups. We test our approach on four different classification algorithms and in most cases it performs good in terms of precision, recall, F-measure, and ROC area. Random Forest algorithm works best with 0.921 ROC. Our experimental analysis demonstrated the capability of our approach for (i) understanding users with similar link sharing behaviors; and (ii) distinguishing the level of manipulated user behaviors of link sharing.

4. COMBATING DISTORTION: IDENTIFYING BS ON SOCIAL MEDIA

In the previous two sections we have studied two specific problems toward tackling user misbehavior on social media — spamming and manipulation. In this section we investigate a newly emerging misbehavior on social media — distortion. We choose a unique perspective of user misbehavior that widely exists on social media: making bullshit (BS) statements. We formulate the problem, build a classification model detecting social media posts that are likely to be called BS, and present our experiment analysis, as follows.

4.1 Introduction

The use of the term "bullshit" (BS) has been on the rise since the mid 20th century.* But what, exactly, constitutes BS? Philosopher Harry Frankfurt addresses this question in his two seminal books: *On Bullshit* and *On Truth* [28, 29]. Frankfurt describes BS as a statement that does not address facts, but rather "misrepresents what the BS-er is up to." Frankfurt explains "one of the most salient features of our culture is that there is so much BS" [28].

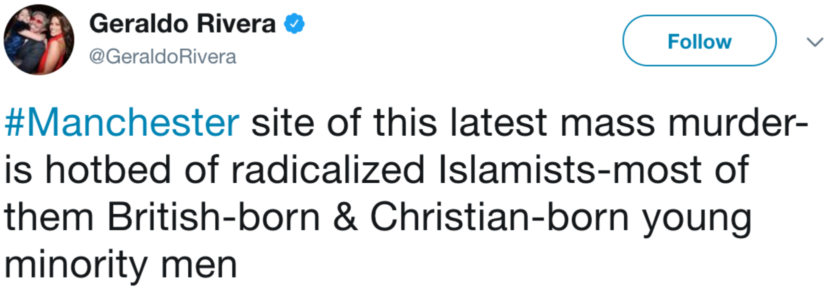
For example, advertisers often rely on BS, as in one case study that shows that many brands of coffee claim they are "99.9% caffeine free," however — due to the serving-size — these brands are as caffeinated as "strong coffee [which is] also 99.9% caffeine free" [31]. This kind of advertising is not only misleading, but can actually cause severe health problems for people who are caffeine intolerant [91]. In a similar vein, researchers have found that political candidates that BS more were seen more favorably during the recent 2016 US Presidential election [17].

Of course, BS reaches beyond politics and advertising. The development of online social media has provided humanity with the ability to share information like never before.

*https://books.google.com/ngrams/graph?content=BS

This has made it trivial to spread false information without accountability. While there are undoubtedly benefits to increased social media options, the current ecosystem has grown to facilitate the production of BS. This trend has not gone unnoticed by online communities and researchers alike. For example, there is a Reddit board dedicated to calling BS.[†] There is a BS-generator that produces BS based on user-submitted and crawled tweets.[‡] There is even a credit-bearing class taught at the University of Washington on “calling BS.” [§]

In this paper, we aim to extend Harry Frankfurt’s definition of BS onto social media, to uncover what is and is not perceived as BS. We define a *BS post on social media* as a claim where the poster does not care if his claim is true or false, but rather uses the post as a story to some other effect. BS can originate from anywhere (i.e. many users post BS; no user posts *only* BS), which makes it very difficult to identify. This is especially true when considering that BS is distinct from lying [28, 92] since most BS statements are not fact-checkable [93]. Although a series of studies suggest that humans have the ability to identify BS [94], the volume of potential BS on social media means traditional manual identification can fail because no service possesses the necessary human capital to label *all* posts. Yet, human judgment is still the most natural approach to identify BS.



#Manchester site of this latest mass murder-is hotbed of radicalized Islamists-most of them British-born & Christian-born young minority men

Figure 4.1: A published controversial tweet that is BS-like

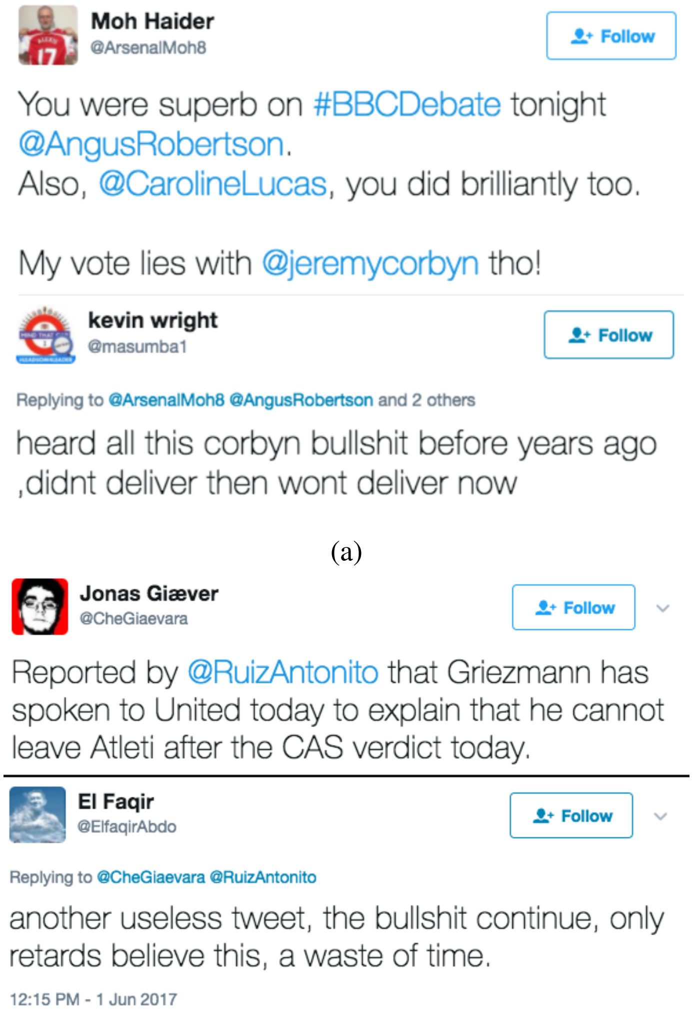
[†] https://www.reddit.com/r/quityourbullshit/

[‡] http://wisdomofchopra.com/

[§] http://callingbullshit.org/

Fortunately we can find footprints of human judgments left by other social media users. One of the most common forms is a reply either using the actual word "BS" or some statement to the same effect. Users can also leave an independent post of their own calling BS, or even report the original post. For example, Figure 4.1 shows an an example tweet that makes a controversial claim; in total, more than 20 unique replies "call" this tweet out as BS. This replying behavior is an imperfect signal, but it provides the tantalizing opportunity to uncover what is and is not perceived as BS by a large social media audience. Hence, in this paper, we view replying behavior as evidence of a "*BS call*". Replies can help us easily track *who* and *what* have been called BS. Also, replies are often topically motivated, meaning repliers care about the content of the post as opposed to other factors (e.g. author of the post) [32].

Of course, there is a gap between *BS-called* posts and *actual* BS posts. Figure 4.2a shows a posted tweet that has been called BS. The tweet praises several English politicians (not typically controversial), yet one user replies, calling the post BS. This post is called BS by another user, but we cannot hastily conclude whether it is "actually" BS tweets or not. In Figure 4.2b, one user shares a piece of (unverified) sports news and another user insists it is a useless BS. These two examples suggest the problem of BS detection is a challenging task, even for humans.



Moh Haider
@ArsenalMoh8

You were superb on #BBCDebate tonight @AngusRobertson.
Also, @CarolineLucas, you did brilliantly too.

My vote lies with @jeremycorbyn tho!

kevin wright
@masumba1

Replying to @ArsenalMoh8 @AngusRobertson and 2 others

heard all this corbyn bullshit before years ago ,didnt deliver then wont deliver now

(a)

Jonas Giæver
@CheGiaevara

Reported by @RuizAntonito that Griezmann has spoken to United today to explain that he cannot leave Atleti after the CAS verdict today.

El Faqir
@ElfaqirAbdo

Replying to @CheGiaevara @RuizAntonito

another useless tweet, the bullshit continue, only retards believe this, a waste of time.

12:15 PM - 1 Jun 2017

(b)

Figure 4.2: Two real posts; are either BS?

Therefore, instead of diving into the extremely nuanced task of detecting “actual” BS, our goal is to first build a model that can automatically determine what social media posts are likely to be *called* BS. We believe our work can serve as a stepping-stone to the ultimate goal of BS detection. For instance, our results can be used to filter posts that are unlikely to be BS from the vast social media stream. This could provide a more narrow search-space to develop a true BS identification method.

In order to determine which posts are likely to be called BS we focus on the audience's perception of BS. We do this by mining how the audience perceives poster's intent of BS-ing through signals from the post itself. We identify four factors within a post that influence a reader's perception: attitude, sentiment, sincerity, and content. Attitude influences intention as part of the *Theory of Planned Behavior* [95, 96]. The sentiment of a post can heavily influence the conflict surrounding it, which often results in calling BS [97, 98]. Sincerity is a measure of how much a poster cares about their claim. Since a BS-er does not care about the factual value of the claim [28, 29], how sincere a post is perceived to be can be a useful metric for determining how likely it is to be called BS. Finally, we examine the contents of each post because posts on certain topics are prone to being called BS.

To the best of our knowledge, this is the first comprehensive study on the topic of BS and BS call on social media. But, how to prepare an appropriate dataset of BS-called posts? Can we design an automatic model to identify posts that are likely to be called BS? To answer these questions, the remainder of this section makes the following contributions:

- A curated crowd-sourced collection of BS-called tweets gathered over a sequence of 100 consecutive days.

- A characterization of four factors impacting user's perception of BS, leading to a classification model to differentiate between posts that are more likely to be called BS and posts that are less likely to be called BS.

- A series of experiments showing our model is capable of leveraging linguistic cues for identifying posts are likely to be called BS, suggesting the great potential of a preliminary BS auto-filter on social media.

4.2 Related Work

There is a rich and growing philosophical foundation for BS. BS is described in classical philosophy. Plato's *Phaedrus* explores and critiques Hellenic Sophists calling them "non-lovers" (i.e. people who do not love the truth) [99]. Plato characterizes non-lovers as people who believe "language is not for telling the truth", but "for finding and strengthening positions, for gaining advantage, and for exerting influence over others" [98]. Most of the post-modern work on BS is from Harry Frankfurt's essay "On Bullshit" [28]. Frankfurt lays out a basic framework for BS, its causes, and the distinction between truth, lies, and BS. Frankfurt explains BS is unique and separate from both truth-telling and lying. Subsequent research and studies further establish that BS-ers do not care about the factual value of their claims, but rather care about some other effect on the audience [100, 98, 94].

Most of the existing research on BS identification is either theoretical or human-focused. George Orwell's essay *Politics and the English Language* (while not mentioning "bullshit" specifically) describes *Dying Metaphors*, *Operators*, *Verbal False Limbs*, *Pretentious Diction*, and *Meaningless Words* as identifiers of BS [101]. More recently, a series of human studies by Pennycook *et al.* focus on the human reception of BS, and the ability for humans to detect BS [102, 94]. The studies report that people tend to be receptive to BS, but are still able to distinguish between BS and meaningful statements. This particular study influenced our data collection method, which we discuss in Section 4.3.

Misbehavior on social media is a rapidly emerging field of study in the last decade [103, 104, 105, 5, 6, 106, 107, 10]. Castillo *et al.* build a credibility classification model [103]. Gupta *et al.* propose a semi-supervised credibility ranking model [105]. Popat *et al.* evaluate the credibility of textual claims on social media and other web services [106]. Lee *et al.* propose a classification model for spam detection on social media [6]. Hu *et al.* make use of sentiment signals in an optimization framework for social spammer detection [5].

Zhao *et al.* present a method for real-time rumor detection on Twitter [10]. Cheng *et al.* look into trolling behavior in the context of online discussion [104]. Rajadesingan *et al.* model the "sarcasm scenario" on Twitter [107]. Ratkiewicz *et al.* and Ferrara *et al.* study manipulated propaganda on social media [15, 108]. Hosseinmardi *et al.* explore cyberbullying behavior on social media [109].

BS presents a unique scenario of misinformation on social media. By definition, BS is a statement that does not address the facts but rather creates a narrative for some effect on the audience. Practically, many BS statements are not fact-checkable [93]. It is important to note that — according to Frankfurt [28, 29] — the BS-er does not simply obscure the facts, but invents new stories to suit themselves. A BS story can be 'true' but still be BS because of the author's intent. If the author intends a story to be for some other effect (other than relaying facts), the story can still be a BS (we will provide more details on the characterization later on). For example, when considering a BS-like example in Figure 4.1, we may ask, is it: Spam? Fraudulent? A rumor? Sarcasm? Propaganda? Cyberbullying? — It seems that it does not belong to any other category other than BS. Although BS can be roughly categorized as a type of misbehavior, we have not seen any existing work focusing specifically on BS as it relates to social media.

4.3 Data Collection

Using Twitter's public Stream API, we collect English tweets with the keyword "bullshit" (and its synonyms such as "bs", "bull$#it", "malarkey", etc.) which reply to another tweet. These are the *BS-calling tweets*. We then back-track and collect the tweet being replied to (the *BS-called tweet*). In this study we focus on the replying behavior as the evidence of BS calling (i.e., any BS-calling tweet must be a reply to another tweet), with two main reasons. First, replies can help to track *who* and *what* are called BS. Second, replies are often topically motivated, meaning the repliers care about the posts they reply [32]. A reply can appear in a "reply chain" (e.g. a reply to a reply). We only consider those BS-called tweets which are not replying to anything other than themselves in order to maintain the context of the BS call. In this way we can automatically gather a collection of pairs of BS-calling tweets and BS-called tweets. These BS-called tweets are potential BS tweets.

We use the "potential" qualifier because we cannot conclude that every BS call is legitimate. We have seen how challenging BS identification can be, and a BS-calling tweet is almost always somewhat subjective (i.e. dependent on the opinions and knowledge of the BS-caller). We are aware that we may include many false positives if we rely on every BS-calling tweet. Thus, to curate our dataset we take into account the number of BS calls on a certain tweet. The more times a tweet gets called BS, the likelihood of the tweet being *actual* BS increases. Does this assumption hold? Can a threshold of a certain *called count* help identify a more "pure" set of BS tweets? Toward answering these question we explore our data as follows.

4.3.1 Filtering the Data: BS vs. Called BS

We conduct a series of data labeling experiments to study what role called count plays in filtering our dataset. Our collection spans a sequence of 100 consecutive days, from April 4, 2017 to July 15, 2017. Table 4.1 provides an overview of BS-called tweets. We can see the majority of BS-called tweets (88%) are called only once, while only a few (1.3%) receive more than 5 BS calls.

Table 4.1: Overview of BS-called tweets

	BS-called	**Called count = 1**	**Called count > 5**
# Tweets	697,865	617,613	9,256

Next, we design a process of manual judgment to see whether we can use called count to obtain a purer set of BS tweets. We prepare seven groups of BS-called tweets, each with a different threshold of called count. In concrete, we create the following seven groups of tweets:

- Group A: We randomly sample 1000 tweets without any query from Twitter's streaming pool. This dataset serves as a general sampling of the whole population of tweets.

- Group B: We first collect all those users whose tweets have been called at least once. Then we find their posted tweets which did not get called BS within our time window.

- Group C: Contains all BS-called tweets whose called count is no more than 4, i.e., $1 \leq$ called count < 5.

- Group D: Includes all tweets that have been BS-called at least 5 times, i.e., called count ≥ 5.

- Group E: Includes all tweets that have been BS-called at least 10 times, i.e., called count ≥ 10.

- Group F: Includes all tweets that have been BS-called at least 15 times, i.e., called count ≥ 15.

- Group G: Includes all tweets that have been BS-called at least 20 times, i.e., called count ≥ 20.

Tweets are put into seven anonymous groups before judging them. We then randomly sample 100 tweets from each group and ask three judges to determine if the tweets are

BS based on their *rational perception*. In order to facilitate judgment the individuals were given the same set of advice:

- A BS post is usually making an assertion about a topic. If there is no assertion (no "story") then it is unlikely that the tweet is BS.
- Is the topic of the assertion clear? If not then the tweet is also unlikely to be BS.
- Does the tweet provide any reasoning or support for its assertion? If not then the tweet is likely to be BS.
- If there is evidence, is the reasoning or evidence itself BS? If so then the Tweet is likely to be BS.

The judges are all fluent in English, without any leaning of politics or religion. Some tweets contain links. Judges are allowed to view the external links, but are instructed to base their judgment primarily on the content of the tweet. The judges are also allowed to examine the users' profiles in order to determine whether or not a user would have a bias or reason to post BS. The judges all evaluated the tweets *independently*, with no knowledge of the other's judgments.

The judges are then asked to judge whether a tweet is BS or not, based on their perception. If the judge is uncertain then they can respond as such. In order to determine whether a tweet is BS, we take the majority of the three responses. If there is no majority (e.g. one 'yes', one 'no, and one 'uncertain') a new tiebreaker judges the tweet to have a majority. Although human-judged BS is not identical to actual BS, it can serve as a high-quality standard of *the perception of BS* for our follow-up studies.

Figure 4.3 presents the results of human judgment over the seven groups of tweets above. As expected, the identified BS tweets monotonically increase and the non-BS tweets monotonically decrease from Group A to G. This shows that having a higher called

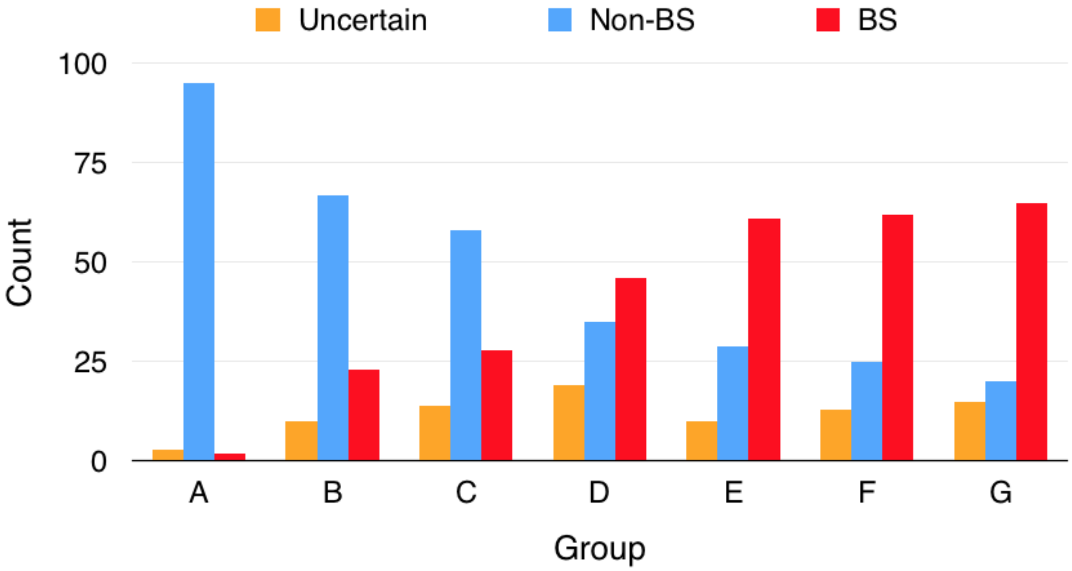
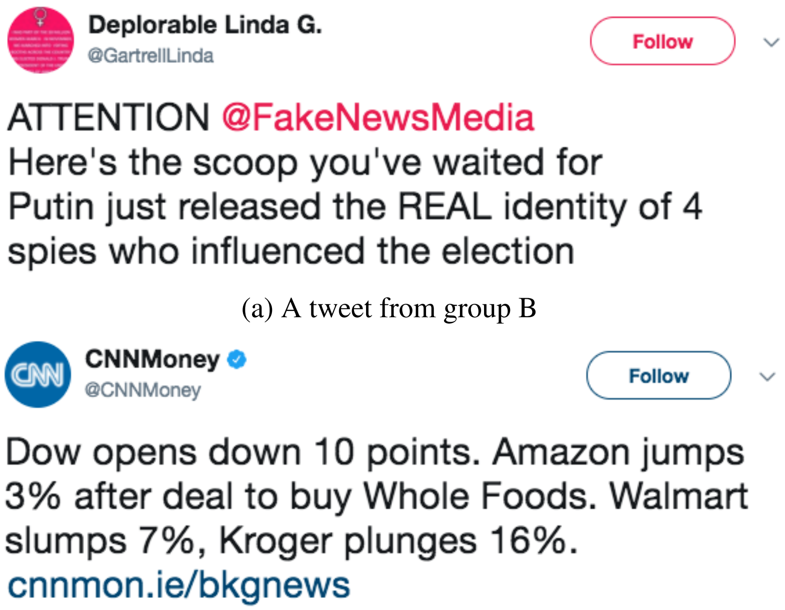


Figure 4.3: Human judgment over six groups of tweets

count threshold tends to filter out more non-BS tweets. On the other hand, we don't see a perfect called count threshold that can give us "100% pure" BS or non-BS tweets. For instance, less than half of the tweets in group D (i.e., being called BS at least five times) are recognized as BS. Also, 25% of the tweets that have not been called BS (Group B) are still identified as BS. Figure 4.4 shows two real tweets that are included in our data. The one in Figure 4.4a is from group B and the one of Figure 4.4b is from group C. After reading them, however, it appears that the top one is a BS tweet and the bottom one is not, even though the former received zero BS calls and the latter got more than one BS call.

4.3.2 Problem Formulation

The observations from Figures 4.3 and 4.4 indicate that while crowd-sourced BS calls can be indicative of BS, there is always a gap of uncertainty between posts that are called BS and actual BS posts. Hence, our goal in this paper is to build a model that can automatically identify what posts are likely to be called BS. Formally, given a post p that has



Deplorable Linda G.
@GartrellLinda

ATTENTION @FakeNewsMedia
Here's the scoop you've waited for
Putin just released the REAL identity of 4
spies who influenced the election

CNNMoney
@CNNMoney

Dow opens down 10 points. Amazon jumps 3% after deal to buy Whole Foods. Walmart slumps 7%, Kroger plunges 16%. cnnmon.ie/bkgnews

(a) A tweet from group B

(b) A tweet from group C

Figure 4.4: Two tweets from our dataset

been published on social media, our goal is to identify whether p is likely to be called BS through a binary classifier $c : p \rightarrow \{\textit{likely BS-called}, \textit{unlikely BS-called}\}$. We require that our model can be generalized for other social media domains even though our data in this study is collected on Twitter. Because BS is defined as a claim [28], we also emphasize our classification model needs to focus on the post rather than user level. The outputs of our proposed model could be used to assess the relative BS-level of users through an analysis of their posts.

We need to prepare two sets of tweets to train our classification model — one set that contains posts most likely to be BS-called (the positive class) and another includes those less likely to be BS-called (the negative class). Of course, a BS-called post is not necessarily an actual BS post, so we want to reduce the gap in our training data as much as possible. According to Figure 4.3, the tweets in group A have zero BS calls and more than

95% are labeled as non-BS. In group E, F, and G are all tweets with at least 10 BS calls, and all have more than 60% tweets that are labeled as BS. We are left with fewer tweets as we increase the threshold of the called count, and yet the proportion of labeled BS does not increase much. Thus, we choose the tweets of group A as the unlikely BS-called (i.e. negative class) and the ones in group E as the likely BS-called (i.e. positive instances). We end up with 3,370 tweets of group E and randomly sample the same amount of tweets from group A to obtain a balanced training set.

4.4 What Tweets are Likely to be Called BS?

4.4.1 Overview

As mentioned earlier, we define a BS post on social media as a claim where the poster does not care if his claim is true or false, but rather uses the post as a story to some other effect. A BS post is not a human-applicable concept (i.e. no human *is* BS), but any BS post is related to the person who creates it in that the person *intends* to post BS. According to Frankfurt [28, 29], the BS-er does not simply obscure the facts (lie), but invents new "facts" to suit themselves. A BS story can be modeled after the truth, but still is BS because the author intends the story for some other effect. Hence, a key step of characterizing BS is to find out whether the speaker (poster) cares if his story (post) is true or not.

Due to the difficulty of assessing a poster's intention we focus on the audience's perception of a story (post). It is unrealistic to fully mine a user's *intent* of posting BS — we will never truly know what a user thinks when he posts. However, we can mine how the audience *perceives* the poster's intent (discussed in Section 4.3). But what makes the audience "perceive" intent? The answer: signals from the post itself [102, 94].

Therefore, we focus on post-level linguistic footprints that can shed light on answering the question: what posts are more likely to be perceived as BS? We operationalize Frankfurt's definition of BS through an initial model of how the audience perceives what is and is not BS. Concretely, we extract multiple linguistic signals from a given post that characterize the four poster-level perspectives shown in Figure 4.5:

- Attitude — Our idea is that a BS claim usually presents a strong attitude of the speaker in order to affect the audience. This inspires us to examine the indicators of poster's attitude based on the writing style of a post.
- Sentiment — Since calling BS is the ultimate outcome of conflicts of stance, we seek for verbal evidence of sentiment and opinions in a post. For example, a post with a strong sentiment can often lead to stronger "counterforce" from the audience.
- Sincerity — According to Frankfurt, BS closely relates to understanding whether the poster cares about the factual value of his post, which, in other words, is poster's sincerity. Thus, from the post we look for textual footprints of "recognizable" sincerity.
- Content — BS-ers usually have a clear motivation on how and to what end they affect the audience who also has motives for why they call a post BS. Thus, our idea is that posts on certain subjects are prone to being called BS.

We explain the motivation and design choices for each feature throughout the remainder of this section.

We do not explicitly include the poster's profile as a fifth perspective, but our linguistic features take into account *audience's perceptions on both post and poster*. A user with a large audience (e.g., millions of followers on Twitter) tends to attract more attention, as do active posters. Yet, normalizing the count of BS call by the total number of followers or received replies will bias the results towards unpopular users. For example, in our data we find President Trump received more than 1,000 BS calls. Thus, considering user profiles as features can lead to an *overfitted* but inaccurate classifier which would lead to a trivial conclusion that users with many followers or posts are more likely to attract BS calls. However, we understand that completely discounting poster-related information can be shortsighted — some users call a social media post BS not by what it says but who says it. Hence, we find those post-level linguistic cues that can also grasp audience's perception on the poster. For instance, the perspective of attitude indicates the poster's writing style. Sentiment captures the poster's strength of stance. Sincerity directly connects to the poster's mental process while writing. Content can reveal poster's motivation of posting BS. These linguistic signals are based on the post but they all influence on audience's perception of the poster's intention.

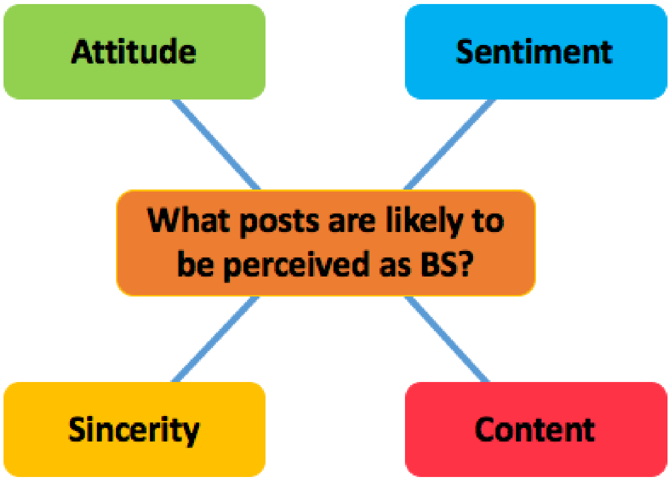


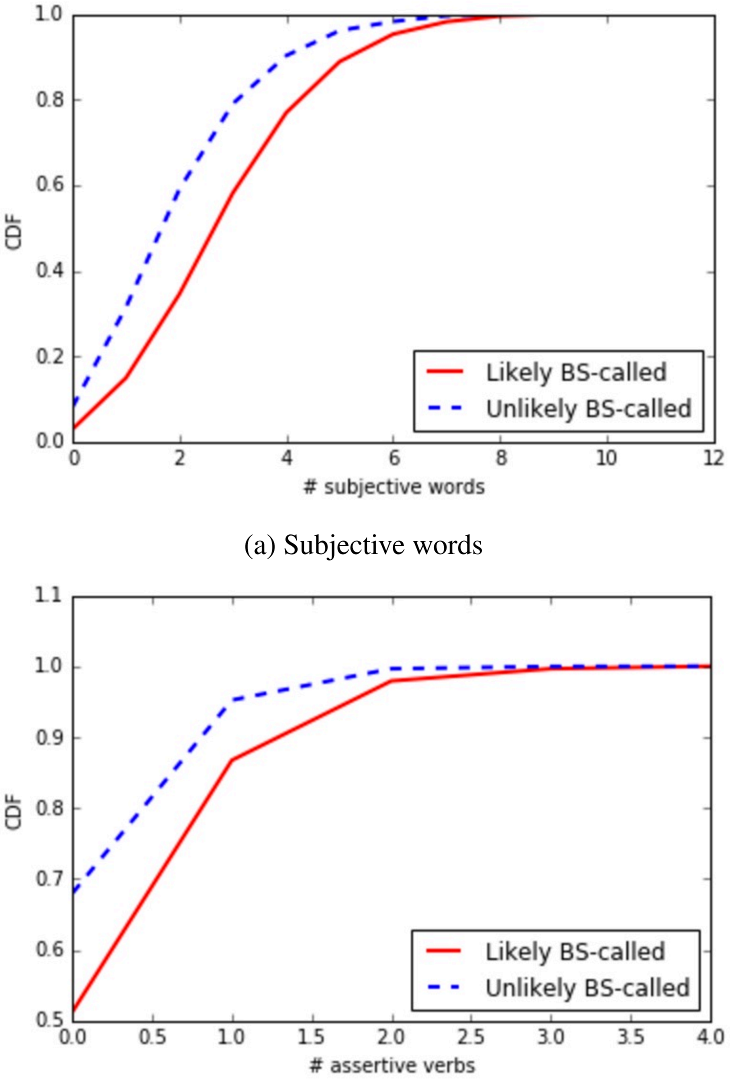
Figure 4.5: Four perspectives characterizing what posts are likely to be perceived as BS

4.4.2 Attitude

In order to impact the audience's perception, a BS claim usually reveals a strong attitude of the BS-er. Attitude, intention, and perception are three cornerstones of the *Theory of Planned Behavior* in psychology [95, 110]. It states that attitude toward behavior, subjective norms, and perceived behavioral control, together shape an individual's behavioral intentions and behaviors [95, 96]. This inspires us to examine the writing style of a post that indicate the poster's attitude.

BS tends to be highly subjective and arbitrarily assertive [97, 29, 98]. In order to affect the audience, BS-ers usually express their subjective feelings or viewpoints with an assertive attitude. In contrast, a typical non-BS (such as scientific articles) usually cautiously report observations. A claim based on a baseless presupposition or preconception is generally seen as less reliable by the audience [98]. Thus, we focus on specific words in a post that reflect subjectivity and assertiveness. For example, assertive verbs (e.g., "assure", "insist") show the level of assertiveness; report verbs (e.g., "deny", "show") and implicative words (e.g., "dare", "preclude") represent the attitude toward preconception. Mitigating words (e.g., "maybe", "possible") soften the commitment of an assertion.

For feature extraction we build a collection of lexicons that have been widely used in other literature [111, 112, 113, 114, 115]. Given a social media post we extract the number of terms occurring in each lexicon (after parsing and stemming). To validate our intuition, for each feature we compare the cumulative distribution functions (CDFs) of the positive tweets and the negative tweets prepared in Section 4.3. Figure 4.6 shows the CDF comparisons for the number of subjective words and assertive verbs. We find those more likely called tweets tend to contain more subjective words and assertive verbs. The median number of subjective words in the positive class is 2.3 while for the negative class it is 1.6. Half of all positive tweets have at least one assertive verb but almost 70% of the negative tweets do not include any assertive verb. We have similar observations for report verbs and implicative words. In the case of mitigating words, the trend is reversed: 14% negative instances have at least one mitigating word while only 8% positive tweets include one.



(a) Subjective words

(b) Assertive verbs

Figure 4.6: Likely called vs. unlikely called: subjectivity and assertiveness

4.4.3 Sentiment

Sentiment plays an import role in the perception of BS. This is especially true in situations where calling BS is the ultimate outcome of conflicts of stance. Posts that are opinionated or have a distinctive bias are more likely to evoke disagreements with others. A stronger sentiment in the post can often result in the stronger “counterforce” from the audience, which is prevalent in BS calling [97, 98]. In contrast, a post with neutral sentiment is usually not divisive, so the likelihood of a neutral post being called BS is lower. Hence, we rate the “strength of sentiment” for a post via seeking verbal evidence of sentiment and opinions.

There are many lexicons of words that indicate bias, opinion, and sentiment. We combine six different lexicons. First, we applied the *SentiStrength* sentiment lexicon which focuses on “short, informal texts,” rating each word from -5 (very negative) to 5 (very positive) [116]. Then we applied AFINN which “is a list of English words rated for valence” with a range of $[-5,\ 5]$ [117]. Next we applied the MPQA Subjectivity (MPQA-S) and the MPQA-Effect (MPQA-E) Lexicons. The MPQA-S lexicon contains a word and the subjective feeling which it implies, which can be either positive or negative [118]. The MPQA-E lexicon provides each word with the effect on the overall sentence: ‘+Effect’ or ‘-Effect’ [119]. We also applied the UIC lexicon which has positive and negative sentiment annotations for 6,800 words [120]. Finally, we added the *SenticNet 4* sentiment lexicon which also provides ‘positive’ or ‘negative’ annotations for each word [121].

All of these lexicons provide a score for each *word*. However, all scores do not necessarily agree. In order to account for the disagreement of our lexicons, we combine the scores of all of the lexicons and take the proportion of negative scores for each word and the proportion of positive scores for each word and treat them as separate entities. This maintains the “strength” of sentiment within a word. If a word has more disagreements, it is more ambiguous (score closer to 0). Whereas, if a word has fewer disagreements it conveys a strong unipolar sentiment (score closer to 1).

In order to score an entire tweet we combine the scores for each word by taking the total negative and positive scores over the sum of the negative and positive scores. Let S be the set of positive and negative scores for each word in a tweet. Let s_p be the positive score of a word and s_n be the negative score of a word. We calculate two measures T_p and T_n as the following:

$$r_p = \sum_{s_p \in S} s_p, \;\; r_n = \sum_{s_n \in S} s_n$$
$$T_p = \frac{r_p}{r_p + r_n}, \;\; T_n = \frac{r_n}{r_p + r_n}.$$

T_p and T_n maintain the strength of sentiment for each tweet because we take the proportion of positivity and negativity to the total sentiment. In this way, a tweet with strong sentiment has positive and negative scores that are vastly different, while a neutral tweet has a positive and negative scores that are closer together.

4.4.4 Sincerity

As discussed, BS characterization closely relates to understanding whether the poster cares about the factual value of what he posts; in other words, poster's sincerity. Here the definition of sincerity is limited to the context of posting. In order to extract the user's sincerity from the post itself, we look for the textual footprints of "recognizable" sincerity. We consider two popular concepts in psychological content analysis — cognition and readability.

The cognitive process behind a text post is important. Analyzing the psychological content in a text can help understand its author's mental process while writing. Frankfurt explains that a BS-er often intends to BS [122], which leads us to examine the cognition behind each post. We use the Regressive Imagery Dictionary (RID) which identifies primordial conceptual content. The RID returns the percentage of the *primary*, *secondary*, and *emotional* thought present in text [123]. Primary thought is generally free-form associative thinking involved in dreams and fantasy. Secondary thought is focused on logical, reality-based problem solving. Emotional thoughts are expressions of of fear, sadness, hate, affection, etc. We collect each as an independent feature through the implementation of NodeBox English Linguistics [124].

The readability of a post is also important to the author's sincerity. Readability is the measure of the approximate level of education necessary to understand a post. The amount of education necessary to understand a post is an important consideration because the less readable a post is (the higher the level of education needed to understand the post), the more it is associated with misbehavior (e.g. deception) [125]. Although credibility and BS are not the same, the research literature has observed a strong association between them [97, 29, 98]. Measuring readability is challenging when it comes to social media due to the brevity of posts, especially since most readability formulas require a minimum sentence count. In order to compensate for this we perform a series of readability tests (*Flesch-Kincaid Grade*, *Coleman-Liau*, *Automated Readability*, *Linsear Write*, and *Gunning-Fog*) [126, 127, 128, 129, 130]. We use their scores (measured in *U.S. grade-levels* [131]) as independent features.

4.4.5 Content

Posts on certain subjects are prone to being called BS. BS-ers usually have a clear motivation on how and to what end they want to affect their audience. The audience also has motives for why they call a post BS. For example, a Pinterest post introducing a recipe for baked salmon is less likely to be called BS (or to even be BS in the first place). On the contrary, a Facebook post by a partisan news outlet reporting a revised health-care bill is more likely to ignite many disputes.

We search for hints of content shift in our prepared likely BS-called and unlikely BS-called tweets. Twitter does not provide explicit topical information for each tweet and it is unrealistic to manually label all tweets. With the idea that a user on social media usually posts with a single focus (usually what she is interested or good at), we investigate the clues of poster's *topical profile*. We leverage *Twitter Lists*, a large publicly-accessible collection of crowd-contributed user tagging. Twitter Lists allows users to label each other with a tag, e.g. politics, sports, art, etc. We query list memberships for each user in our data via the Twitter API. Finding many nonsense tags and near-synonyms, we manually compile a leaner list of the 200 most popular tags by filtering out nonsense and merging variants (e.g. excluding "cool-people"; merging "breaking-news" and "noticias" into "news"). Note that the derivation of these tags is completely independent from the formation of BS and calling BS.

Next, for each tweet we count how many times its poster is labeled by a tag related to politics. We choose politics in the first place because politics-related content is usually controversial [132] which — as we have discussed — are favorites of BS callers and BS-ers. Figure 4.7 shows the CDFs for posters of likely BS-called posts and unlikely BS-called posts. We see a big difference: more than 60% of the posters of likely BS-called tweets are tagged at least once with politics-related tags as their topic focus and more than 20% are labeled at least 50 times; but almost 80% posters of unlikely BS-called tweets have not been tagged with "politics." Similarly we observed posts whose authors are tagged by "entertainment news" are more likely to be called BS compared to users who focus on "technology" or "art".

All of these observations imply some content is more likely to be perceived as BS than other content. However, our classification model would lose the generality if we only preselect certain tags as the "vulnerable topics" and exclude the others. Hence, we should avoid directly using the measure in Figure 4.7 as a feature of our model. But how can we capture the topical distribution only based on the text of a social media post? Recent advances in distributed text representation learning have offered effective text embedding methods that capture syntactic and semantic features [133]. In this way, for each post we have a "vector of content" without specifying the topics so that all nuances of content difference can be

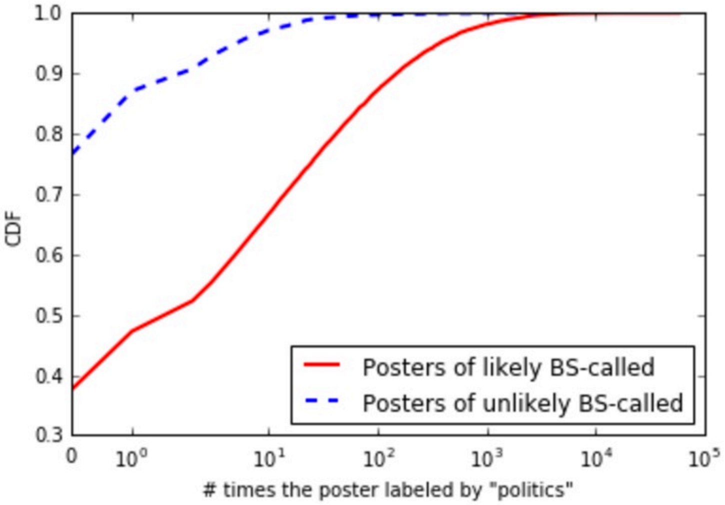


Figure 4.7: Likely called vs. unlikely called: politics-related poster

captured in the same high-dimension space. We train a word embedding model with the whole collection of text posts using the popular implementation *Word2Vec*,[¶] and average the learned latent vectors of term as features for each post. Parameters such as window size and dimensionality are empirically determined.

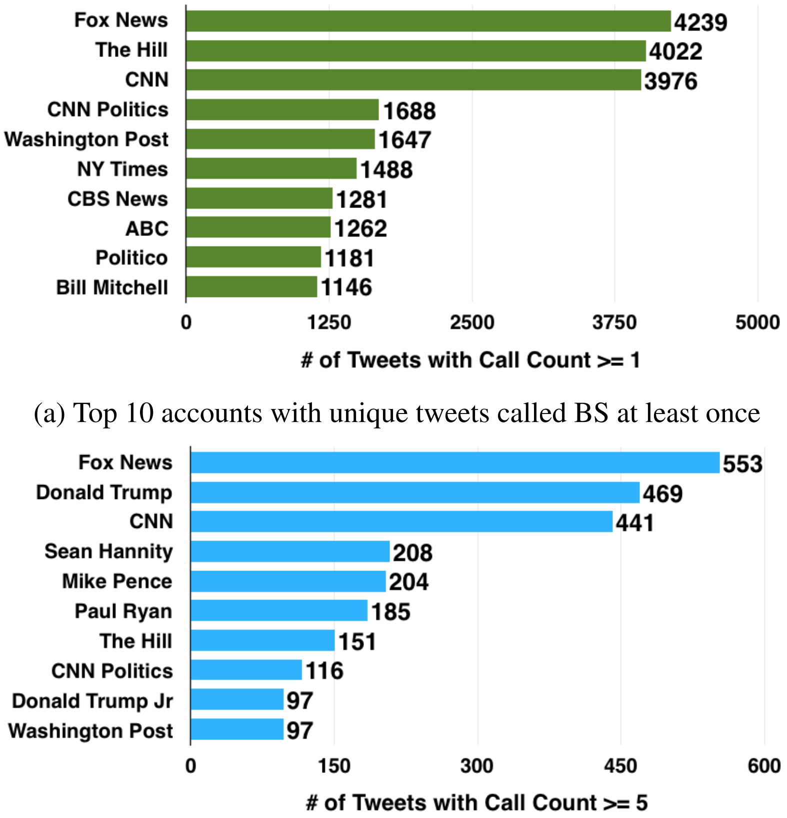
4.5 Experiments

4.5.1 Data Exploratory Analysis

We perform a series of data exploratory analyses in order to know more about our data. First, we attempt to have a better idea of who has been called BS the most. Next, we explore what topics users who are called BS mostly post about. Finally, we explore where BS called tweets and BS calls originate.

Who are called BS the most? This is one of the most natural questions to ask given the whole dataset. To answer this, we rank all users by two metrics. We first rank them by how many of their unique tweets have been called BS, regardless of how many times they

[¶] https://www.tensorflow.org/tutorials/word2vec have been called. The top-10 accounts are shown in Figure 4.8a. Next we rank them by the number of unique tweets that have been called at least 5 times, shown in Figure 4.8b. This metric allows us to examine more closely which users may be actually BS-ing more than others due to the increased purity of the dataset (as discussed in Section 4.3).



(a) Top 10 accounts with unique tweets called BS at least once

(b) Top 10 accounts with unique tweets called BS ≥ 5 times

Figure 4.8: Accounts that have been called BS the most

We see nearly all of these users have a close connection to politics. “Fox News” has 4,239 tweets that have been called BS, and 553 of them have been called BS more than 5 times. “CNN” has 3,976 tweets being called at least once. “The Hill” is a newspaper covering the U.S. Congress and Presidency which has been called 4,022 times. Sean

Hannity is a "conservative political commentator" who is employed by Fox News. Paul Ryan is the current Speaker of the United States House of Representatives. Bill Mitchell is a YouTube and talk show personality. Another observation is the ranking in Figure 4.8a mostly contains news outlets while the ranking in Figure 4.8b includes more individuals related to politics. This could mean that in spite of the fact individuals tend to post fewer tweets than news organizations, important political figures attract more BS calls.

What is called BS the most? A challenging task is to create a representation of the topic focuses of all the posts that are called BS in our dataset. In other words, *what* topics have been called BS? This gives us a better understanding of what topics are likely to evoke BS calls. Since it is challenging (and potentially noise inducing) to manually label thousands of users, we instead leverage Twitter Lists as discussed in Section 4.4.5. There is no set limit on how many tags a user can be labeled with or how many times they can be labeled with that tag. We collect the tags and choose the mode as the representative tag for each user, leaving each account with one topic focus.

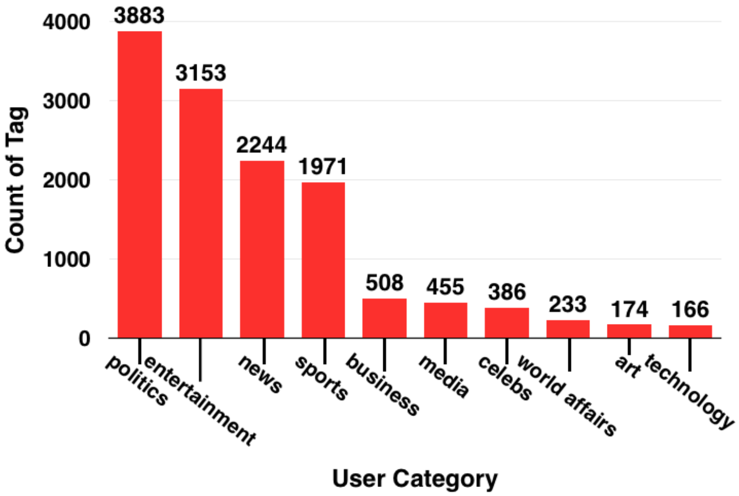
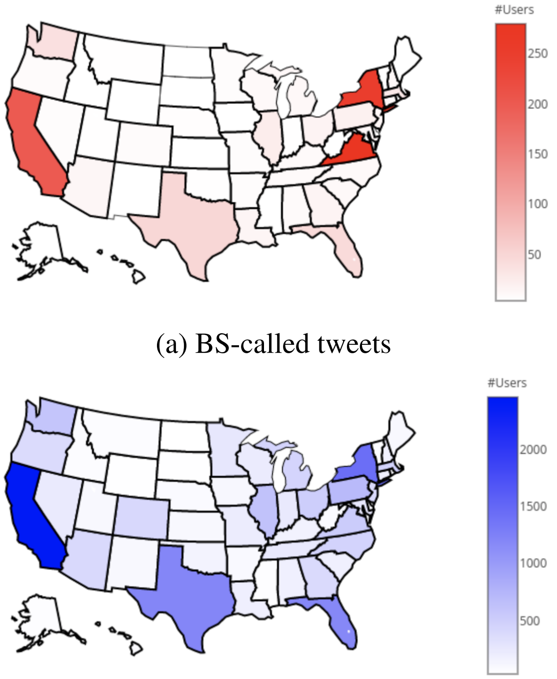


Figure 4.9: Top focuses of users with BS-called tweets

Figure 4.9 presents the distribution of tags over all of the users who had at least one unique tweet called BS at least once. The largest category is “politics,” followed closely by “entertainment.” Politics has been a scrutinized subject recently especially since the controversial 2016 United States Presidential Election. Politics has also been previously identified as a hub for BS-related activity [122, 17]. The entertainment industry is notorious for spreading many dramatized and invented stories (i.e. BS) [134], so it is not surprising to see that many entertainment-centric users have been called BS. The whole distribution shows the dominance of politics, entertainment, and news outlets as the most common users that are called BS. This makes sense considering all of these users are primarily spreading information which — as we discussed previously — they are not held accountable for.

Where do callers and their targets come from? Location is a fundamental demographic for any population. We want to investigate the geographic distribution of callers as to whom they are calling. To this end, we ask: Where do callers and their targets come from? In order to answer this question we leverage the information found in each tweet as well as the account which posted the tweet. We gather the geographic (latitude and longitude) information from each tweet if available, otherwise we extract the location mentioned in the description provided by users themselves (after manually filtering out nonsensical locations). We focus on the United States in this study.

Figure 4.10a is a heat map showing the geographic distribution of tweets called BS. A large portion of tweets that are called BS are from three states — Washington D.C. (plotted as a part of Virginia), New York (NY), and California (CA). Many politicians, news outlets, and celebrities — all of which are targets of BS-calls (as discussed) — reside in these three states. This geographic distribution seems to confirm the analysis in Figure 4.9. Many of the BS-ers are labeled as “politics” and most of them also are



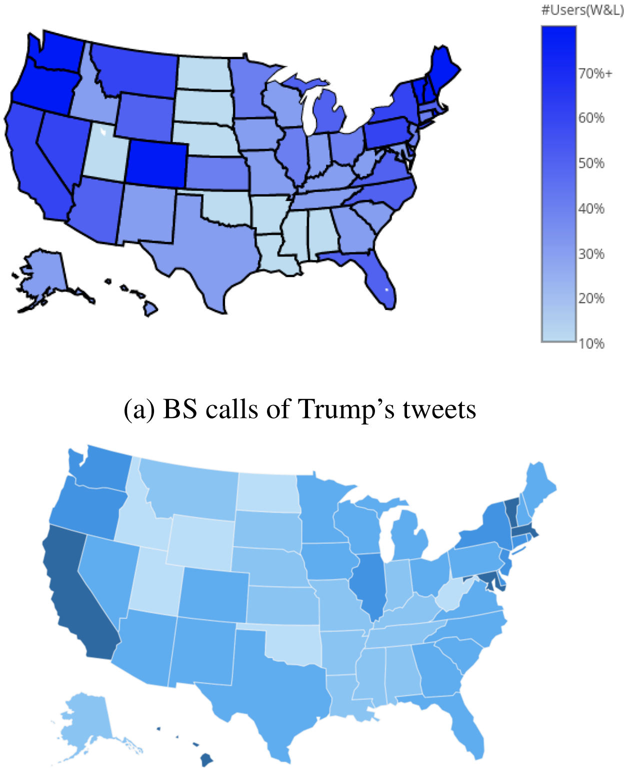
(a) BS-called tweets

(b) BS-calling tweets

Figure 4.10: Geographic distributions of BS-called tweets and BS calling tweets

based in Washington D.C. Figure 4.10b shows the distribution of tweets calling BS. We find two distinct differences when comparing this distribution to Figure 4.10a. First, not many callers come from DC. Second, the distribution of BS-callers is more reminiscent of the Unites States' general population distribution [135]. These observations seem to point to the fact that BS-calling tweets (and BS-callers in general) are unclustered and representative of the population distribution of the United States.

We have observed that many BS-called tweets are based in Washington D.C. and many BS-callers are based in CA and NY. We aim to determine whether or not the general political leaning of a location has an impact on the BS callers in that area. To answer this question, we carry out a case study on the 2016 Presidential Election. To begin we collect all of the tweets which are posted by President Trump and called BS and plot the distribution of all tweets calling Trump's posts BS. To avoid bias toward states with bigger populations, we normalize the number of callers in a state by that state's actual population density [135].



(a) BS calls of Trump's tweets

(b) Clinton's supporters in 2016 Election

Figure 4.11: Geographic distributions of Trump's BS calls and Clinton's supporters

Figure 4.11b is the normalized geographic distribution of Trump's BS-callers. We find most of the regions saturated with Donald Trump's BS-called tweets are traditionally Democratic-voting states. Figure 4.11a is the density of votes cast for Hillary Clinton in the 2016 presidential election.|| We clearly see a relationship between the geographic

|| `http://www.bbc.com/news/election-us-2016-37889032`

distributions of Trump's callers and Clinton's supporters.

4.5.2 Classification

We are ready to build a classification model considering all the four perspectives introduced in Section 4.4. To choose appropriate classification algorithms, we prefer that: (i) the algorithm has been widely used and maturely implemented; and (ii) we need to test on a variety of representatives. Concretely, we choose four well-known classification algorithms with different theoretical foundations: logistic regression (linear model), random forest (tree-based ensemble), support vector machine (kernel function), and multi-layer perception (neutral network). We abbreviate them *LR*, *RF*, *SVM*, and *MLP*, respectively.

We also alter various settings of each classification algorithm. For instance, we test two split criteria in RF (*Gini impurity* and *information gain*), four solvers in LR (*Newton-CG* [136], *L-BFGS* [137], *LIBLINEAR* [138], *stochastic average gradient* [139]), four kernels in SVM (*linear*, *polynomial*, *RBF*, *sigmoid*), four activation functions in MLP (*sigmoid*, *tanh*, *relu*, *identity*), and three solvers in MLP (*L-BFGS*, *stochastic gradient descent*, *ADAM* [140]). We also test standardization and implement feature selection based on *F-value* or *Chi-square* (if applicable). To evaluate, we do 10-fold cross validation over the 6,740 tweets (3,370 for each class) prepared in Section 4.3. We measure the classification performance by three popular metrics: micro-averaged F1, macro-averaged F1, and macro-averaged AUC of ROC curve.

Table 4.2 shows the results (95% confidence interval (CI)) of three metrics over four algorithms and four feature sets introduced in Section 4.4. For each algorithm we report the best performance among all tested settings. We draw several observations from Table 4.2. Horizontally, SVM and MLP tend to outperform the other two in general, and in some cases the differences are statistically significant. For instance, SVM and MLP work equally well across all three metrics, and they significantly outperform RF in most cases.

Table 4.2: Classification performances (95% CIs) over four algorithms and five feature sets

Feature Set	**F1 Macro (mean ± margin of error)**			
	LR	**RF**	**SVM**	**MLP**
Attitude	0.641 (0.007)	0.633 (0.011)	0.639 (0.008)	0.643 (0.008)
Sentiment	0.665 (0.004)	0.620 (0.010)	0.665 (0.007)	0.670 (0.008)
Sincerity	0.713 (0.010)	0.708 (0.006)	0.724 (0.009)	0.728 (0.009)
Content	0.707 (0.012)	0.710 (0.010)	0.715 (0.010)	0.733 (0.009)
Together	0.790 (0.007)	0.770 (0.008)	**0.798** (0.011)	0.797 (0.008)

Feature Set	**F1 Micro (mean ± margin of error)**			
	LR	**RF**	**SVM**	**MLP**
Attitude	0.641 (0.007)	0.634 (0.010)	0.640 (0.008)	0.643 (0.007)
Sentiment	0.665 (0.004)	0.621 (0.006)	0.665 (0.007)	0.668 (0.007)
Sincerity	0.713 (0.009)	0.705 (0.008)	0.725 (0.009)	0.726 (0.009)
Content	0.707 (0.012)	0.711 (0.010)	0.716 (0.010)	0.733 (0.008)
Together	0.790 (0.007)	0.770 (0.008)	**0.798** (0.011)	0.796 (0.010)

Feature Set	**AUC (mean ± margin of error)**			
	LR	**RF**	**SVM**	**MLP**
Attitude	0.690 (0.010)	0.679 (0.013)	0.690 (0.011)	0.691 (0.011)
Sentiment	0.721 (0.006)	0.660 (0.009)	0.721 (0.006)	0.724 (0.007)
Sincerity	0.785 (0.010)	0.772 (0.010)	0.785 (0.010)	0.796 (0.009)
Content	0.787 (0.012)	0.784 (0.010)	0.804 (0.012)	0.817 (0.010)
Together	0.867 (0.008)	0.848 (0.009)	0.876 (0.008)	**0.878** (0.007)

The neural network model MLP, having been a rising power this decade, works consistently well for our task. Vertically, we find features of sincerity and content statistically work better than those of attitude and sentiment. Sincerity is the measure of how much a person "cares" or how genuine they are about what they are talking about. This plays into their intention, so it makes sense that this feature set works great. Similarly — as we have discussed in Section 4.4.5 — the content of a post connects to the motivations of both BS-ing and calling BS, so it is consistent with our observations to see this feature set working well. Besides, we find the RID to be a very telling feature, especially because it allows us to peer into the cognition of the post. Readability also increases the results of our classifier, due to its direct effect on perception. Sentiment encapsulates only one portion of perception, so it works better in conjunction with other features.

We always see noticeable improvement as we consider all types of feature sets together. For example, SVM outputs a 95% CI of F1 Micro 0.798 ± 0.011 when considering all features together. The highest AUC scores we have achieved is 0.878 ± 0.007, when we apply the neural network model and consider all four aspects at the same time. These promising results show the effectiveness of our classifier that leverages linguistic cues to identify likely BS-called posts and the great potential of serving as a preliminary BS auto-filter on social media.

4.6 Summary

BS has become a prominent societal issue, especially in this social media driven time. BS identification can be very tricky for even humans, so we focus on BS-calling and the perception of BS on social media as a stepping-stone. In this dissertation we build a system that automatically determines what social media posts are likely to be called BS. First, we gather a curated crowd-sourced collection of BS-called tweets from Twitter over a 100 day period. Next, we argue that while it is difficult to assess a poster's intent of posting BS, we can mine the audience's perceptions of the poster's intent. We introduce four perspectives – attitude, sentiment, sincerity, and content – that can be leveraged through a set of linguistic footprints which we extract from each post. Our data-driven exploration of this curated dataset finds who and what topics have been called BS the most, and where BS callers and their targets come from. Finally, we develop a classification model that uses the four perspectives to differentiate between posts that are likely to be called BS and posts that are not. The results show the promising potential of our system as a preliminary BS discovery tool on social media.

5. A CROSS-CUTTING COMPONENT: LEARNING USER TOPICAL PROFILES*

In the previous three sections we have studied three concrete applications against three user misbehaviors on social media — spamming, manipulation, and distortion. As discussed in Section 1, all these applications of user misbehavior relate to the question of who those users are. With the aim of better understanding the relationship between users and misbehavior, we identify the problem of learning user topical profile on social media as a *cross-cutting* component that can improve our understanding of each specific misbehavior (while also providing the foundation for a number of other studies). Hence, we turn in this section to propose a generalized framework for learning user topical profiles, as follows.

5.1 Introduction

In social media systems, *demographic profiles* — often including name, age, gender, and location — provide an important first step toward creating rich user models for information personalization. For example, a user's location can be a signal to surface local content in the Facebook newsfeed. These demographic profiles typically reveal very little about a user's topical interests (what she likes) or expertise (what she is known for). Hence, there is great effort toward building high-quality *user topical profiles*, toward improving user experience and powering important applications like personalized web search [33], recommendation system [34, 35], expert mining [36], and community detection [37].

Indeed, there are two major approaches to build the topical profiles for social media users. One thread of methods seeks to uncover latent factors that may be descriptive of a

*Part of this section is reprinted with permission from "What Are You Known For?: Learning User Topical Profiles with Implicit and Explicit Footprints" by Cao C., Ge H., Lu H., Hu X., and Caverlee J. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). ACM, New York, NY, USA, 743-752. Copyright 2017 by ACM.

user. For example, running Latent Dirichlet Allocation (LDA) over a user's posts in social media can reveal the topics of interest of the user [141, 142, 143]; similarly, matrix factorization approaches have proven popular at capturing user factors, often for personalization purposes [144, 145, 146, 35, 33, 147, 148]. Aside from such recommendation applications, latent factor models have also been used to find influential users, mine communities, and predict review quality [149, 143, 37]. Another thread of methods seeks to encourage social media users to directly assess each other's interests and expertise, providing a partial perspective on user topical profiles. For example, LinkedIn users can choose *skill tags* for their own profiles and can endorse these tags on the profiles of others. *Twitter Lists* allow users to organize others according to user-selected keywords, e.g., placing a group of popular chefs on the list "Top Chefs". In this way, some list names can be viewed as a topical tag for list members. In the aggregate, this crowd-contributed tagging knowledge can be viewed as explicit evidence for capturing user interests and expertise [150, 36, 151].

Both approaches, however, face great challenges. Approaches that identify latent topics (often, as a distribution over features in some lower dimensional space) are typically trained only over content (ignoring other important footprints) and are difficult to directly interpret. Methods that only use crowdsourced tags typically suffer from limited coverage; that is, while the hand-curated tags may be of high-quality, very few users actually have descriptive topical tags associated with them. For example, in a random sample of 3.5 million Twitter users, we find that only 2% have been labeled with a topical tag (more details in Section 5.5). Moreover, to better understand user topical interests and expertise, a more comprehensive profiling framework is necessary. For instance, it is unclear what kind of evidence is useful for user topical profiling. And how can such potentially heterogeneous evidence be modeled for user topical profiling?

Hence, we propose to exploit heterogeneous footprints (e.g., tags, friends, interests, behavior) for intelligently learning user topical profiles. Based on a small set of explicit user tags, our goal is to extend this known set to the wider space of users who have no explicit tags. The key intuition is to identify "similar" users in terms of their topical profiles by exploiting their similarity in a *footprint space*. For instance, Twitter users who post similar hashtags may have similar interests, and YouTube users who upvote the same videos may have similar preferences. Such evidence of *homophily* has been widely studied in the sociological literature [38] and repeatedly observed in online social media, e.g., [39, 40, 41, 42, 43]. But what footprint spaces are appropriate for finding this homophily? What impact do they have on the discovery of user topical profiles? And which footprints are more effective at uncovering topical profiles?

Toward answering these questions, the rest of this section makes the following main contributions:

- First, we formulate the problem of learning user topical profiles in social media, with a focus on leveraging heterogeneous footprints.

- Second, we demonstrate how to model different footprints (e.g., like interests, social, and behavioral footprints) under this framework, and we present a unified 2-D factorization model in which we simultaneously consider all of these footprints (called *UTop*).

- Third, we then extend this initial approach through a generalized model that integrates the pairwise relations across all potential footprints via a tensor-based model (called *UTop+*), which provides a more robust framework for user profile learning.

- Finally, through extensive experiments, we find the proposed UTop+ model is capable of learning high-quality user topical profiles, and leads to a 10-15% improvement in precision and mean average error versus a state-of-the-art baseline. We find that behavioral footprints are the single strongest factor, but that intelligent integration of multiple footprints leads to the best overall performance.

5.2 Related Work

Finding User Interests and Expertise. Finding user interests and expertise has numerous applications, and one of the most popular tasks is personalized search and recommendation. Considerable research [34, 144, 146, 35, 142, 33, 152, 148] has been dedicated to uncover users' latent interests or expertise as their personal preferences for building recommender systems in different domains, such as web search [142, 33], web content [148], rating systems [153, 35], and social media [144, 145, 146, 152].

For social media research, the latent factor model is a state-of-the-art method for user recommendation. Interpreting the latent factors as topics, approaches based on such a model usually avoid explicitly identifying user interests but instead integrate the factors into a recommendation task. For example, Hong et al. applied matrix factorization on both users and tweets and focused on recommending user's retweeting behavior [145]. Similarly, Jiang et al. presented a probabilistic matrix factorization method to recommend whether a user adopts an item on a social network [146]. Zhong et al. collected user's webpage views to build a matrix factorization profile for web content recommendation [148].

Leveraging Footprints. A sequence of research has focused on using various footprints to learn user interests. One of the most traditional approaches is to model text-based footprints to obtain users' latent topical preferences, as in the case of PLSA and LDA [154, 155, 143, 37]. Another popular footprint is social (often via friendships) [146, 156, 157], with the natural homophily assumption that friends tend to have similar profiles. In addition, behavioral footprints have become a newer factor; for example, Guy et al. used a user's tagging behavior as evidence for content recommendation [34]. Lappas et al. considered *user endorsement* as a behavioral signal [158]. In [152], Zhao et al. focused on the behaviors of commenting, “+1”, and “like” on Google+. Some of the other footprints that have been explored in previous works include user’s emotions and sentiment, geo-location, temporal context and linguistic activity. For example, Hu et al. [159, 160] proposed an unsupervised factorization approach for user sentiment analysis through emotional signals. Lu et al. [161] considered user’s geographical footpints to discover what people are known for. Yin et al. [147] proposed a probabilistic graphical method to model user’s temporal interest for item recommendation. Hu et al. [162] applied a factorization method to infer linguistic properties of user’s documents. However, typically, these different footprints have been treated separately.

Factorization Models. Technically, it is challenging to embed users’ heterogeneous footprints into a factorization model. A handful of studies have adopted a regularization model [163, 149, 35] for personalized recommendation, though typically focusing on only one footprint. In [153], latent spaces are learned separately for each footprint through probablisitc matrix factorization assuming they are not independent. Tensor-based factorization methods [164, 165] have been used in many applications such as behavior modeling, healthcare, and urban planning [166, 167, 168]. A more comprehensive survey of tensor factorization and its applications can be found in [169]. In contrast, we first propose a factorization model in which we simultaneously consider multiple contexts via linearly weighted regularization. We then extend the model with a generalized tensor-based factorization so that not only different types of footprints can be considered together but their multi-linear interactions with each other can be exploited.

Several studies have focused on heterogeneous domains or entities, instead of contexts. Yu et al. put mutiple types of entities into a heterogeneous network and used a Bayesian ranking process to estimate user preferences [170]. Similarly, Hu et al. looked into a traditional user-item recommendation problem, presenting a factorization model across heterogeneous items. However, the network will quickly grow when users and items increase. Singh and Gordon proposed a framework to learn different types of relations, where they iteratively do matrix factorization between all pairs of domains [171]. Hu et al. [172] adopted the existing PARAFAC2 factorization algorithm on a tensor model, which is obtained by combining user ratings of different merchandises like book, music, and movie. Zhong et al. [148] directly applies a matrix factorization model on Web users and their clicked content items. However, in this work, we focus on learning user topical profiles rather than recommending item ratings for users.

Personalized Tag Recommendation. Another related research line focuses on personalized tag recommendation for users in social tagging systems [173, 174, 175, 176, 177, 178]. For example, Rendle et al. [175, 176] proposed tensor factorization to suggest tags to users for annotation on different items. Feng et al. [173] modeled social tagging as a multi-type graph and proposed random walk with restart for tag recommendation. Konstas et al. [174] also proposed a modified random walk with restart by exploiting social relationships and tagging for item recommendation. Our work is different from personalized tag recommendation in two aspects. The first is that we use crowdsourced tags to represent user's interests and expertise instead of annotating items in social systems. The second is that our problem is to infer users' topical profiles through tags for unknown users based on their different footprints rather than recommend tags based on partial knowledge of a user's profile.

5.3 Preliminaries

Explicit Footprints. Let $\boldsymbol{\mathcal{U}} = \{u_1, u_2, \ldots, u_N\}$ be a set of users where N is the number of users, and $\boldsymbol{T} = \{t_1, t_2, \ldots, t_M\}$ is a set of M tags each of which is associated with a particular topic. Suppose we have a subset of users $\boldsymbol{\mathcal{S}} \subset \boldsymbol{\mathcal{U}}$ where each user $u_i \in \boldsymbol{\mathcal{S}}$ has been labeled with a subset of $\boldsymbol{T}$, typically based on the collective efforts of the crowd. We refer to such labels as *explicit footprints*. Practical examples of explicit footprints include LinkedIn Skill Tags and Twitter Lists, wherein users can provide a crowdsourced summary of a user's interests and expertise [150, 36, 151]. We denote the explicit footprints as the user-tag matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}| \times M}$ in which element $\boldsymbol{P}(i, j)$ represents the number of times u_i is labeled by t_j.

Learning User Topical Profiles. Given a set of users $\boldsymbol{\mathcal{U}}$, a set of tags $\boldsymbol{T}$, and a subset of users $\boldsymbol{\mathcal{S}} \subset \boldsymbol{\mathcal{U}}$ for whom we know their user topical profiles $\boldsymbol{P}$, the problem of *Learning User Topical Profiles* is the task of inferring the unknown tags from $\boldsymbol{\mathcal{T}}$ for users in $\boldsymbol{\mathcal{U}} - \boldsymbol{\mathcal{S}}$.

An Initial Attempt with Explicit Footprints Only. A natural choice for attacking the challenge of learning user topical profiles is the matrix completion approach, which has been adopted in many related works [145, 171, 170, 148]. Under a matrix completion approach, we can extend $\boldsymbol{P}$ to a larger matrix $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ by including all users of $\boldsymbol{\mathcal{U}}$. Then, we can formulate the learning user topical profiles problem as a matrix completion problem:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V}} \quad & \frac{1}{2}\|\boldsymbol{\Omega} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2, \\ \text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \end{aligned} \tag{5.1}$$

where $\boldsymbol{X}$ is a user-tag matrix, and $\boldsymbol{U} \in \mathbb{R}^{N \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{M \times K}$ are latent representations of users and tags, respectively. $K \ll min(N, M)$ is the number of latent dimensions. Since the given $\boldsymbol{X}$ is naturally non-negative, we add the same constraints for $\boldsymbol{U}$ and $\boldsymbol{V}$ so that we can better interpret the values in them. $\boldsymbol{\Omega}$ is a non-negative matrix with the same size of $\boldsymbol{X}$:

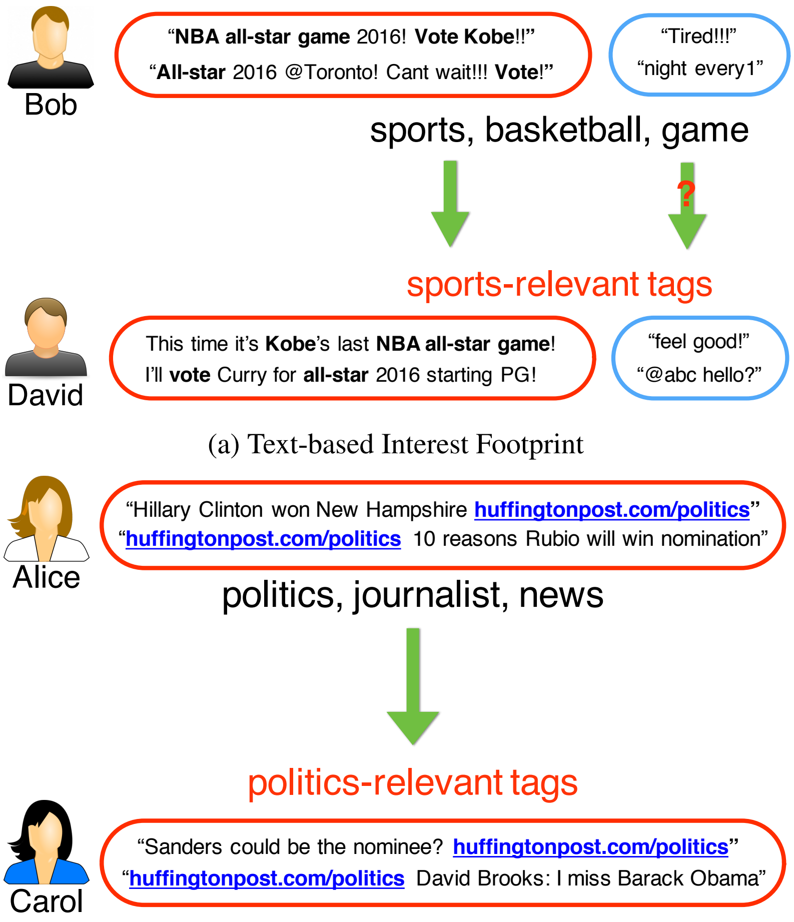
$$\boldsymbol{\Omega}(i, j) = \begin{cases} 1 & \text{if } \boldsymbol{X}(i, j) \text{ is observed,} \\ 0 & \text{if } \boldsymbol{X}(i, j) \text{ is unobserved.} \end{cases}$$

The basic matrix completion model above learns an optimal set of $\{\boldsymbol{U}, \boldsymbol{V}\}$ to approximate the original matrix $\boldsymbol{X}$, estimating for unobserved users through observed user-tag pairs. However, as in many linear-inverse problems, there may not be sufficient information to estimate the original matrix $\boldsymbol{X}$ based only on the partially observed data. The problem of learning user topical profiles is one such case, since most of our target users do not have any partially explicit footprint.

Implicit Footprints. With the scarcity of explicit footprints in mind, we are interested to explore the potential of *implicit footprints* for learning unknown user topical profiles. Implicit footprints may indirectly reflect user interests or expertise. Typical implicit footprints, for example, could include user behaviors, the social circle of a user, sentiment-based features of a user's posts, the geo-location of a user, emotional cues, and temporal dynamics, among many others [34, 159, 160, 162, 146, 158, 161, 156, 157, 147, 152]. The key intuition is to identify "similar" users in terms of their topical profiles by exploiting their similarity via these implicit footprints. Since evidence from these heterogeneous implicit footprints may provide conflicting evidence, potentially leading to lower quality user profiles than considering footprints in isolation, we propose a generalized optimization framework that takes into account pairwise relations among all possible implicit footprints for learning user profiles. In this way, the benefits of each footprint may be intelligently combined to find the best evidence across multiple implicit footprints for learning high-quality user profiles.

5.4 Learning User Topical Profiles

We turn in this section to propose a generalized model for learning user topical profiles. We first identify multiple implicit footprints and demonstrate how to model them. We then introduce a matrix factorization based approach — called UTop, before extending this version to a more general tensor-based approach — called UTop+.


Bob
"NBA all-star game 2016! Vote Kobe!!"
"All-star 2016 @Toronto! Cant wait!!! Vote!"
"Tired!!!"
"night every1"
sports, basketball, game
?
sports-relevant tags
David
This time it's Kobe's last NBA all-star game!
I'll vote Curry for all-star 2016 starting PG!
"feel good!"
"@abc hello?"
Alice
"Hillary Clinton won New Hampshire huffingtonpost.com/politics"
"huffingtonpost.com/politics 10 reasons Rubio will win nomination"
politics, journalist, news
politics-relevant tags
Carol
"Sanders could be the nominee? huffingtonpost.com/politics"
"huffingtonpost.com/politics David Brooks: I miss Barack Obama"

(a) Text-based Interest Footprint

(b) Behavioral Footprint

Figure 5.1: Examples of different implicit footprints on learning user topical profiles

5.4.1 Modeling Implicit Footprints

We aim to integrate many different kinds of implicit footprints into the framework for learning user topical profiles. For the concreteness in our discussion, we focus in this section on three specific types of implicit footprints that capture three different perspectives on user topical profiles. The three footprints are: *social*, based on the friends (via the social graph) around the user; *interest*, based on the text posts made by the user; and *behavioral*, based on the link sharing activities of the user. The intuition is that these varied implicit footprints can connect related users, such that user topical profiles can be propagated from user to user. But how should we model these kinds of implicit footprints? And how can we integrate them into a matrix completion model? Note that the proposed model can be easily extended to incorporate additional footprints.

Social Footprints. Social footprints — directly suggested by homophily — naturally indicate that connected users may share common interests, and hence can be used for inferring user topical profiles [34, 146, 156, 157]. For example, if Carol and David are following each other on Twitter, the social footprint suggests that it is more likely for them to share common interests.

These social network connections between users can be naturally modeled as a matrix. We denote the matrix as $\boldsymbol{E} \in \mathbb{R}^{N \times N}$ in which the binary element $E(i, j)$ represents if user u_i and user u_j have a connection on a social network. We can model this social footprint as a regularization term:

$$\mathcal{L}_1 = \frac{1}{2}\|\boldsymbol{E} - \boldsymbol{U}\boldsymbol{U}^T\|_F^2.$$

Our goal is to optimize the user latent matrix $\boldsymbol{U}$ in order to minimize $\mathcal{L}_1$, with the intuition that friends are likely to have similar profiles. Of course, users may form relationships in social media for many diverse reasons, and so these relationships may not be appropriate for inferring similar topical profiles. As one example, family members may be "friends" in a social network but can have distinct topical profiles (e.g., sister vs brother, grandson vs grandfather). Hence, we next consider additional implicit footprints that may serve to mitigate these challenges.

Interest Footprints. The second footprint we consider is based on user interests. Texts posted by users can semantically reflect related subjects associated with their interests or expertise. Thus, many studies have directly applied LDA on posted texts, assuming the (latent) topics in user's posts are their topical profiles [141, 142, 143]. In Figure 5.1a, Alice is a basketball fan and she has posted many tweets talking about the upcoming NBA all-star game. We find Bob's tweets share many of the same words as Alice's. Hence, their posted texts demonstrate their shared interests in basketball, suggesting that Alice's user topical profile may be similar to Bob's.

We can model this text-based interest footprint like so: let $\boldsymbol{w} = \{w_1, w_2, \ldots, w_L\}$ be the set of words, where L denotes the number of words. $\boldsymbol{A} \in \mathbb{R}^{N \times L}$ is a user-word matrix in which $\boldsymbol{A}(i,j)$ is the frequency of word w_j appearing in user u_i's posts. Similarly, $\boldsymbol{B} \in \mathbb{R}^{M \times L}$ is a tag-word matrix where $\boldsymbol{B}(i,j)$ represents the frequency of word w_j posted by all users who have tag t_i. We propose to leverage a user's interest footprint as the following loss function:

$$\mathcal{L}_2 = \frac{1}{2}\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2 + \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T\|_F^2,$$

where $\boldsymbol{W} \in \mathbb{R}^{L \times K}$ represents word's latent topics. Our goal is to minimize $\mathcal{L}_2$ so that two users who are "nearby" in the interest footprint space tend to have similar topical profiles. However, a user's posts are often short (like on Twitter) and may contain many nonsense or off-topic texts, which can interfere with clearly revealing user topical profiles. Hence, we next turn to a third footprint for overcoming these issues.

Behavioral Footprints. Finally, we propose to augment the social and textual footprints with behavioral footprints [34, 158, 152]. According to the homophily evidence in the behavior dimension [38], for instance, two YouTube users may have close tastes if they usually "like" or "dislike" the same videos. A retweet on Twitter is a strong indication of the retweeter's personal endorsement, so two users can have similar preferences if they often retweet the same tweets. Hence, these behavioral footprints may provide strong evidence beyond who users are connected to (social) and what they post (interests).

In this dissertation we adopt link sharing as a public, observable behavior that may serve as a first step toward improving the learning of user topical profiles. Other behavioral footprints are possible, and we anticipate revisiting these in our future work. Link sharing behavior for topical profiles has received some attention in social media research. Previous work looked into why and what content people share via links in social media [71, 69]. Some other work has mentioned the role of link sharing in social spamming [76]. Through link sharing, users can concisely express their viewpoints, interests, and professional expertise. For instance, a person who works in the IT industry may usually post link of `engadget.com`. A user who likes sports may often share links of `espn.com`. In Figure 5.1b, Carol is a political journalist so she regularly posts some links of `huffingtonpost.com`, and we see David also usually shares the same links. In this case we may infer politics-relevant tags for David.

Concretely, let $\boldsymbol{Z} = \{z_1, z_2, \ldots, z_P\}$ be the set of links posted by users. Similar to the interest footprint, we define $\boldsymbol{C} \in \mathbb{R}^{N\times P}$ as a user-link matrix where $\boldsymbol{C}(i,j)$ is the frequency of link z_j posted by user u_i. Also, $\boldsymbol{D} \in \mathbb{R}^{M\times P}$ is a tag-link matrix with $\boldsymbol{D}(i,j)$ as the frequency of link z_j appearing in all posts from users having tag t_i. As a result, we leverage link sharing via the following loss function:

$$\mathcal{L}_3 = \frac{1}{2}\|\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T\|_F^2 + \frac{1}{2}\|\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T\|_F^2,$$

where $\boldsymbol{G} \in \mathbb{R}^{P\times K}$ represents link's latent topical spaces. Our goal is to minimize $\mathcal{L}_3$, with the idea that users may have similar topical profiles if they behave similarly when posting links.

5.4.2 Learning User Topical Profiles: A 2-D Model

Since evidence from multiple implicit footprints may provide conflicting evidence, potentially leading to lower quality user profiles than considering footprints in isolation, we turn in this section to developing a unified model that can integrate all possible het-

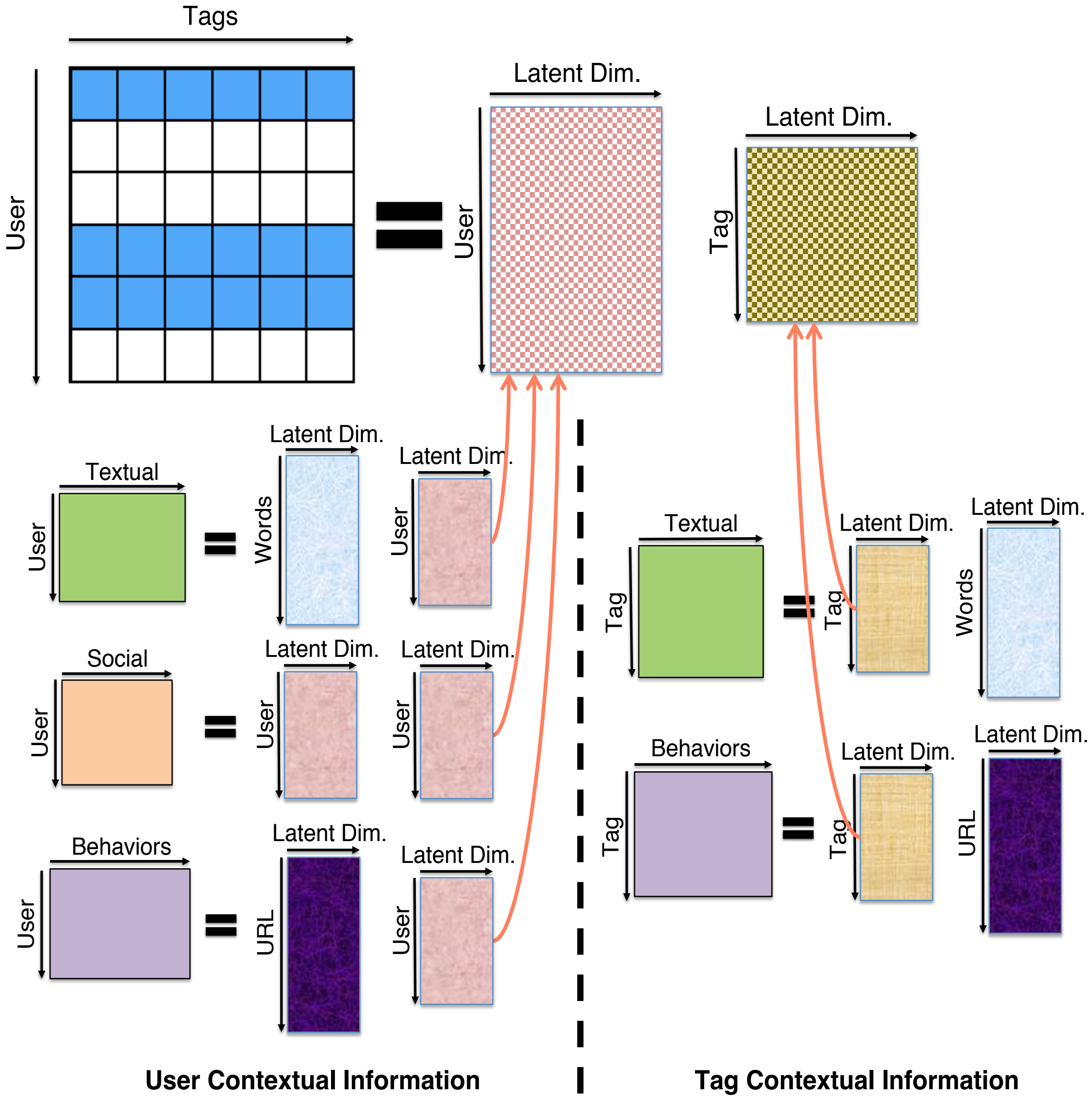


Figure 5.2: An overview of the 2-D model (UTop)

erogeneous footprints together into a matrix (2-D) completion model. Since all implicit footprints are modeled as regularization terms in Section 5.4.1, intuitively we can linearly incorporate them into the proposed *UTop* model. Again, recall that we focus our presentation here on those three specific footprints (social, interest, behavioral), but the model is designed to generalize to more alternative footprints as well.

Figure 5.2 gives an overview of UTop. In general, we factorize each of the social, interest, and behavioral footprint matrices, and assume that the objective user-tag matrix shares the same latent user dimensions with them. This is the fundamental assumption in most factorization-based methods for solving matrix completion problems. We also consider explicit footprints. Similarly, we collect each tag's latent representation, and multiply them with each user's latent factor for estimating the objective matrix.

Concretely, we formulate the following optimization problem as following:

$$\begin{aligned} \min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{G}} \quad & \mathcal{F} = \frac{1}{2}\|\boldsymbol{\Omega} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2 \\ & + \frac{\lambda}{2}(\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2 + \|\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T\|_F^2) \\ & + \frac{\gamma}{2}(\|\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T\|_F^2 + \|\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T\|_F^2) \\ & + \frac{\delta}{2}\|\boldsymbol{E} - \boldsymbol{U}\boldsymbol{U}^T\|_F^2 \\ & + \frac{\alpha}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 + \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{G}\|_F^2) \\ \text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0, \boldsymbol{W} \geq 0, \boldsymbol{G} \geq 0, \end{aligned} \tag{5.2}$$

where λ, γ, δ and α are positive regularization parameters controlling the contributions of different implicit footprints. $\|\boldsymbol{U}\|_F^2$, $\|\boldsymbol{V}\|_F^2$, $\|\boldsymbol{W}\|_F^2$ and $\|\boldsymbol{G}\|_F^2$ are deployed to avoid overfitting. Similar to Equation 5.1, we insert the non-negative constraints for $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$, and $\boldsymbol{G}$.

The derivation of the objective function in Eq.(5.2) regarding four variables $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$

and $\boldsymbol{G}$ are demonstrated as:

$$\begin{aligned}
\frac{\partial \mathcal{F}}{\partial \boldsymbol{U}} =& -\boldsymbol{\Omega} \odot \boldsymbol{\Omega} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} - \lambda(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T) \\
& - \gamma(\boldsymbol{C} - \boldsymbol{U}\boldsymbol{G}^T) - 2\delta(\boldsymbol{E} - \boldsymbol{U}\boldsymbol{U}^T) + \alpha\boldsymbol{U}, \\
\frac{\partial \mathcal{F}}{\partial \boldsymbol{V}} =& -\boldsymbol{\Omega}^{\boldsymbol{T}} \odot \boldsymbol{\Omega}^{\boldsymbol{T}} \odot (\boldsymbol{X}^T - \boldsymbol{V}\boldsymbol{U}^T)\boldsymbol{U} - \lambda(\boldsymbol{B} - \boldsymbol{V}\boldsymbol{W}^T) \\
& - \gamma(\boldsymbol{D} - \boldsymbol{V}\boldsymbol{G}^T) + \alpha\boldsymbol{V}, \\
\frac{\partial \mathcal{F}}{\partial \boldsymbol{W}} =& -\lambda(\boldsymbol{A}^T - \boldsymbol{W}\boldsymbol{U}^T)\boldsymbol{U} - \lambda(\boldsymbol{B}^T - \boldsymbol{W}\boldsymbol{V}^T)\boldsymbol{V} + \alpha\boldsymbol{W}, \\
\frac{\partial \mathcal{F}}{\partial \boldsymbol{G}} =& -\gamma(\boldsymbol{C}^T - \boldsymbol{G}\boldsymbol{U}^T)\boldsymbol{U} - \gamma(\boldsymbol{D}^T - \boldsymbol{G}\boldsymbol{V}^T)\boldsymbol{V} + \alpha\boldsymbol{G}.
\end{aligned} \tag{5.3}$$

Based upon these derivations, we then apply stochastic gradient descent to iteratively update each variable by taking a step η along its gradient ascending. The algorithm details are presented in **Algorithm 1** in which learning steps η_u, η_v, η_w and η_g are chosen based upon the Goldstein Conditions [179]. We implement the non-negative constraints on $\boldsymbol{U}$ and $\boldsymbol{V}$ through forcing their negative values to 0 in each iteration. As shown, this algorithm considers all three footprints together to estimate the topical profiles for each user.

```
Algorithm 1: UTop solver
Input: user-tag matrix $\boldsymbol{X}$, user-word matrix $\boldsymbol{A}$, tag-word matrix $\boldsymbol{B}$, user-url matrix
       $\boldsymbol{C}$, tag-url matrix $\boldsymbol{D}$, user friendship matrix $\boldsymbol{E}$, observation indication
       matrix $\boldsymbol{\Omega}$ and parameters $\{\lambda, \gamma, \delta, \rho, \eta\}$
Output: $\boldsymbol{U}, \boldsymbol{V}$
Initialize $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$ and $\boldsymbol{G}$ randomly, $t = 0$
while Not Converged do
    Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, $\frac{\partial \mathcal{F}}{\partial \boldsymbol{V}}$, $\frac{\partial \mathcal{F}}{\partial \boldsymbol{W}}$ and $\frac{\partial \mathcal{F}}{\partial \boldsymbol{G}}$ in Eq.(5.3)
    Update $\boldsymbol{U}_{t+1} \leftarrow$ max($\boldsymbol{U}_t - \eta_u \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    Update $\boldsymbol{V}_{t+1} \leftarrow$ max($\boldsymbol{V}_t - \eta_v \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    Update $\boldsymbol{W}_{t+1} \leftarrow$ max($\boldsymbol{W}_t - \eta_w \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    Update $\boldsymbol{G}_{t+1} \leftarrow$ max($\boldsymbol{G}_t - \eta_g \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    $t = t + 1$
return $\boldsymbol{U}$ and $\boldsymbol{V}$
```

Though unifying all three heterogeneous implicit footprints, this initial UTop approach has two main drawbacks. First, it will become complex if we introduce additional footprints, as we bring in more controlling parameters of new footprints to be tuned. In addition, UTop does not take into account the relations between those heterogeneous footprints which could be jointly explored in the latent space. Given these concerns, can we find a generalized model that can jointly leverage all potential heterogeneous footprints? We turn in the following section to answering this question.

5.4.3 Learning User Topical Profiles: A Generalized Model

In this section, we augment UTop with a generalized approach toward jointly exploring the relationships across footprints for more robust user topical profile learning. First, to relieve the dramatic increase of parameters when introducing more regularization terms, we need to replace the linear combination model in UTop by a more compact factorization model without manually tuning tradeoff parameters from different new footprints. Second, such a compact factorization model should consider all possible pairwise interactions between footprints to exploit their multi-linear relationships. Therefore, we adopt a *tensor factorization* model which explicitly takes into account the multi-way structure of data. Moreover, the factorization will only happen once even if we introduce additional heterogeneous footprints.

Figure 5.3 shows an overview of UTop+. In general, we model all implicit footprints in one tensor via calculating the user similarity in each footprint space. There can be many options for measuring the user similarity in every footprint space. We test many of them and report the one providing the best performance in Section 4.5. Then, we factorize the tensor and obtain a matrix of latent representations for all users, upon which we extract a user similarity matrix to estimate the original user-tag matrix.

Concretely, we denote the tensor as $\mathcal{C} \in \mathbb{R}^{N \times N \times R}$ which is a multidimensional array

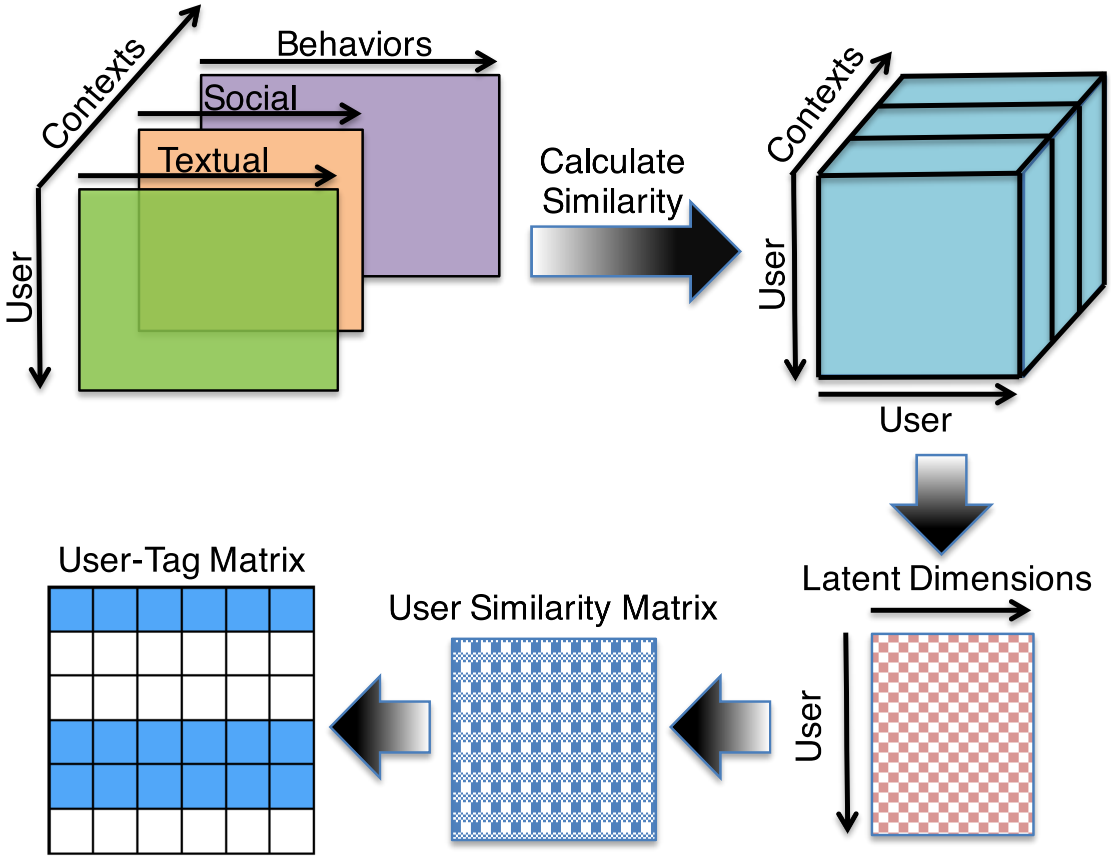


Figure 5.3: An overview of the generalized model (UTop+)

where R is the number of implicit footprints and N is the size of the user set. We can factorize the tensor $\boldsymbol{\mathcal{C}}$ to one latent user matrices $\boldsymbol{Q} \in \mathbb{R}^{N \times K}$ and one latent context matrix $\boldsymbol{Y} \in \mathbb{R}^{R \times K}$, where K is the number of latent dimensions. The tensor factorization is to solve the optimization problem defined below:

$$\min_{\boldsymbol{Q},\boldsymbol{Y}} \quad \frac{1}{2}\|\boldsymbol{\mathcal{C}} - [\![\boldsymbol{Q}, \boldsymbol{Q}, \boldsymbol{Y}]\!]\|_F^2 + \frac{\alpha}{2}(\|\boldsymbol{Q}\|_F^2 + \|\boldsymbol{Y}\|_F^2), \tag{5.4}$$

where $[\![\boldsymbol{Q}, \boldsymbol{Q}, \boldsymbol{Y}]\!] \in \mathbb{R}^{N \times N \times R}$ is given by

$$[\![\boldsymbol{Q}, \boldsymbol{Q}, \boldsymbol{Y}]\!] = \sum_{k=1}^{K} \boldsymbol{q}_k \circ \boldsymbol{q}_k \circ \boldsymbol{y}_k.$$

Here $\boldsymbol{q}_k$ and $\boldsymbol{y}_k$ are the k^{th} column vectors of $\boldsymbol{Q}$ and $\boldsymbol{Y}$, respectively. To solve Eq.(5.4), we adopt the existing CPOPT method [180] — a fitting approach for the CP (Canonical-decomposition / Parallel-factor-analysis (PARAFAC)) model. The latent footprint matrix $\boldsymbol{Y}$ represents the contribution of each type of footprint to latent dimensions.

The next natural question is how to leverage the new latent space $\boldsymbol{Q}$ of all users. The basic idea is that two users tend to have similar topical profiles if they have similar latent representations derived by jointly considering all their implicit footprints. Thus, we first calculate the user similarity matrix denoted as $\boldsymbol{\Psi}$ computed from latent features of users $\boldsymbol{Q}$ by the cosine similarity. We can see $\boldsymbol{Q}$ as a "new footprint" and formulate it as the new loss function:

$$\begin{aligned}
\boldsymbol{\Theta} &= \frac{1}{2} \sum_{i,j} \Psi(i,j) \|\boldsymbol{U}_i - \boldsymbol{U}_j\|^2 \\
&= \sum_{i,j} \boldsymbol{U}_i \Psi(i,j) \boldsymbol{U}_i^T - \sum_{i,j} \boldsymbol{U}_i \Psi(i,j) \boldsymbol{U}_j^T \\
&= \sum_{i} \boldsymbol{U}_i \boldsymbol{D}(i,i) \boldsymbol{U}_i^T - \sum_{i,j} \boldsymbol{U}_i \Psi(i,j) \boldsymbol{U}_j^T \\
&= \text{tr}(\boldsymbol{U}^T (\boldsymbol{D} - \boldsymbol{\Psi}) \boldsymbol{U}) \\
&= \text{tr}(\boldsymbol{U}^T \boldsymbol{\mathcal{L}} \boldsymbol{U}),
\end{aligned} \tag{5.5}$$

where $\boldsymbol{U}_i$ is the ith row of $\boldsymbol{U}$, $tr(\cdot)$ denotes the matrix trace, and $\boldsymbol{D}$ is a diagonal matrix in which $\boldsymbol{D}(i,i) = \sum_j \boldsymbol{\Psi}(i,j)$, and $\boldsymbol{\mathcal{L}} = \boldsymbol{D} - \boldsymbol{\Psi}$ is the graph Laplacian of the user similarity matrix $\boldsymbol{\Psi}$.

How can we utilize the new implicit footprint $\boldsymbol{\Theta}$ to learn user topical profiles? Similarly, we are able to use $\boldsymbol{\Theta}$ to regulate latent representations of two similar users to make them as close as possible. Hence, we can build the generalized UTop+ by solving the following optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{U},\boldsymbol{V}} \quad & \frac{1}{2}\|\boldsymbol{\Omega} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T)\|_F^2 + \frac{\beta}{2}\text{tr}(\boldsymbol{U}^T \boldsymbol{\mathcal{L}} \boldsymbol{U}) \\
& + \frac{\alpha}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2), \\
\text{s. t.} \quad & \boldsymbol{U} \geq 0, \boldsymbol{V} \geq 0,
\end{aligned}
\tag{5.6}
$$

where β is the controlling parameter. This optimization problem can be solved similarly as introduced in Section 5.4.2. The detailed solver is presented in **Algorithm 2**.

Algorithm 2: UTop+ solver

```
Input: user-tag matrix $\boldsymbol{X}$, user-word matrix $\boldsymbol{A}$, user-url matrix $\boldsymbol{C}$, user friendship
       matrix $\boldsymbol{E}$, observation indication matrix $\boldsymbol{\Omega}$ and parameters $\{\alpha, \beta, \eta_u, \eta_v\}$
Output: $\boldsymbol{U}, \boldsymbol{V}$
Calculate the tensor $\boldsymbol{\mathcal{C}}$ from $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{E}$
Calculate $[\boldsymbol{Q}, \boldsymbol{Y}] \leftarrow$ CPOPT($\boldsymbol{\mathcal{C}}$)
Calculate the user similarity matrix $\boldsymbol{\Psi}$ based on $\boldsymbol{Q}$
Construct the graph Laplacian matrix $\boldsymbol{\mathcal{L}}$ for $\boldsymbol{\Psi}$
Initialize $\boldsymbol{U}$ and $\boldsymbol{V}$, randomly, $t = 0$
while Not Converged do
    Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{U}} = -(\boldsymbol{\Omega} \odot \boldsymbol{\Omega})(\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^T))\boldsymbol{V} + \beta \boldsymbol{\mathcal{L}} \boldsymbol{U}$
    Compute $\frac{\partial \mathcal{F}}{\partial \boldsymbol{V}} = -(\boldsymbol{\Omega}^T \odot \boldsymbol{\Omega}^T)(\boldsymbol{X}^T - \boldsymbol{V}\boldsymbol{U}^T))\boldsymbol{U}$
    Update $\boldsymbol{U}_{t+1} \leftarrow$ max($\boldsymbol{U}_t - \eta_u \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    Update $\boldsymbol{V}_{t+1} \leftarrow$ max($\boldsymbol{V}_t - \eta_v \frac{\partial \mathcal{F}}{\partial \boldsymbol{U}}$, 0)
    $t = t + 1$
return $\boldsymbol{U}$ and $\boldsymbol{V}$
```

In summary, we first present a 2-D model for learning user topical profiles (called UTop) in which each of three heterogeneous implicit footprints is modeled as regularization terms. We provide Algorithm 1 to solve the optimization problem in Equation 5.2.

Then we extend UTop to a compact generalized model (called UTop+). Based on a tensor decomposition method, UTop+ can jointly handle relationships across multiple footprints without introducing new parameters. The complete overview of UTop+ is shown in Figure 5.3, and we propose Algorithm 2 to solve Equation 5.6.

5.5 Experiments

In this section, we conduct a series of experiments to answer the following questions: (i) How well do the proposed UTop and UTop+ models work? (ii) Which implicit footprints are most effective? (iii) How does UTop+ compare with other alternatives? Does it really improve upon the simpler UTop approach? (iv) How do the proposed approaches compare to other variants?; and (v) What impact do the model parameters have on the ultimate performance? We begin by introducing the experimental setup including dataset collection and evaluation method.

5.5.1 Experiment Setup

In this section, we start with describing the data we collect. Next, we introduce the metrics we use for evaluation. Finally, we provide the details of three baselines, and show the parameter settings we choose in our proposed models.

Twitter Lists. We adopt Twitter Lists, a large publicly-accessible collection of crowd-contributed tagging knowledge for social media users. Recall that these lists allow one user to annotate another with a list name (or tag), e.g., politics, music, art. Via the public Twitter API, we randomly sample a set of 3.468 million Twitter users, and crawl the list membership information for each of them. We identify 977,000 users who have ever been included in some list, but we find a huge amount of noise. For instance, nonsense tags (like numbers, unicode characters, single letters) take up a major proportion. Many tags (e.g., “friend”, “love”, and “amigo”) are not reflective of topical profiles. Also, there exist many near-synonyms and variants such as “writer-author” and “news-noticia”. To obtain high-quality tags for our problem, we rank all tags by the number of labeled users, and manually curate the top-500 tags through merging variants and filtering noise.

Implicit Footprints. For interest footprints, we aggregate all terms each user has posted and adopt the standard LDA topic model after filtering stopwords and stemming. We further measure user similarity by calculating the pairwise Jensen-Shannon divergence. For social footprints, we crawl the friendship connection information for each user. Following a user can be quite casual on Twitter, so we focus on mutual followings as the basis of user similarity in the social footprint space. For behavioral footprints, we aggregate all links a user has posted in her tweets and obtain the posting counts. We resolve all crawled URLs (most are shortened) to take care of URL variants, and focus on the URL domain name which conceptually represents a website. For quantifying similar link sharing patterns, we test a set of measurements (e.g., intersection, cosine, jaccard) and find the one in [181] works best.

Users. We collect a set of 72,096 users who have all those three types of implicit footprints and have been labeled by at least one of the candidate tags. Since many of them have sparse tagging information, we rank all users by the number of tags they have. We look into the top 50,000 users, and randomly select 10,000 users for training and evaluation.

In our proposed models, we end up with scores of all candidate tags for each user. Since we should take those most associated tags as user topical profiles, we rank them in descending order and focus on the top-k ranked tags. Our evaluation is based on ten-fold cross validation.

Metrics. We pick several metrics which can cover different evaluation aspects. On the one side, we would like to see the ratio of correct inferences for learning user topical profiles. And on the other side, we want to measure the prediction error. Thus, we adopt *precision@k* which measures the percentage of correctly estimated top-k tags, and *Mean*

Absolute Error (MAE) which quantifies the prediction quality in terms of errors. Note that a lower MAE means a better performance.

Furthermore, besides the absolute measurement in accuracy, the relative ranking order is another important perspective, especially in some recommendation scenarios. The rank correlation coefficients of both *Kendall's* τ and *Spearman's* ρ are two prevalent metrics for measuring rank-based agreement across two lists. We use them both to measure the number of pairs of tags that are correctly ordered from our results. Their values both range from -1 to 1, with the higher the more relevant.

Baselines. We select three baselines as alternatives to the proposed UTop+ approach. To be fair, we incorporate all three proposed footprints and maintain the same experimental setup for all the following approaches:

- **Nearest Neighborhood (NN).** An intuitive solution is based on the traditional nearest neighborhood model. A user is modeled by a vector extracting from the corresponding row in the context matrix, i.e., $\boldsymbol{A}$, $\boldsymbol{C}$, or $\boldsymbol{E}$. Then, for each target user, we separately find a set of closest seed users in each context, and pick the intersected neighbors from whom we propagate their tags and scores and take the average for each tag.

- **Cross-domain Triadic Factorization (CTF)** [172]. This state-of-the-art method directly combines user ratings of different merchandise (e.g., book, music, movie) into one tensor model, in which all the values are user ratings. Then, it extends the existing PARAFAC2 model [182] that transforms heterogeneous user-rating matrices of different lengths into one cubical tensor and factorizes it. Here in our problem setting, this approach can also be applied on those heterogeneous user-footprint matrices; the subsequent steps follow Equation 5.5 in order to solve Equation 5.6.

- **UTop.** Introduced in Section 5.4.2, this model is a basic version that considers each footprint as a regularization term and linearly adds them together.

Parameter Settings. To determine the number of latent dimensions in both UTop and UTop+, we experiment with a sequence of settings {5, 10, 20, 30, 40, 50, 100} and empirically select 20 for both UTop and UTop+, as a trade-off between accuracy and efficiency. In Algorithm 1, there are five parameters λ, γ, δ, α, and η. The first four parameters are used to control the contributions of various footprints. The last one is a step along its gradient ascending. As is commonly done, we iteratively employ cross-validation to tune these parameters. Specifically, we empirically set $\lambda = 0.02$, $\gamma = 0.7$, $\beta = 0.1$, $\alpha = 0.4$ and $\eta = 0.05$ for general experiments, respectively. In UTop+, we choose 10 for the number of latent dimension in tensor factorization. The step size η is set to 0.05. In addition, two positive parameters α and β in Eq. (5.6) are involved in the experiments. Concretely, we empirically set $\alpha = 0.3$ and $\beta = 0.02$ via cross-validation.

5.5.2 The Impact of Different Footprints

In general, interest, social, and behavioral footprints have different emphases on user topical profiles. Hence, which footprints work better (or best) is one of the most compelling questions to answer. Hence, we compare different combinations of all footprints in both NN and UTop. The reason we do not test them in UTop+ is that the multi-way manner of UTop+ may not clearly tell which footprint contributes more. We show the results in Table 5.1 in which T is for text-based interest, S is for social, and B is for behavioral.

When individually using each implicit footprint, we find the behavioral footprint (link sharing) always performs the best in any setting. Moreover, combining it with other footprints always bring the biggest improvement in these experiments. For instance, within the NN method, the behavioral footprint has up to 24$\%$ larger Spearman correlation than the social footprint. In UTop, the MAE@10 decreases by 8$\%$ when the behavioral footprint is added with the interest footprint. These results indicate the importance of capturing actual

Table 5.1: The impact of different implicit footprints for learning user topical profiles

Method	**Precision**			**MAE**			**Kendall's τ**			**Spearman's ρ**		
	Top 5	**Top 10**	**Top 15**	**Top 5**	**Top 10**	**Top 15**	**Top 5**	**Top 10**	**Top 15**	**Top 5**	**Top 10**	**Top 15**
NN (T)	0.2113	0.2356	0.2673	0.2914	0.2692	0.2432	0.2460	0.1687	0.1531	0.3054	0.2262	0.1784
NN (S)	0.1920	0.2153	0.2330	0.3048	0.2791	0.2642	0.2110	0.1420	0.1289	0.2670	0.1852	0.1682
NN (B)	0.2423	0.2629	0.3155	0.2650	0.2342	0.2110	0.2826	0.2044	0.1834	0.3314	0.2429	0.2106
UTop (T)	0.3438	0.3791	0.4668	0.2264	0.2069	0.1897	0.3221	0.2464	0.2031	0.4163	0.2987	0.2409
UTop (S)	0.3390	0.3837	0.4561	0.2298	0.2093	0.1887	0.3172	0.2421	0.2003	0.4135	0.2916	0.2341
UTop (B)	0.3556	0.3980	0.4733	0.2275	0.1982	0.1699	0.3286	0.2557	0.2067	0.4302	0.3015	0.2426
UTop (T+S)	0.3494	0.3847	0.4657	0.2300	0.2107	0.1872	0.3205	0.2516	0.2085	0.4189	0.2970	0.2378
UTop (T+B)	0.3587	0.4132	0.4758	0.2193	0.1894	0.1909	0.3329	0.2606	0.2197	0.4348	0.3071	0.2535
UTop (S+B)	0.3544	0.4069	0.4729	0.2238	0.1930	0.1852	0.3272	0.2588	0.2185	0.4322	0.3054	0.2561
UTop (T+S+B)	**0.3616**	**0.4189**	**0.4931**	**0.2137**	**0.1861**	**0.1772**	**0.3403**	**0.2746**	**0.2267**	**0.4414**	**0.3104**	**0.2682**

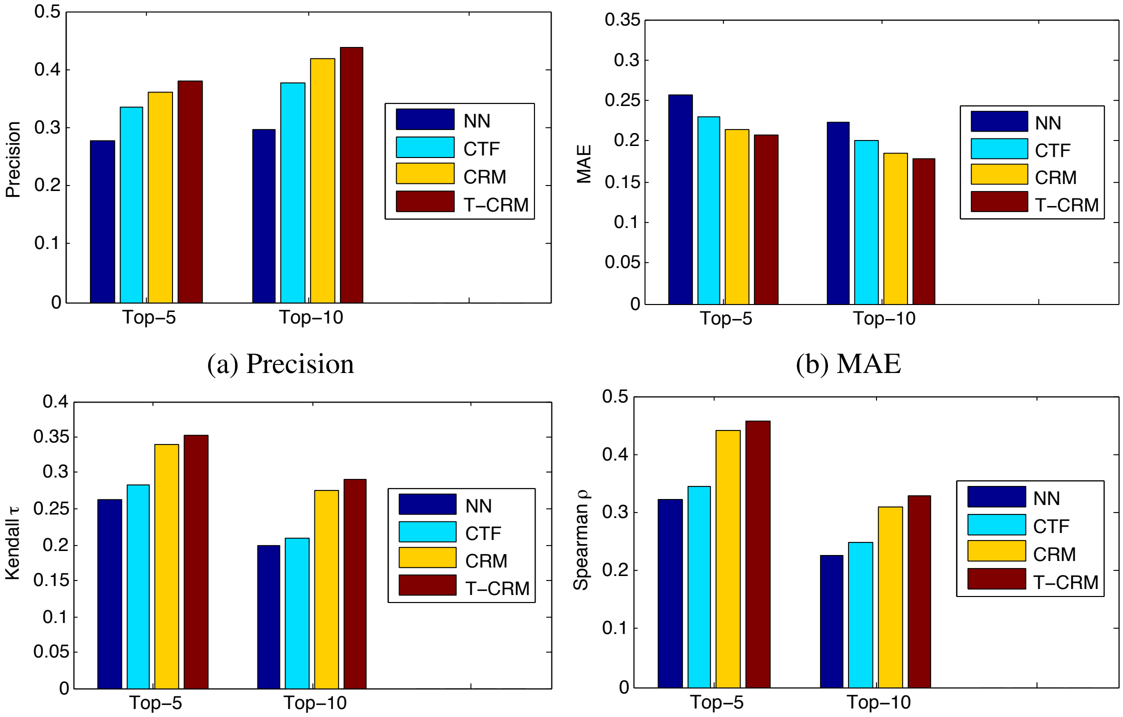
user behaviors as a critical step for identifying user topical profiles (in contrast, to relying purely on social connections or on the content of what users post). These results support the intuition that social footprints may capture spurious user similarities (e.g., linking two very different users) and that text-based interest footprints may insert noise into learning user topical profiles. In contrast, behavioral cues provide a clearer perspective on user interests and expertise.

What if behavioral data is scarce? Link sharing is one of the few publicly-available sources of behavioral information, but sometimes it can still be a scarce resource because not all users will share many links on social media. In contrast, social and interest-based footprints are typically more universally available. We see in Table 5.1 that interest and social footprints can still work well even without access to behavioral footprints. For example, in UTop, the interest footprint is only 5% behind behavioral in precision@10, and the social footprint has just 1% larger MAE@5 than behavioral. These observations show that our model can still achieve a good performance even when we have scarce behavioral evidence. But that together, the three different footprints can complement each other, leading to even better user topical profiles.

5.5.3 Evaluating UTop and UTop+

Given the evidence of the importance of different footprints, we now turn to evaluating the two proposed models — UTop and UTop+ — versus alternatives. As we can see in Figure 5.4, both UTop and UTop+ perform better than the Nearest Neighbor (NN) and the Cross-domain Triadic Factorization (CTF) across all four evaluation metrics. For precision@5, UTop+ is 36% and 13% better than NN and CTF with p-values of 0.001 and 0.003 under McNemar's test, respectively. For MAE@10, UTop+ outperforms NN by 20% with the p-value of 0.002 and CTF by 11.8% with the p-value of 0.001. The gaps become even larger for the two ranking correlation coefficients, as we can see in Figure 5.4c and 5.4d. These results suggest that the proposed learning models can better leverage all footprints together than either the neighborhood-based propagation or the immediate tensor decomposition. Note that the CTF method is fundamentally different from our problem setting where we cannot simply put together all heterogeneous footprints. In contrast, we exploit latent factors to build a user similar matrix and find its graph Laplacian as a new regularization term. We show the effectiveness of this step in Section 5.5.4.

Recall that we introduced UTop+ as an extension to UTop to provide a more compact factorization and to jointly handle relationships across multiple implicit footprints. In Figure 5.4 we find UTop+ surpasses UTop in all settings. UTop+ has an improvement of 4.2% in precision@10, 3% in MAE@5, 5.9% in Kendall correlation@10, and 3.8% in Spearman correlation@5. These findings indicate that the proposed UTop+ can better exploit the joint correlations between all heterogeneous footprints for improved learning of user topical profiles. All these finding are conducted under McNemar's test with p-values less than 0.01.



(a) Precision

(b) MAE

(c) Kendall's τ

(d) Spearman's ρ

Figure 5.4: Comparisons between proposed models and alternative baselines

5.5.4 Considering Other Variants

Why We Need Regularization? A natural question is why we need a regularization model. Why not just put all footprints into one large matrix and directly apply state-of-the-art matrix factorization methods? To investigate this question, we put them into one matrix upon which we adopt the standard factorization technique, where we denote such a method MF. We do normalization for the data of each footprint since their values can have distinct scales. We follow the same evaluation methodology and show the comparisons in Figure 5.5. All results are measured at the top-10. We clearly see the proposed UTop results in better performances than MF in every metric. These results suggest that heterogeneous footprints require careful integration, and that the proposed UTop approach is a good solution in comparison.

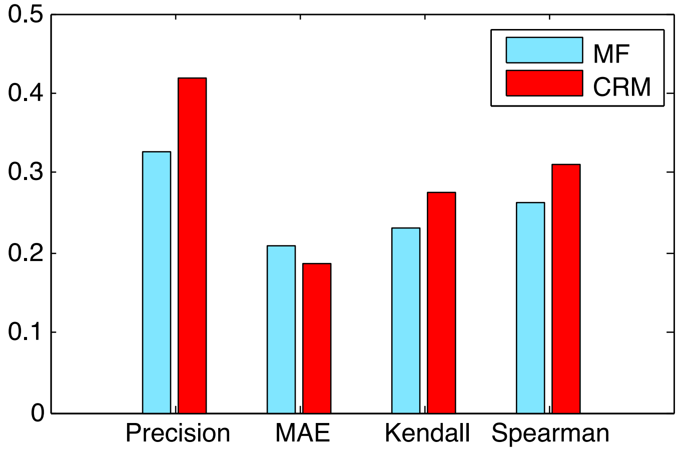


Figure 5.5: Comparisons between UTop and standard MF

Why We Do Regularization After Tensor Factorization? In UTop+, after having the latent factors of users from tensor factorization, we build a user similar matrix and find its graph Laplacian as the new regularization term. Why not just directly replace the user's latent matrix $\boldsymbol{U}$ in $\boldsymbol{X}$ after factorizing the tensor? We call such a scheme Tensor Factorization-based Matrix Factorization (TFMF), and we show the comparison results in Figure 5.6 for all metrics at top 10. Our UTop+ outperforms TFMF in all settings (e.g., 68% precison, 38% MAE, 45% Kendall correlation). These outcomes show that regularization after tensor factorization can significantly improve the performance.

Impact of Parameters. Finally, two critical parameters in UTop+ are α and β. Recall that α is used to avoid overfitting; β is to control the contribution of the user similarity derived from three types of footprints. In order to better understand the impacts of these two parameters, we evaluate the performance of UTop+ across various parameter settings. We vary values of these parameters in [0.001 0.01, 0.1, 1, 10] and present the results of

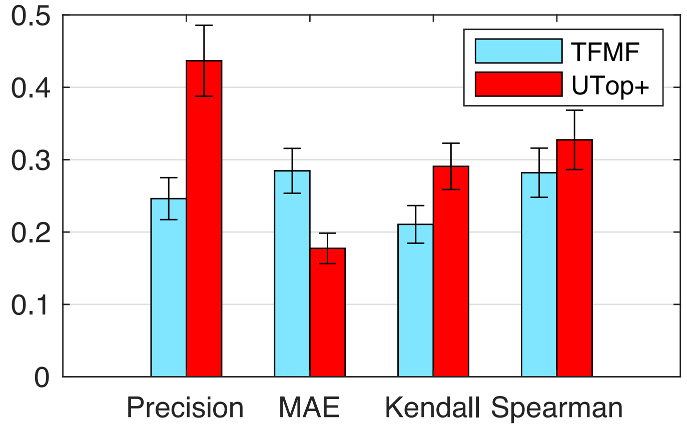
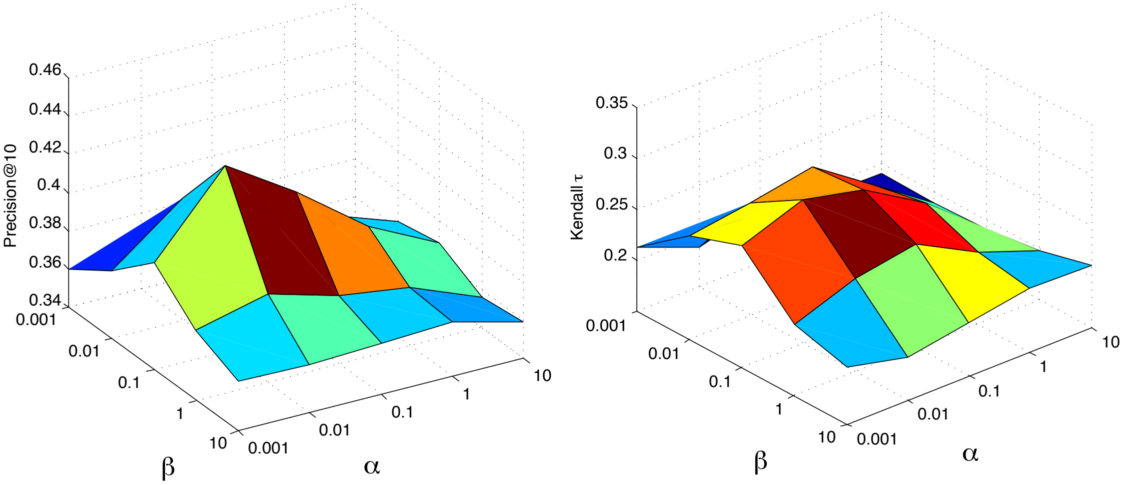


Figure 5.6: Comparisons between UTop+ and TFMF

precision and Kendall's τ in Figure 5.7 for learning the top-10 tags. As we can see, UTop+ achieves relatively consistent performance across a wide range. Particularly, we find the setting $\alpha = 0.1$ and $\beta = 0.01$ gives the best performance. These results indicate the stability of UTop+ to these parameters.

5.6 Summary

Mining user topical profiles (e.g., user interests and expertise) has important applications in diverse domains such as personalized search and recommendation, as well as expert detection. We tackled the problem of learning user topical profiles. In particular, we investigated how to leverage user-generated information in heterogeneous and diverse footprints. Concretely, we proposed UTop+ — a generalized model that integrates multiple implicit footprints with explicit footprints for learning high-quality user topical profiles. By taking into account pairwise relations among multiple footprints, the proposed UTop+ intelligently combines the potential benefits of each footprint to find the best evidence across footprints for learning high-quality user profiles. And indeed, extensive experiments demonstrate the effectiveness of UTop+. For instance, it surpasses other alternatives up to 36% in precision@5 and 20% in MAE@10. Link sharing, as one type



(a) Precision@10

(b) Kendall's τ

Figure 5.7: Impact of α and β on UTop+

of publicly-accessible user behavior, brings better results than other implicit footprints in every evaluation setting. Moreover, compared with other variants in terms of modeling, our model also has the best performance, e.g., up to 68% for precision@10 and Kendall correlation@10.

6. SUMMARY AND FUTURE WORK

This chapter concludes this dissertation with a summary of the contributions and highlighting some future directions for continued research.

6.1 Summary of This Dissertation

Social media facilitates user's self-expression and information spread like never before. Massive user behavior information are generated everyday on social media through varying user activities there. Unfortunately, many social media platforms have become breeding grounds for user misbehavior. In this dissertation we focus on three specific threads of user misbehaviors that commonly exist on social media — spamming, manipulation, and distortion. Toward tackling these misbehaviors we look into one concrete application for each misbehavior.

We first address the challenge of detecting spam links, which is an important task for shielding users from links associated with phishing, malware, and other low-quality, suspicious content. Rather than rely on traditional blacklist-based filters or content analysis of the landing page, we examine the behavioral factors of both who is posting the link and who is clicking on the link. The core intuition is that these behavioral signals may be more difficult to manipulate than traditional signals. Concretely, we propose and evaluate fifteen click and posting-based features. Through extensive experimental evaluation, we find that this purely behavioral approach can achieve high precision (0.86), recall (0.86), and area-under-the-curve (0.92), suggesting the potential for robust behavior-based spam link detection.

Next we uncover manipulated behavior of link sharing on social media. We investigate the individual-based and group-based user behavior pattern of link sharing toward organic versus organized user groups. Concretely, we propose a four-phase approach to model, identify, characterize, and classify organic and organized groups who engage in link sharing. The key motivating insights of this approach are (i) that patterns of individual-based behavioral signals embedded in link posting activities can uncover groups whose members engage in similar behaviors; and (ii) that group-level behavioral signals can distinguish between organic and organized user groups. Through extensive experiments, we find that levels of organized behavior vary by link type and that the proposed approach achieves good performance – an F-measure of 0.836 and Area Under the Curve of 0.921.

Finally we investigate a particular distortion behavior: posting BS on social media. Many case studies indicate that BS affects our government, news, health, advertising, as well as our social media. Because of the volume of content and the nuance of BS identification, building an auto-detector of BS on social media is extremely challenging and has not been extensively explored. We explore the factors impacting the perception of BS on social media and what leads users to ultimately perceive and call a post BS. We begin by preparing a reasonable, crowd-sourced collection of tweets that have been called BS. We then build a post level classification model that can determine what posts are more likely to be called BS. Our experiments suggest our classifier has the potential of leveraging linguistic cues for detecting social media posts that are likely to be called BS. We believe our work can serve as a stepping-stone to the ultimate goal of automated BS detection.

Each of these applications has its own scope in the context of user misbehavior, but user is the cross-cutting subject. Knowing the profiles of users can shed light into what subjects each user is associated with, which can benefit the understanding of the connection between user and misbehavior.

Therefore, we study a fourth task of learning social media user's topical profile — interests (i.e., what she likes) or expertise (i.e., what she is known for). It serves as a cross-cutting component toward better understanding those applications of user misbehavior.

User interests and expertise are valuable but often hidden resources on social media. To solve this problem, our main idea is to intelligently learn user topical profiles by exploiting information from multiple, heterogeneous footprints. We propose a unified model for learning user topical profiles that simultaneously considers multiple footprints. We show how these footprints can be embedded in a generalized optimization framework that takes into account pairwise relations among all footprints for robustly learning user profiles. Through extensive experiments, we find the proposed model is capable of learning high-quality user topical profiles, and leads to a 10-15% improvement in precision and mean average error versus a cross-triadic factorization state-of-the-art baseline.

6.2 Future Work

This dissertation can be extended along two thrusts in the future:

- **Thrust 1: Exploiting User Topical Profiles for Combating User Misbehavior**. In Figure 1.1 we have shown four problems studied in this dissertation — three threads of user misbehavior and a cross-cutting component related to user profile. In previous chapters we have presented our detailed solution for each of those four problems, mainly focusing on the patterns of misbehavior without emphasizing the user profile. Thus, one future direction is to introduce the cross-cutting component toward better combating those three user misbehaviors.

  In particular, we are interested in intelligently exploiting the learned user topical profiles for better solving those three applications. First, besides the behavioral signals (posting and clicking) explored in Section 2, we will incorporate user's topical profile into a model of detecting spam links. For example, those links posted by a user whose topical profiles associate with low-quality content (such as *ads* and *virus*) will become more suspicious in the new model. Second, for identifying manipulated link sharing behavior we will consider the user's group-level affinity of both sharing pattern and topical profile. For instance, for a group of users who share many same politics related links, we tend to have more evidence that these users are in a political campaign if we find the topical profiles of these users are all politics based. Thirdly, the learned user topical profile can also benefit BS identification. Recall that BS is a claim where the poster does not care if his claim is true or false, but rather uses the post as a story to some other effect. Hence, the learned topical profiles of a user can shed light into this user's knowledge domains, which can help judging whether this user makes BS statement in many circumstances.

- **Thrust 2: Automated BS Detection**. In Section 5 we build a system that automatically determines what social media posts are likely to be called BS. As discussed, this is our initial attempt toward building a BS auto-detector on social media. It can serve as a stepping-stone to the ultimate goal of automated BS detection.

  In concrete, we will continue our work through several avenues. First, we will temporally (considering a longer collection duration) and spatially (e.g., broadening to the whole world) extend our dataset. Second, we will incorporate more indicators to help us better characterize BS (when and where a post has been called as BS). Third, we will keep looking for existing related models from other disciplines (like philosophy and sociology) and explore whether they can be applied in our problem setting. Finally, we will pursue formally modeling of the motivation of BS-ing and BS calling, i.e. why some people are keen on saying or calling BS on social media.

REFERENCES

[1] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 49–62, ACM, 2009.

[2] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos, "Understanding user behavior in online social networks: A survey," *IEEE Communications Magazine*, vol. 51, no. 9, pp. 144–150, 2013.

[3] G. Seidman, "Self-presentation and belonging on facebook: How personality influences social media use and motivations," *Personality and Individual Differences*, vol. 54, no. 3, pp. 402–407, 2013.

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

[5] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging.," in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 13, pp. 2633–2639, 2013.

[6] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots+ machine learning," in *Proceedings of the 33rd international ACM International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 435–442, ACM, 2010.

[7] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web (WWW)*, pp. 61–70, ACM, 2012.

[8] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy,” in *Proceedings of the 22nd international conference on World Wide Web (WWW)*, pp. 729–736, ACM, 2013.

[9] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng, “Rumor cascades.,” in *International AAAI Conference on Web and Social Media (ICWSM)*, 2014.

[10] Z. Zhao, P. Resnick, and Q. Mei, “Enquiring minds: Early detection of rumors in social media from enquiry posts,” in *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pp. 1395–1405, International World Wide Web Conferences Steering Committee, 2015.

[11] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders, “Social networks and context-aware spam,” in *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work (CSCW)*, pp. 403–412, ACM, 2008.

[12] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, “Social phishing,” *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007.

[13] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting burstiness in reviews for review spammer detection.,” *International AAAI Conference on Web and Social Media (ICWSM)*, vol. 13, pp. 175–184, 2013.

[14] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, “Spotting opinion spammers using behavioral footprints,” in *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 632–640, ACM, 2013.

[15] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, “Detecting and tracking political abuse in social media.,” *International*

AAAI Conference on Web and Social Media (ICWSM), vol. 11, pp. 297–304, 2011.

[16] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[17] S. Pfattheicher and S. Schindler, "Misperceiving bullshit as profound is associated with favorable views of cruz, rubio, trump and conservatism," *PloS one*, vol. 11, no. 4, p. e0153419, 2016.

[18] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[19] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the web," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 381–396, ACM, 2011.

[20] A. Cui, M. Zhang, Y. Liu, and S. Ma, "Are the urls really popular in microblog messages?," in *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on*, pp. 1–5, IEEE, 2011.

[21] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis, "we. b: The web of short urls," in *Proceedings of the 20th international conference on World Wide Web*, pp. 715–724, ACM, 2011.

[22] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi. sh/$ ocial: the phishing landscape through short urls," in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pp. 92–101, ACM, 2011.

[23] F. Klien and M. Strohmaier, "Short links under attack: geographical analysis of spam in a url shortener network," in *Proceedings of the 23rd ACM conference on*

Hypertext and social media, pp. 83–88, ACM, 2012.

[24] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, “Knowing your enemy: understanding and detecting malicious web advertising,” in *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 674–686, ACM, 2012.

[25] K. Lee, P. Tamilarasan, and J. Caverlee, “Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media.,” in *International AAAI Conference on Web and Social Media (ICWSM)*, 2013.

[26] G. Stringhini, C. Kruegel, and G. Vigna, “Shady paths: Leveraging surfing crowds to detect malicious web pages,” in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 133–144, ACM, 2013.

[27] M. C. Forelle, P. N. Howard, A. Monroy-Hernandez, and S. Savage, “Political bots and the manipulation of public opinion in venezuela,” *SSRN Electronic Journal*, 2015.

[28] H. G. Frankfurt, *On bullshit*. Princeton University Press, 2009.

[29] H. G. Frankfurt, *On truth*. Random House, 2010.

[30] J. Bordeau, *Xenophobia: The violence of fear and hate*. The Rosen Publishing Group, 2009.

[31] C. Bergstrom and J. West, “Case study: 99.9% caffeine free,” 2017. `http://callingbullshit.org/case_studies/case_study_caffeine_free.html`.

[32] D. Sousa, L. Sarmento, and E. Mendes Rodrigues, “Characterization of the twitter replies network: are user ties social or topical?,” in *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC ’10, (New York, NY, USA), pp. 63–70, ACM, 2010.

[33] A. Sieg, B. Mobasher, and R. Burke, “Web search personalization with ontological user profiles,” in *ACM International Conference on Information and Knowledge Management (CIKM)*, 2007.

[34] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, “Social media recommendation based on people and tags,” in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[35] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, “Recommender systems with social regularization,” in *ACM WSDM Conference Series Web Search and Data Mining*, 2011.

[36] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, “Cognos: crowdsourcing search for topic experts in microblogs,” in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012.

[37] X. Zhang, J. Cheng, T. Yuan, B. Niu, and H. Lu, “Toprec: domain-specific recommendation through community topic mining in social network,” in *International World Wide Web Conference (WWW)*, 2013.

[38] M. McPherson, L. Smith-Lovin, and J. M. Cook, “Birds of a feather: Homophily in social networks,” *Annual Review of Sociology*, 2001.

[39] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao, “Happiness is assortative in online social networks,” *Artificial life*, 2011.

[40] D. Centola, “An experimental study of homophily in the adoption of health behavior,” *Science*, 2011.

[41] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?,” in *International World Wide Web Conference (WWW)*, 2010.

[42] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: a survey," *SIGKDD Conference on Knowledge Discovery and Data Mining Explorations Newsletter*, 2014.

[43] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *International World Wide Web Conference (WWW)*, 2010.

[44] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, ACM, 2007.

[45] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pp. 1–10, IEEE, 2010.

[46] D. Antoniades *et al.*, "we.b: the web of short urls," in *International World Wide Web Conference (WWW)*, 2011.

[47] A. Neumann, J. Barnickel, and U. Meyer, "Security and privacy implications of url shortening services," in *WEB 2.0 SECURITY and PRIVACY (W2SP)*, 2010.

[48] F. Maggi *et al.*, "Two years of short urls internet measurement: security threats and countermeasures," in *International World Wide Web Conference (WWW)*, 2013.

[49] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2010.

[50] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," in *IMC*, 2011.

[51] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *ACSAC*, 2010.

[52] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@spam: the underground on 140 characters or less,” in *CCS*, 2010.

[53] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, “Know your neighbors: web spam detection using the web topology,” in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2007.

[54] J. Song, S. Lee, and J. Kim, “Spam filtering in twitter using sender-receiver relationship,” in *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2011.

[55] C. Yang, R. C. Harkreader, and G. Gu, “Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers,” in *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2011.

[56] S. Lee and J. Kim, “WarningBird: Detecting suspicious URLs in Twitter stream,” in *Network and Distributed System Security Symposium (NDSS)*, 2012.

[57] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious urls,” in *Proceedings of the Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[58] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Identifying suspicious urls: an application of large-scale online learning,” in *International Conference on Machine Learning (ICML)*, 2009.

[59] D. K. McGrath and M. Gupta, “Behind phishing: an examination of phisher modi operandi,” in *LEET*, 2008.

[60] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: a fast filter for the large-scale detection of malicious web pages," in *International World Wide Web Conference (WWW)*, 2011.

[61] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drive-by-download attacks and malicious javascript code," in *International World Wide Web Conference (WWW)*, 2010.

[62] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in *Security and Privacy (SP), 2011 IEEE Symposium on*, pp. 447–462, IEEE, 2011.

[63] Y. Wang *et al.*, "Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities," in *Network and Distributed System Security Symposium (NDSS)*, 2006.

[64] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale automatic classification of phishing pages," in *Network and Distributed System Security Symposium (NDSS)*, 2010.

[65] G. Wang *et al.*, "You are how you click: Clickstream analysis for sybil detection," in *USENIX Security Symposium*, 2013.

[66] C. Wei *et al.*, "Fighting against web spam: A novel propagation method based on click-through data," in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2012.

[67] G. Wang *et al.*, "Serf and turf: crowdturfing for fun and profit," in *International World Wide Web Conference (WWW)*, 2012.

[68] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference*

(CEAS), vol. 6, p. 12, 2010.

[69] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in *Social computing (socialcom), 2010 ieee second international conference on*, pp. 177–184, IEEE, 2010.

[70] L. M. Smith, L. Zhu, K. Lerman, and Z. Kozareva, “The role of social media in the discussion of controversial topics,” in *Social Computing (SocialCom), 2013 International Conference on*, pp. 236–243, IEEE, 2013.

[71] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” in *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528, ACM, 2012.

[72] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, “Transient news crowds in social media.,” in *International AAAI Conference on Web and Social Media (ICWSM)*, 2013.

[73] J. Park, M. Cha, H. Kim, and J. Jeong, “Managing bad news in social media: A case study on domino’s pizza crisis.,” in *International AAAI Conference on Web and Social Media (ICWSM)*, 2012.

[74] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 1–9, ACM, 2010.

[75] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@ spam: the underground on 140 characters or less,” in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37, ACM, 2010.

[76] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, “Detecting and characterizing social spam campaigns,” in *Proceedings of the 10th ACM SIGCOMM*

conference on Internet measurement, pp. 35–47, ACM, 2010.

[77] N. Nikiforakis, F. Maggi, G. Stringhini, M. Z. Rafique, W. Joosen, C. Kruegel, F. Piessens, G. Vigna, and S. Zanero, "Stranger danger: exploring the ecosystem of ad-based url shortening services," in *Proceedings of the 23rd international conference on World wide web*, pp. 51–62, International World Wide Web Conferences Steering Committee, 2014.

[78] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks," in *ACM SIGCOMM Computer Communication Review*, vol. 36, pp. 267–278, ACM, 2006.

[79] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pp. 3–17, IEEE, 2008.

[80] G. Danezis and P. Mittal, "Sybilinfer: Detecting sybil nodes using social networks.," in *Network and Distributed System Security Symposium (NDSS)*, 2009.

[81] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 363–374, 2011.

[82] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[83] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[84] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[85] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 202–207, 1996.

[86] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in kernel methods*, pp. 185–208, MIT press, 1999.

[87] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[88] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[89] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning (ICML)*, (San Francisco), pp. 148–156, Morgan Kaufmann, 1996.

[90] M. Brito, E. Chavez, A. Quiroz, and J. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Statistics & Probability Letters*, vol. 35, no. 1, pp. 33–42, 1997.

[91] C. I. Staff, "Caffeine sensitivity." Available at `https://www.caffeineinformer.com/caffeine-sensitivity`.

[92] M. Caminada, "lies and bullshit; distinguishing classes of dishonesty," in *In: Social Simulation Workshop at the International Joint Conference on Artificial Intelligence (SS@ International Joint Conference on Artificial Intelligence (IJCAI)*, Citeseer, 2009.

[93] N. B. Macintosh, "Commentary accounting truth, lies, or âĂIJbullshitâĂİ? a philosophical investigation," *Accounting and the Public Interest*, vol. 6, no. 1, pp. 22–36, 2006.

[94] G. Pennycook, *Bullshit Detection and Cognitive Reflection.* 2016.

[95] I. Ajzen, "The theory of planned behavior," *Organizational behavior and human decision processes*, vol. 50, no. 2, pp. 179–211, 1991.

[96] "Theory of planned behavior," 2017. `https://en.wikipedia.org/wiki/Theory_of_planned_behavior` [Online; accessed 7-August-2017].

[97] I. Bogost, "Gamification is bullshit," *The gameful world: Approaches, issues, applications*, pp. 65–80, 2011.

[98] J. Fredal, "Rhetoric and bullshit," *College English*, vol. 73, no. 3, pp. 243–259, 2011.

[99] H. Yunis *et al.*, *Plato: Phaedrus.* Cambridge University Press, 2011.

[100] M. Caminada, "lies and bullshit; distinguishing classes of dishonesty," in *In: Social Simulation Workshop at the International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.

[101] G. Orwell, "Politics and the english language," *New York*, vol. 68, 1945.

[102] G. Pennycook, J. A. Cheyne, N. Barr, D. J. Koehler, and J. A. Fugelsang, "On the reception and detection of pseudo-profound bullshit," *Judgment and Decision Making*, vol. 10, no. 6, p. 549, 2015.

[103] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, ACM, 2011.

[104] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Anyone can become a troll: Causes of trolling behavior in online discussions,” *arXiv preprint arXiv:1702.01119*, 2017.

[105] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, “Tweetcred: Real-time credibility assessment of content on twitter,” in *International Conference on Social Informatics*, pp. 228–243, Springer, 2014.

[106] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, “Where the truth lies: Explaining the credibility of emerging claims on the web and social media,” in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1003–1012, International World Wide Web Conferences Steering Committee, 2017.

[107] A. Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm detection on twitter: A behavioral modeling approach,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 97–106, ACM, 2015.

[108] E. Ferrara, O. Varol, F. Menczer, and A. Flammini, “Detection of promoted social media campaigns.,” in *ICWSM*, pp. 563–566, 2016.

[109] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, “Towards understanding cyberbullying behavior in a semi-anonymous social network,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pp. 244–252, IEEE, 2014.

[110] M. Fishbein and I. Ajzen, “Belief, attitude, intention, and behavior: An introduction to theory and research,” 1977.

[111] S. Mukherjee and G. Weikum, “Leveraging joint interactions for credibility analysis in news communities,” in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 353–362, ACM, 2015.

[112] L. Karttunen, “Implicative verbs,” *Language*, pp. 340–358, 1971.

[113] J. Kimball, *Syntax and Semantics*. No. v. 3 in Syntax and Semantics, Academic Press, 1975. https://books.google.com/books?id=1bnlAAAAMAAJ.

[114] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, “Linguistic models for analyzing and detecting biased language.,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1650–1659, 2013.

[115] H. Rodina, “Metadiscourse: Exploring interaction in writing by hyland, ken,” *The Modern Language Journal*, vol. 91, no. 3, pp. 479–480, 2007.

[116] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *Journal of the Association for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[117] F. Å. Nielsen, “Afinn,” mar 2011. Available at http://www2.imm.dtu.dk/pubdb/p.php?6010.

[118] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354, Association for Computational Linguistics, 2005.

[119] Y. Choi and J. Wiebe, “+/-effectwordnet: Sense-level lexicon acquisition for opinion inference.,”

[120] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[121] E. Cambria and A. Hussain, "Sentic computing," *Cognitive Computation*, vol. 7, no. 2, pp. 183–185, 2015.

[122] H. Frankfurt, "On bullshit," *RARITAN-A QUARTERLY REVIEW*, vol. 6, no. 2, pp. 81–100, 1986.

[123] C. Martindale and A. Dailey, "Creativity, primary process cognition and personality," *Personality and individual differences*, vol. 20, no. 4, pp. 409–414, 1996.

[124] L. N. Frederik De Bleser, Tom De Smedt, "Nodebox," 2002. `http://nodebox.net`.

[125] K. Moffitt and M. B. Burns, "What does that mean? investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports," *AMCIS 2009 Proceedings*, p. 399, 2009.

[126] R. Flesch, "A new readability yardstick.," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[127] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring.," *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

[128] R. Senter and E. A. Smith, "Automated readability index," tech. rep., DTIC Document, 1967.

[129] "Linsear write," 2017. `https://en.wikipedia.org/wiki/Linsear_Write`.

[130] R. Gunning, *The technique of clear writing*. McGraw-Hill, New York, 1952.

[131] "Structure of u.s. education," 2017. `https://www.ed.gov/about/offices/list/ous/international/usnei/us/edlite-structure-us.html`.

[132] "List of controversial issues — Wikipedia, the free encyclopedia," 2017. https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues#Politics_and_economics.

[133] T. Mikolov, 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.

[134] N. DiFonzo and P. Bordia, "Rumor, gossip and urban legends," *Diogenes*, vol. 54, no. 1, pp. 19–35, 2007.

[135] "List of u.s. states by population density," Aug 2017. https://en.wikipedia.org/wiki/List_of_U.S._states_by_population_density.

[136] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *The computer journal*, vol. 7, no. 2, pp. 149–154, 1964.

[137] J. Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.

[138] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[139] N. L. Roux, M. Schmidt, and F. R. Bach, "A stochastic gradient method with an exponential convergence _rate for finite training sets," in *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

[140] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[141] Q. Liu, E. Chen, H. Xiong, C. H. Ding, and J. Chen, “Enhancing collaborative filtering by user interest expansion via personalized ranking,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 218–233, 2012.

[142] A. Majumder and N. Shrivastava, “Know your personalization: learning topic level personalization in online services,” in *International World Wide Web Conference (WWW)*, 2013.

[143] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” in *ACM WSDM Conference Series Web Search and Data Mining*, 2010.

[144] J. Hannon, M. Bennett, and B. Smyth, “Recommending twitter users to follow using content and collaborative filtering approaches,” in *Proceedings of the fourth ACM conference on Recommender systems*, pp. 199–206, ACM, 2010.

[145] L. Hong, A. S. Doumith, and B. D. Davison, “Co-factorization machines: modeling user interests and predicting individual decisions in twitter,” in *ACM WSDM Conference Series Web Search and Data Mining*, 2013.

[146] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, “Social contextual recommendation,” in *ACM International Conference on Information and Knowledge Management (CIKM)*, 2012.

[147] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang, “A temporal context-aware model for user behavior modeling in social media systems,” in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1543–1554, ACM, 2014.

[148] E. Zhong, N. Liu, Y. Shi, and S. Rajan, “Building discriminative user profiles for large-scale content recommendation,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

[149] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, “Exploiting social context for review quality prediction,” in *International World Wide Web Conference (WWW)*, 2010.

[150] P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly, and K. P. Gummadi, “Deep twitter diving: Exploring topical groups in microblogs at scale,” in *ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2014.

[151] V. Rakesh, D. Singh, B. Vinzamuri, and C. K. Reddy, “Personalized recommendation of twitter lists using content and network information,” in *International AAAI Conference on Web and Social Media (ICWSM)*, 2014.

[152] Z. Zhao, Z. Cheng, L. Hong, and E. H. Chi, “Improving user topic interest profiles by behavior factorization,” in *International World Wide Web Conference (WWW)*, 2015.

[153] M. Jamali and L. Lakshmanan, “Heteromf: recommendation in heterogeneous information networks using context dependent factor models,” in *International World Wide Web Conference (WWW)*, 2013.

[154] R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida, “Of pins and tweets: Investigating how users behave across image- and text-based social networks,” *International AAAI Conference on Web and Social Media (ICWSM)*, 2014.

[155] M. Qiu, F. Zhu, and J. Jiang, “It is not just what we say, but how we say them: Lda-based behavior-topic model,” in *Proceedings of the 2013 SIAM International*

Conference on Data Mining, pp. 794–802, SIAM, 2013.

[156] H. Ma, "On measuring social friend interest similarities in recommender systems," in *International ACM International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2014.

[157] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *ACM WSDM Conference Series Web Search and Data Mining*, 2010.

[158] T. Lappas, K. Punera, and T. Sarlos, "Mining tags using social endorsement networks," in *International ACM International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2011.

[159] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *International World Wide Web Conference (WWW)*, 2013.

[160] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *ACM WSDM Conference Series Web Search and Data Mining*, 2013.

[161] H. Lu, J. Caverlee, and W. Niu, "Discovering what you're known for: A contextual poisson factorization approach," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2016.

[162] Y. Hu, K. Talamadupula, S. Kambhampati, *et al.*, "Dude, srsly?: The surprisingly formal nature of twitter's language," in *International AAAI Conference on Web and Social Media (ICWSM)*, 2013.

[163] X. Liu and K. Aberer, "Soco: a social network aided context-aware recommender system," in *International World Wide Web Conference (WWW)*, 2013.

[164] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, “Tensor decompositions for learning latent variable models,” *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.

[165] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A multilinear singular value decomposition,” *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[166] M. Jiang, P. Cui, F. Wang, X. Xu, W. Zhu, and S. Yang, “Fema: Flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.

[167] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun, “Rubik: Knowledge guided tensor factorization and completion for health data analytics,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

[168] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban computing: concepts, methodologies, and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[169] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM review*, 2009.

[170] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han, “Personalized entity recommendation: A heterogeneous information network approach,” in *ACM WSDM Conference Series Web Search and Data Mining*, 2014.

[171] A. P. Singh and G. J. Gordon, “Relational learning via collective matrix factorization,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.

[172] L. Hu, J. Cao, G. Xu, L. Cao, Z. Gu, and C. Zhu, “Personalized recommendation via cross-domain triadic factorization,” in *International World Wide Web Conference (WWW)*, 2013.

[173] W. Feng and J. Wang, “Incorporating heterogeneous information for personalized tag recommendation in social tagging systems,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[174] I. Konstas, V. Stathopoulos, and J. M. Jose, “On social networks and collaborative recommendation,” in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2009.

[175] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme, “Learning optimal ranking with tensor factorization for tag recommendation,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[176] S. Rendle and L. Schmidt-Thieme, “Pairwise interaction tensor factorization for personalized tag recommendation,” in *ACM WSDM Conference Series Web Search and Data Mining*, 2010.

[177] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, “Tag recommendations based on tensor dimensionality reduction,” in *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 43–50, ACM, 2008.

[178] D. Yin, Z. Xue, L. Hong, and B. D. Davison, “A probabilistic model for personalized tag prediction,” in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[179] A. A. Goldstein, *Constructive real analysis*. Courier Corporation, 2013.

[180] E. Acar, D. M. Dunlavy, and T. G. Kolda, “A scalable optimization approach for fitting canonical tensor decompositions,” *Journal of Chemometrics*, vol. 25, pp. 67–

86, February 2011.

[181] C. Cao, J. Caverlee, K. Lee, H. Ge, and J. Chung, “Organic or organized?: Exploring url sharing behavior,” in *ACM International Conference on Information and Knowledge Management (CIKM)*, 2015.

[182] H. A. Kiers, J. M. Ten Berge, and R. Bro, “Parafac2-part i. a direct fitting algorithm for the parafac2 model,” *Journal of Chemometrics*, vol. 13, no. 3-4, pp. 275–294, 1999.