Anomalous Information Detection in Social Media

Rongrong Tao

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Narendran Ramakrishnan (Chair)
Chang-Tien Lu
Christopher North
Chandan Reddy
Feng Chen

Dec 11, 2020
Arlington, Virginia

Keywords: Social Media, Anomalous Information, Event Detection
Copyright 2020, Rongrong Tao
Anomalous Information Detection in Social Media

Rongrong Tao

(ABSTRACT)

This dissertation focuses on identifying various types of anomalous information pattern in social media and news outlets. We focus on three types of anomalous information, including (1) media censorship in news outlets, which is information that should be published but is actually missing, (2) fake news in social media, which is unreliable information shown to the public, and (3) media propaganda in news outlets, which is trustworthy information but being over-populated.

For the first problem, existing approaches on censorship detection mostly rely on monitoring posts in social media. However, media censorship in news outlets has not received nearly as much attention, mostly because it is difficult to systematically detect. The contributions of our work include: (1) a hypothesis testing framework to identify and evaluate censored clusters of keywords, (2) a near-linear-time algorithm to identify the highest scoring clusters as indicators of censorship, and (3) extensive experiments on six Latin American countries for performance evaluation.

For the second problem, existing approaches studying fake news in social media primarily focus on topic-level modeling or prediction based on a set of aggregated features from a collection of posts. However, the credibility of various information components within the same topic can be quite different. The contributions of our work in this space include: (1) a new benchmark dataset for fake news research, (2) a cluster-based approach to improve instance-level prediction of information credibility, and (3) extensive experiments for performance evaluations.

For the last problem, existing approaches to media propaganda detection primarily focus
on investigating the pattern of information shared over social media or evaluation from domain experts. However, these approaches cannot be generalized to a large-scale analysis of media propaganda in news outlets. The contributions of our work include: (1) non-parametric scan statistics to identify clusters of over-populated keywords, (2) a near-linear-time algorithm to identify the highest scoring clusters as indicators of propaganda, and (3) extensive experiments on two Latin American countries for performance evaluation.
Nowadays, massive information is available through a variety of social media platforms. However, the information accessed by the audience might be not exactly correct in different ways. In order for the audience being able to get access to the correct information, we develop various machine learning algorithms to uncover the anomalous information pattern in social media and explain the reason behind this behavior. Our algorithms can be used to learn what different information patterns can exist in the open data source.
Acknowledgements

I would like to express my gratitude for all of the help I have received through working on this PhD dissertation.

First of all, I want to thank my advisor, Dr. Naren Ramakrishnan, for his guidance in both my academic research and life. I still remember Dr. Ramakrishnan introducing me the field of data mining and machine learning when I joined the lab. He always encourages me to explore research problems of interest and provides valuable suggestions when I faced bottlenecks. His wisdom and passion in academic research not only provides me valuable support through my PhD work, but also motivates me to pursue a career in data mining and machine learning.

I also want to thank Dr. Chang-Tien Lu, Dr. Chris North, Dr. Chandan Reddy, and Dr. Feng Chen for serving as my committee members. They provide valuable suggestions through my work to help me broaden research directions, strengthen the completeness of research work and improve the presentation.

In addition, I want to thank my collaborators Dr. David Mares and Dr. Ryan Kennedy for helping to propose research directions and validate the correctness of experiment with their domain expertise.

Moreover, I want to thank the members and alumni of the Discovery Analytics Center: Sathappan Muthiah, Dr. Yue Ning, Dr. Rupinder Paul, Dr. Wei Wang, Dr. Huijuan Shao, Dr. Fang Jin, Ping Wang, Dr. Tian Shi, Dr. Parang Saraf, Juanita Victoria, Wanawsha Hawrami, and Joyce Newberry.

Furthermore, I want to thank my parents and parents-in-law for their support.

Finally, I want to thank my husband, Dr. Xu Shi, for his company and love.
# Contents

1 Introduction .................................................. 1

1.1 Research Issues ............................................. 3

1.1.1 Detecting Media Self-Censorship Using Social Media ........ 3

1.1.2 Predicting Information Credibility in Social Media .......... 4

1.1.3 Detecting Propaganda in News Media ...................... 5

1.2 Contributions ................................................. 5

1.3 Dissertation Organization .................................... 7

2 Related Work .................................................. 9

2.1 Detecting Media Self-Censorship Using Social Media .......... 11

2.2 Predicting Information Credibility in Social Media ............ 12

2.3 Detecting Propaganda in News Media ........................ 13

3 Detecting Media Self-Censorship Using Social Media .......... 15

3.1 Introduction .................................................. 15
3.2 Data and Processing ........................................... 17
   3.2.1 Data Preprocessing ....................................... 18
   3.2.2 Pattern Analysis .......................................... 19
3.3 Methodology .................................................. 22
   3.3.1 Problem Formulation ..................................... 22
   3.3.2 GraphDPD Algorithm .................................... 25
3.4 Experiments ................................................... 29
   3.4.1 Experimental Design ..................................... 30
   3.4.2 Semi-synthetic Data Evaluation ....................... 32
   3.4.3 Real Data Evaluation .................................... 38
   3.4.4 Case Studies ............................................. 39

4 Predicting Information Credibility in Social Media 43
   4.1 Introduction ................................................. 43
   4.2 Methodology ................................................ 45
      4.2.1 Dataset ................................................ 45
      4.2.2 Feature Extraction ................................... 47
      4.2.3 Cluster-based Prediction .............................. 51
   4.3 Experiments ................................................ 53
      4.3.1 Quantitatively Evaluation .............................. 53
5 Detecting Propaganda in News Media

5.1 Methodology ................................................................. 59
   5.1.1 Pattern Analysis ....................................................... 59
   5.1.2 Problem Formulation ............................................... 60
   5.1.3 GraphMP-FS Algorithm ............................................. 65

5.2 Experiments ................................................................. 68
   5.2.1 Data ................................................................. 68
   5.2.2 Evaluation .......................................................... 69

6 Conclusion and Future Work ............................................ 71

6.1 Conclusion ................................................................. 71

6.2 Future Work ............................................................... 72
   6.2.1 Media Censorship Forecasting ................................. 72
   6.2.2 Predicting Information Credibility in News Media ........ 73
   6.2.3 Study of Social Bots .............................................. 75
# List of Figures

1.1 Overview of research problems in this dissertation. .......................... 2

3.1 Example TSDF in *News* vs. TSDF in *Twitter* for a set of connected keywords. These keywords are relevant to the 43 missing students from Ayotzinapa Rural Teachers’ College on Sep 26th, 2014 in Mexico. We can find consistent censorship pattern in El Mexicano Gran Diario Regional (el-mexicano.com.mx) shortly after the students are missing. .............................................. 35

3.2 Anomaly detection results (mean precision (left), recall (center), and F-measure (right) vs. perturbation intensity) for different anomaly subgraph sizes (increased size from top to bottom) in semi-synthetic data. X-axis represents $q_t$, which implies the scale of anomaly injected in *Twitter*. $q_n$, which implies the scale of anomaly injected in *News*, is varied following $q_t \times q_n = 1.0$. .......................... 36

3.3 The density of the proportion of sources reporting events that are detected as censored by our model. The results suggest that we are capturing very few events that are only of interest to social media. ................................. 40

3.4 Word cloud representing censored keywords in *News* identified by our method. 42

4.1 ROC curve of our proposed approach with baseline methods SVM and MISVM. 56
4.2 Example tweets of different credibility level in the same news topic.
List of Tables

3.1 Summary of Twitter Transparency Report for Year 2014 on selected countries. 18

3.2 Different patterns of co-occurrence observed between social media and news
media sources. ................................................................. 20

3.3 Description of major notation. .................................................. 22

3.4 Real-world dataset used in this work. Tweets: average number of daily tweets.
News Articles: average number of daily local news articles. News Outlets:
number of news outlets used in this study. ........................................ 30

3.5 Example indicators of censorship identified by our approach in Mexico and
Venezuela during Year 2014 (with significance level $\leq 0.05$) .............. 37

3.6 Comparison of false positive rates in censorship detection between GRAPHDPD
and three baseline methods: LTSS, NPHGS, and EventTree on real data of
Mexico and Venezuela during year 2014. ........................................ 38

4.1 Description of the dataset we used in this work. ............................. 47

4.2 Features extracted at topic level and grouped into three categories: Message,
User, and Propagation. .......................................................... 49
4.3 Features extracted at instance level and grouped into two categories: Message and User. ................................................................. 50

4.4 Performance of our proposed approach with baseline methods on a 5-fold cross validation. ......................................................... 54

4.5 Performance of our proposed approach with baseline methods on instance-level credibility prediction. .............................................. 55

5.1 Comparison of false positive rates in propaganda detection between GraphMP-FS and three baseline methods: LTSS, NPHGS, and EventTree on real data of Mexico and Venezuela during year 2014. .............................. 70
Chapter 1

Introduction

Online social networks (OSNs) continue to play an important role in information sharing, including Twitter, Facebook, Instagram, online news, and etc. Hence, OSNs have become a popular source for many data mining, machine learning, and information retrieval problems. This dissertation’s motivation arises from one of the classical research problems in this domain, viz. event modeling, which includes event analysis, event detection, event forecasting [50, 75, 74, 73, 32, 2, 76, 49], and etc. (see Figure 1.1).

Classical event detection research has focused on a range of mining and machine learning algorithms. Some are supervised algorithms, e.g., regression models [55, 56], classification models [42], and topic models including Latent Dirichlet Allocation (LDA) [15, 59, 22]. Other classification models proposed for event detection problems include support vector machine (SVM) [62, 5], and Naive Bayes [63]. Unsupervised methods include clustering models [70, 40, 1, 77, 54, 16], and extraction models including PCA, SVD [46]. Example clustering models discussed for event detection problems include K-means [39], hierarchical clustering [26, 43], agglomerative clustering [30], density-based clustering (DBSCAN) [12], and etc. There are also other techniques explored for this research problem, for example,
Figure 1.1: Overview of research problems in this dissertation.

As event modeling becomes mainstream we face a rapid explosion in the amount of data generated by individuals. This leads to a range of anomalous information patterns that should be modeled and analyzed. Furthermore, the quality, volume, and trustworthiness of the information is also called into question. Quality of information can be disrupted by either manipulating the volume of information or by manipulating the content of the information. Issues associated with volume of information can entail: (1) information missing from the internet as it is hidden by the publisher for specific purpose, and (2) information overemphasized on behalf of specific viewpoints. Issues associated with the content of information involve information becoming misleading due to either lack of relevant domain knowledge or because it has been repurposed to suit specific purposes.
As a result, it is not easy for the general public or even domain experts to delineate these aspects of online information. Although the social aspects of news media censorship have been deeply discussed and analyzed in the field of social sciences, there is currently no efficient and effective approach to automatically detect and track self-censorship in news media in real time. Social media censorship often takes the form of identifying posts and deleting them and therefore tracking post deletions supports the use of supervised learning approaches. Censorship in news media typically has no labeled information and must rely on unsupervised techniques instead. In addition, analysis of information credibility mostly focus on identifying the non-trustworthy news topics [21, 57, 78, 3, 28, 37, 27, 8] and credibility checking of single piece of information has received relatively less attention. However, credibility of pieces of information can be quite different depending on the opinion and sentiment even if they are discussing about the same news topic.

1.1 Research Issues

This research aims to investigate and develop robust models for detection of missing information, misinformation, and over-populated information in large-scale datasets. The major research issues are stated as follows:

1.1.1 Detecting Media Self-Censorship Using Social Media

The motives and means of explicit state censorship have been well studied, both quantitatively and qualitatively. Self-censorship by media outlets, however, has not received nearly as much attention, mostly because it is difficult to systematically detect. We develop a novel
approach, using social media to identify likely instances of self-censorship. To achieve this, we develop a hypothesis testing framework to identify and evaluate censored clusters of keywords and a near-linear-time algorithm (called GraphDPD) to identify the highest scoring clusters as indicators of censorship. We evaluate the accuracy of our framework, in rivalry with several other state-of-the-art algorithms, using both semi-synthetic and real-world data from six Latin American countries from 2013-2017. These tests demonstrate the capacity of our framework to identify self-censorship, and provide an indicator of broader media freedom. The results of this study lay the foundation for detection, study, and policy-response to self-censorship.

1.1.2 Predicting Information Credibility in Social Media

The analysis of non-trustworthy information has been well studied, both in traditional news media and fast-growing social media. Analysis of information credibility at the lower level, however, has not received nearly as much attention, mostly because the lower-level information has less features to extract and is more subject to noise. We develop a cluster-based approach, grouping instances with potentially similar credibility level into clusters to improve accuracy of credibility prediction. To achieve this, we develop a framework for instance clustering based on propagation network and linguistic characteristics. The accuracy of our proposed approach is compared with several baseline methods, using real-world Twitter data on a variety of news topics. These tests demonstrate the capacity of our framework to effectively predict the information credibility at the instance level in social media.
1.1.3 Detecting Propaganda in News Media

Media propaganda, which exists in a manner of amplified or over-populated information on purpose in news media, has been well studied by domain experts. However, investigating the pattern of information being shared or evaluation from experts or profiles are limited to a range of specific event topics and not easy to generalize to a large-scale analysis. We develop a novel approach based on non-parametric scan statistics to identify highest scoring clusters as indicators of propaganda pattern. We evaluate the performance of our proposed approach, in comparison to several state-of-the-art algorithms, using real-world data from Mexico and Venezuela in 2014. The evaluation demonstrates that our proposed approach is capable to efficiently and effectively detect propaganda in news media.

1.2 Contributions

The major proposed research contributions can be stated as follows:

Detecting Media Self-Censorship Using Social Media

- Analysis of censorship patterns between news media and Twitter: We carried out an extensive analysis of information in Twitter deemed relevant to censored information in news media. In doing so, we make important observations that highlight the importance of our work.

- Formulation of an unsupervised censorship detection framework: We propose a novel hypothesis-testing-based statistical framework for detecting clusters of co-occurred keywords that demonstrate statistically significant differences between the information published in news media and the correlated information published in a
uncensored source (e.g., Twitter). To the best of our knowledge, this is the first unsupervised framework for automatic detection of censorship events in news media.

- **Optimization algorithms**: The inference of our proposed framework involves the maximization of a GLRT statistic function over all clusters of co-occurred keywords, which is hard to solve in general. We propose a novel approximation algorithm to solve this problem in nearly linear time.

- **Extensive experiments to validate the proposed techniques**: We conduct comprehensive experiments on real-world Twitter and local news articles datasets to evaluate our proposed approach. The results demonstrate that our proposed approach outperforms existing techniques in the accuracy of censorship detection. In addition, we perform case studies on the censorship patterns detected by our proposed approach and analyze the reasons behind censorship from real-world data of six Latin American countries from 2013 to 2017.

Predicting Information Credibility in Social Media

- **New Benchmark Dataset for fake news research**: Existing research on fake news have made a few datasets publicly available, however, there are no agreed upon benchmark datasets. The reason behind this issue is mainly because of the topics included in the datasets or the way that ground-truth labels are collected. We build a new benchmark dataset on a wide range of topics with ground truth labels collected from fact-checking website Snopes.

- **Formulation of a cluster-based credibility prediction framework**: We propose a cluster-based framework to predict information credibility at the instance level by grouping instances with potentially similar credibility level into clusters. The instances are clustered based on their linguistic and propagation network.
• **Extensive experiments to validate the proposed techniques**: We conduct comprehensive experiments on real-world Twitter data to evaluate our proposed approach. The results demonstrate that our proposed approach outperforms baseline methods in the accuracy of predicting information credibility at the instance level.

Detecting Propaganda in News Media

• **Formulation of an unsupervised propaganda detection framework**: We propose a novel non-parametric scan statistics based framework to detect clusters of co-occurred keywords and sets of news outlets that demonstrate statistically significant differences among the information published in news media. The problem of propaganda detection is reformulated as composed of two major components: highest scoring clusters detection and statistical significance analysis.

• **Optimization algorithms**: We propose a novel approximation algorithm to maximize the scoring function over clusters of keywords and sets of news outlets at the same time subject to a connectivity constraint in graphs in nearly linear time.

• **Extensive experiments for performance evaluation**: We conduct comprehensive experiments on real-world Twitter and local news articles datasets of Mexico and Venezuela in 2014 to evaluate our proposed approach. The results demonstrate that our proposed approach outperforms state-of-the-art algorithms in propaganda detection.

### 1.3 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 discusses literature work relevant to this research area and research problems, including event detection and event
forecasting. Chapter 3 presents (1) a hypothesis testing framework to identify and evaluate censored clusters of keywords, and (2) a near-linear-time algorithm to identify the highest scoring clusters as indicators of censorship. The accuracy of our framework is evaluated using both semi-synthetic and real-world data from six Latin American countries against several other state-of-art algorithm. Chapter 4 presents (1) a new benchmark dataset for fake news research, and (2) a cluster-based approach to improve prediction of information credibility at the instance level in social media. Chapter 5 presents (1) a non-parametric scan statistics framework to identify indicators of propaganda in news media, and (2) a near-linear-time algorithm to maximize scoring function over clusters of keywords and sets of news outlets at the same time subject to a connectivity constraint in graphs. Chapter 6 concludes this dissertation and discusses future work for this dissertation.
Chapter 2

Related Work

This research focuses on the development of robust models to identify missing information in news media, misleading information of social media at the instance level, and over-populated information in news media. These tasks are relevant to classic research on event detection, anomaly detection and misinformation detection. A summary of relevant research are summarized as follows.

- **Analysis of the coverage of various topics across social media and news media** has been well established in many studies. [53] studies topic and timing overlapping in newswire and Twitter and concludes that Twitter covers not only topics reported by news media during the same time period, but also minor topics ignored by news media. Through analysis of hundreds of news events, [51] observes both similarities and differences of coverage of events between social media and news media. In this dissertation, we uncover indicators of censorship pattern in news media from various interactive patterns between social media and news media. The role of social media in news reporting is analyzed in [67].
• **Event detection** in social media has been studied in many recent works. Watanabe et al. [69] develop a system, which identifies tweets posted closely in time and location and determine whether they are mentions of the same event by co-occurring keywords. Ritter et al. [59] presents the first open-domain system for event extraction and an approach to classify extracted events based on latent variable models. Rozenshtein et al. [60] formulates event detection in activity networks as a graph mining problem and proposes effective greedy approaches to solve this problem. In addition to textual information, Gao et al. [25] propose an event detection method which utilizes visual content and intrinsic correlation in social media.

• **Censorship Detection** is a critical problem in many countries across the world and most of the existing studies on censorship analysis are focused on Twitter. It has been studied for censorship topics by applying topic extraction and clustering on a collection of censored tweets. However, most of the existing approaches are supervised or semi-supervised, which rely on collections of censored posts, and highlight the necessity of unsupervised approaches to uncover self censorship in news media.

• **Misinformation Detection** has been well studied in research. Depending on the data source of interest, we can categorize existing studies into two major types. On one hand, [9] and [61] aim to identify misinformation in online news media with set of features extracted from headlines and news contents. On the other hand, identification of misinformation in social media has received more attention in this field. Example social media platform of interest include Facebook and Twitter. Although features that can be extracted from these platforms are slightly different, we can categorize the features into three major types: user features about authors of tweets, topic features about contents of tweets, and propagation features about interaction of tweets. Many studies view misinformation detection as a binary classification problem and propose
set of features [21, 57, 78, 3, 28, 37, 27, 8]. Furthermore, many studies also dug into
the propagation pattern [72, 24, 71] and temporal diffusion pattern [44, 45, 38, 36] of
rumor cascades in the social media.

2.1 Detecting Media Self-Censorship Using Social Media

Previous studies have focused on explicit censorship of posts on a variety of social networking
sites, for example, Twitter, Facebook, and Instagram. The authors in [19] studied patterns
of censorship by collecting English posts from 3.9 million Facebook users over 17 days.
They proposed a list of behavioral, demographic, and social graph features of users and
constructed a data-driven model of censorship. The study in [66] performed a user study
to explore censorship behavior. The authors discussed the types of missing content and the
reasons for censorship.

These methods are not easily adapted for the study of self-censorship, since they require that
the story or post be published (or submitted) and removed, allowing for direct observation of
explicit censorship. To detect self-censorship using social media, we need to be able to detect
major events in social media apriori, i.e. events the media would have reported with a high
likelihood if not for self-imposed restrictions. The detection of such events has largely been
done in the field of event detection. [69], for example, developed a system which identifies
tweets posted closely in time and location, and determined whether they are mentions of
the same event by co-occurring keywords. [59] presented the first open-domain system for
event extraction and an approach to classify events based on latent variable models. [60]
formulated event detection in activity networks as a graph mining problem and proposed
effective greedy approaches to solve this problem. In addition to textual information, [25] proposed an event detection method which utilizes visual content and intrinsic correlation in social media.

We must be careful not to overstate the utility of social media for detecting major events. Analysis of the coverage of various topics across social media and news media have found many similarities, but also some systematic differences. [53] studied topic and timing overlapping in newswire and Twitter and concluded that Twitter covers not only topics reported by news media during the same time period, but also minor topics ignored by news media. Through analysis of hundreds of news events, [51] observed both similarities and differences of coverage of events between social media and news media.

### 2.2 Predicting Information Credibility in Social Media

Existing studies of information credibility analysis can be summarized into two major categories based on different interests of data sources. On one hand, [9] and [61] aim to detect misinformation in online news media, where [9] utilizes a variety of features in headlines and [61] makes use of linguistic features in news context.

On the other hand, propagation of misinformation/rumor in social news media, including Facebook, Twitter, and Sina Weibo, has received more attention in relevant studies and most of these studies view misinformation detection as a classification problem. According to different types of features proposed, existing studies can be summarized into three major categories. First of all, [21, 57, 78, 3, 28, 37, 27, 8] proposes to extract user features and topic features. User features mainly focus on the authors of tweets, including status of their friends, followers, and etc. Topic features mainly focus on the contents of tweets, including information of retweets, embedded urls, sentiment, user mentions, hashtags, and etc.
Specifically, [34] proposes a novel method, which utilizes conflicting viewpoints in tweets to evaluate information credibility.

In addition, [72, 24, 71] studies how rumor cascades in the social network and proposes to model the pattern of message propagation as a tree, from where the propagation of false rumors can be distinguished from other messages. Furthermore, [44, 45, 38, 36] studies the temporal diffusion pattern of rumors and proposed to distinguish misinformation based on the following observation: time series of rumors tend to have multiple and periodic spikes, whereas non-rumors typically have a single prominent spike.

Instead of the most widely used classification methods, a few different approaches has been proposed. [33] proposes a three-layer hierarchical credibility network from message to sub-event, and sub-event to event and formulate credibility propagation as a graph optimization problem. [13] views rumor detection as an anomaly detection problem and perform FAMD on proposed features to detect the anomalies.

### 2.3 Detecting Propaganda in News Media

Existing research on propaganda study can be summarized into two major categories depending on the analysis strategy. On one hand, many studies have been investigating the pattern of news and information being shared over social media including Twitter and Facebook. Existing studies discussed the differential media sharing patterns on the social media platforms and analyzed the formation and function of media structures. [23] investigated into the spreading of a viral meme via (1) cluster analysis of network using a label propagation method, and (2) use kNN to predict labels based on the similarity calculated by Dynamic time warping (DTW) between time series of features. On the other hand, evaluations from domain expert include: (1) country profile for each country drafted by a team of research
assistants who consulted country-specific experts on the accuracy of reliability of the publicly available information, and (2) consulting with country experts to check facts, find additional sources in multiple languages and assist in evaluating the quality of sources.
Chapter 3

Detecting Media Self-Censorship Using Social Media

3.1 Introduction

News media censorship is generally defined as a restriction on freedom of speech to prohibit access to public information, and is taking place more than ever before. The Freedom of the Press Report categorizes the level into ”free”, ”partly free”, and ”not free”. According to the Freedom of the Press Report, there are a few nations fit into the ”not free” category in 2014.

One of the responses to this environmental context is self-censorship, i.e., the act of deciding not to publish about certain topics. However, there is currently no efficient and effective approach to automatically detect and track self-censorship events in real time. We can draw some parallels to social media censorship. Here, censorship often takes the form of active

\[\text{\url{https://freedomhouse.org/event/new-challenges-freedom-expression-latin-america}}\]
censors identifying offending posts and deleting them and therefore tracking post deletions supports the use of supervised learning approaches. On the other hand, censorship in news media typically has no labeled information and must rely on unsupervised techniques instead.

In this chapter, we present a novel unsupervised approach that views social media as a sensor to detect censorship in news media wherein statistically significant differences between information published in the news media and the correlated information published in social media are automatically identified as candidate censored events. A generalized log-likelihood ratio test (GLRT) statistic is formulated for hypothesis testing, and the problem of censorship detection is cast as the maximization of the GLRT statistic over all possible clusters of keywords. We propose a near-linear-time algorithm called GraphDPD to identify the highest scoring clusters as indicators of censorship events in the local news media, and further apply randomization testing to estimate the statistical significance of these clusters. We consider the detection of censorship in the news media of Mexico and Venezuela, and utilize Twitter as the uncensored source.

Starting in January 2012, a “Country-Withheld Content” policy has been launched by Twitter, with which governments are able to request withholding and deletion of user accounts and tweets. At the same time, Twitter started to release a transparency report, which provided worldwide information about such removal requests. The Transparency Report lists information and removal requests from Year 2012 on a half-year basis. Table 3.1 summarizes the information and removal requests for Year 2014 on our selected countries. Turkey is the country issuing the largest number of censorship requests to Twitter (see Table 3.1). For Mexico and Venezuela, Twitter did not participate in any social media censorship. Based on this observation, Twitter can be considered as a reliable and uncensored source to detect news self censorship events in Latin America. The main contributions of this chapter are summarized as follows:
• **Analysis of censorship patterns between news media and Twitter**: We carried out an extensive analysis of information in Twitter deemed relevant to censored information in news media. In doing so, we make important observations that highlight the importance of our work.

• **Formulation of an unsupervised censorship detection framework**: We propose a novel hypothesis-testing-based statistical framework for detecting clusters of co-occurred keywords that demonstrate statistically significant differences between the information published in news media and the correlated information published in an uncensored source (e.g., Twitter). To the best of our knowledge, this is the first unsupervised framework for automatic detection of censorship events in news media.

• **Optimization algorithms**: The inference of our proposed framework involves the maximization of a GLRT statistic function over all clusters of co-occurred keywords, which is hard to solve in general. We propose a novel approximation algorithm to solve this problem in nearly linear time.

• **Extensive experiments to validate the proposed techniques**: We conduct comprehensive experiments on real-world Twitter and News datasets. The results demonstrate that our proposed approach outperforms existing techniques in the accuracy of censorship detection. In addition, we perform case studies on the detected censorship patterns and analyze the reasons behind censorship.

### 3.2 Data and Processing

Table 3.3 summarizes the notation used in this work. The EMBERS project [58, 31] provided a collection of Latin American news articles and Twitter posts. The news dataset was sourced
Table 3.1: Summary of Twitter Transparency Report for Year 2014 on selected countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Requests (Court Order)</th>
<th>Requests (Govt, Police, etc.)</th>
<th>Accounts Withheld</th>
<th>Tweets Withheld</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Australia</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brazil</td>
<td>35</td>
<td>0</td>
<td>5</td>
<td>101</td>
</tr>
<tr>
<td>Colombia</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Greece</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Japan</td>
<td>6</td>
<td>21</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>Mexico</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Turkey</td>
<td>393</td>
<td>270</td>
<td>79</td>
<td>2,003</td>
</tr>
<tr>
<td>Venezuela</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

from around 6000 news agencies during 2014 across the world. From “4 International Media & Newspapers”, we retrieved a list of top newspapers with their domain names in the target country. News articles are filtered based on the domain names in the URL links. Twitter data was collected by randomly sampling 10% (by volume) tweets during Year 2014. Mexico and Venezuela were chosen as target countries in this work since they had no censorship in Twitter (as shown in Table 3.1) but severe level of censorship in news media according to the Freedom of the Press Report.

3.2.1 Data Preprocessing

The inputs to our proposed approach are keyword co-occurrence graphs. Each node represents a keyword associated with four attributes: (1) time-series daily frequency (TSDF) in Twitter; (2) TSDF in News, (3) expected daily frequency in Twitter; and (4) expected daily frequency in News. Each edge represents the co-occurrence of connecting nodes in Twitter, or News, or both. However, constructing such graphs is not trivial due to data integration.
One challenge is to handle the different vocabularies used in *Twitter* and *News*, with underlying distinct distributions.

To find words that behave differently in *News* comparing to *Twitter*, we only retained keywords which are mentioned in both *Twitter* and *News*. For each keyword, linear correlation between its TSDF in *Twitter* and *News* during Year 2014 is required to be greater than a predefined threshold (e.g. 0.15) in order to guarantee the keyword is well correlated in two data sources. TSDF in *Twitter* and *News* for each node are normalized with quantile normalization. An edge is removed if its weight is less than Γ, where Γ is the threshold used to tradeoff graph sparsity and connectivity. Empirically we found Γ = 10 to be an effective threshold. A keyword co-occurrence graph for a continuous time window is defined as the maximal connected component from a union of daily keyword co-occurrence graph during the time window.

### 3.2.2 Pattern Analysis

It’s challenging to claim that any deviation between social media and news media is evidence of censorship or different topics of interest. Table 3.2 summarizes various co-occurring patterns between Twitter and news media that we are able to observe from our real world dataset and more details are discussed as follows.

**Topic is of interest both in social media and news media:** In early March 2014, Malaysia Airlines Flight MH370 disappeared while flying. We are able to observe sparks in mentions of relevant keywords across both social media and news media.

**Topic is of interest only in social media:** In late June 2014, there is one soccer game between Mexico and Holland during the 2014 FIFA World Cup. We are able to observe spikes in mentions of relevant keywords across Twitter in Mexico while Mexican news outlets do
Table 3.2: Different patterns of co-occurrence observed between social media and news media sources.

<table>
<thead>
<tr>
<th>Topic is of interest in both social media and news media.</th>
<th>Topic is of interest in social media but not in news media.</th>
<th>Topic is of interest in news media but not in social media.</th>
<th>Censorship in one news media source.</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

Example: In early March 2014, Malaysia Airlines Flight MH 370 went missing.

Example: Late June 2014 featured a soccer game between Mexico and Holland as part of the 2014 FIFA World Cup.

Example: In late September 2014, 125 heads of state and governments attended the Global Climate Summit, which was seen as a milestone to a new legal agreement on climate change.

Example: In late September 2014, 43 students from Ayotzinapa Rural Teachers’ College went missing in Mexico.

not depict significant changes.

**Topic is of interest only in news media:** In late September 2014, heads of state and governments attended the global Climate Summit. This incident is widely discussed in news media, while relatively less attention in social media.

**Topic is censored in news media:** Fig. 3.1 compares TSDF in El Mexicano Gran Diario Regional (el-mexicano.com.mx) and TSDF in Twitter during a 2-month period on a connected set of keywords. All the example keywords are relevant to the 43 missing students from Ayotzinapa in the city of Iguala protesting the government’s education reforms. The strong connectivity of these keywords, as shown in Fig. 3.1e, guarantees that they are mentioned together frequently in Twitter and local news media. The time region during
which anomalous behavior is detected is highlighted with two yellow markers. Since volume of Twitter is much larger than volume of News, TSDF in Fig. 3.1a to Fig. 3.1d are normalized to $[0, 500]$ for visualization. Fig. 3.1a to Fig. 3.1d depict that TSDF in El Mexicano Gran Diario Regional is well correlated with TSDF in Twitter except during the highlighted time region, where abnormal absenteeism in El Mexicano Gran Diario Regional can be observed for all example keywords. In order to validate the deviation between them is not due to difference in topics of interests, we also compare with a number of other local news outlets. Fig. 3.1a to Fig. 3.1d shows that TSDF in El Universal in Mexico City is consistent with TSDF in Twitter and does not depict an abnormal absenteeism during the highlighted time period. Using Twitter and El Universal in Mexico City as sensors, we can conclude an indicator of self-censorship in El Mexicano Gran Diario Regional with respect to the 43 missing students during the highlighted time region.

Inspired by these observations, we say that a censorship pattern exists if for a cluster of connected keywords,

1. Their TSDF in at least one local news media is consistently different from TSDF in Twitter during a time period,

2. Their TSDF in local news media are consistently well correlated to TSDF in Twitter before the time period, and

3. Their TSDF in at least one different local news outlet does not depict abnormal absenteeism during the time period.
Table 3.3: Description of major notation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {a^t(v)}_{t=1}^T )</td>
<td>time series of daily frequency of node ( v ) in uncensored Twitter dataset</td>
</tr>
<tr>
<td>( \lambda_a(v) )</td>
<td>expected daily frequency of node ( v ) in the Twitter dataset.</td>
</tr>
<tr>
<td>( {b^t(v)}_{t=1}^T )</td>
<td>time series of daily frequency of node ( v ) in the censored news dataset</td>
</tr>
<tr>
<td>( \lambda_b(v) )</td>
<td>expected daily frequency of node ( v ) in data source ( b )</td>
</tr>
<tr>
<td>TSDF</td>
<td>time series of daily frequency</td>
</tr>
</tbody>
</table>

3.3 Methodology

3.3.1 Problem Formulation

Suppose we have a dataset of news reports and a dataset of tweets within a shared time period in a country of interest. Each news report or tweet is represented by a set of keywords and is indexed by a time stamp (e.g., day). We model the joint information of news reports and tweets using an undirected keyword co-occurrence graph \( G = (V, E) \), where \( V = \{1, 2, \ldots, n\} \) refers to the ground set of nodes/keywords, \( n \) refers to the total number of nodes, and \( E \subseteq V \times V \) is a set of edges, in which an edge \((i, j)\) indicates that the keywords \( i \) and \( j \) co-occur in at least one news report or tweet. Each node \( v \in V \) is associated with four attributes: \( \{a^t(v)\}_{t=1}^T \), \( \lambda_a(v) \), \( \{b^t(v)\}_{t=1}^T \), and \( \lambda_b(v) \) as defined in Table 3.3. As our study is based on the analysis of correlations between frequencies of keywords in the news and Twitter datasets, we only consider the keywords whose frequencies in these two datasets are well correlated (with correlations above a predefined threshold 0.15). Our goal is to detect a cluster (subset) of co-occurred keywords and a time window as an indicator of censorship.
pattern, such that the distribution of frequencies of these keywords in the news dataset is significantly different from that in the Twitter dataset.

Suppose the chosen time granularity is day and the shared time period is $\{1, \cdots, T\}$. We consider two hypotheses: under the null ($H_0$), the daily frequencies of each keyword $v$ in the news and Twitter datasets follow two different Poisson distributions with the mean parameters $\lambda_a(v)$ and $\lambda_b(v)$, respectively; under the alternative ($H_1(S,R)$), there is a connected cluster $S$ of keywords and a continuous time window $R \subseteq \{1, \cdots, T\}$, in which the daily frequencies of each keyword $v$ in the Twitter dataset follow a Poisson with an elevated mean parameter $q_a \cdot \lambda_a(v)$, but those in the news dataset follows a Poisson with a down-scaled mean parameter $q_b \cdot \lambda_b(v)$. Formally, they can be defined as follows:

Null hypothesis $H_0$:

$$a^t(v) \sim \text{Pos}(\lambda_a(v)), \forall v \in V, t \in \{1, \cdots, T\}$$
$$b^t(v) \sim \text{Pos}(\lambda_b(v)), \forall v \in V, t \in \{1, \cdots, T\}$$

Alternative hypothesis $H_1(S,R)$:

$$a^t(v) \sim \text{Pos}(q_a \cdot \lambda_a(v)), b^t(v) \sim \text{Pos}(q_b \cdot \lambda_b(v)), \forall v \in S, t \in R$$
$$a^t(v) \sim \text{Pos}(\lambda_a(v)), b^t(v) \sim \text{Pos}(\lambda_b(v)), \forall v \notin S \text{ or } t \notin R$$

where $q_a > 1, q_b < 1, S \subseteq V$, the subgraph induced by $S$ (denoted as $G_S$) must be connected to ensure that these keywords are semantically related, and $R \subseteq \{1, 2, \cdots, T\}$ is a continuous time window defined as $\{i, i + 1, \cdots, j\}, 1 \leq i \leq j \leq T$. Given the Poisson probability mass function denoted as $p(x; \lambda) = \lambda^x e^{-\lambda}/x!$, a generalized log likelihood ratio test (GLRT)
statistic can then be defined to compare these two hypotheses, and has the form:

\[
F(S, R) = \log \frac{\max_{q_a > 1} \prod_{t \in R} \prod_{v \in S} p(a^t(v); q_a \lambda_a(v))}{\prod_{t \in R} \prod_{v \in S} p(a^t(v); \lambda_a(v))} + \log \frac{\max_{q_b < 1} \prod_{t \in R} \prod_{v \in S} p(b^t(v); q_b \lambda_b(v))}{\prod_{t \in R} \prod_{v \in S} p(b^t(v); \lambda_b(v))}.
\]

(3.1)

In order to maximize the GLRT statistic, we need to obtain the maximum likelihood estimates of \(q_a\) and \(q_b\), which we set \(\partial F(S, R) / \partial q_a = 0\) and \(\partial F(S, R) / \partial q_b = 0\), respectively and get the best estimate \(\hat{q}_a = C_a / B_a\) of \(q_a\) and \(\hat{q}_b = C_b / B_b\) of \(q_b\) where \(C_a = \sum_{v \in S, t \in R} a^t(v)\), \(C_b = \sum_{v \in S, t \in R} b^t(v)\), \(B_a = \sum_{v \in S, t \in R} \lambda_a(v)\), \(B_b = \sum_{v \in S, t \in R} \lambda_b(v)\). Substituting \(q_a\) and \(q_b\) with the best estimations \(\hat{q}_a\) and \(\hat{q}_b\), we obtain the parametric form of the GLRT statistic as follows:

\[
F(S, R) = \left( C_a \log \frac{C_a}{B_a} + B_a - C_a \right) + \left( C_b \log \frac{C_b}{B_b} + B_b - C_b \right).
\]

(3.2)

Given the GLRT statistic \(F(S, R)\), the problem of censorship detection can be reformulated as Problem 1 that is composed of two major components: 1) **Highest scoring clusters detection.** The highest scoring clusters are identified by maximizing the GLRT statistic \(F(S, R)\) over all possible clusters of keywords and time windows; 2) **Statistical significance analysis.** The empirical p-values of the identified clusters are estimated via a randomization testing procedure [47], and are returned as significant indicators of censorship patterns in the news dataset, if their p-values are below a predefined significance level (e.g., 0.05).

**Problem 1 (GLRT Optimization Problem).** Given a keyword co-occurrence graph \(G(V, E)\) and a predefined significance level \(\alpha\), the GLRT optimization problem is to find the set of highest scoring and significant clusters \(\mathcal{O}\). Each cluster in \(\mathcal{O}\) is denoted as a specific pair of connected subset of keywords \((S_i \subseteq V)\) and continuous time window \((R_i \subseteq \{1, \cdots, T\})\), in
which \( S_i \) is the highest scoring subset within the time window \( R_i \):

\[
\max_{S \subseteq V} F(S, R_i) \quad \text{s.t.} \quad S \text{ is connected},
\]

(3.3)

and is significant w.r.t the significance level \( \alpha \).

3.3.2 GraphDPD Algorithm

Our proposed algorithm GraphDPD decomposes Problem 1 into a set of sub-problems, each of which has a fixed continuous time window, as shown in Algorithm 1. For each specific day \( i \) (the first day of time window \( R \) in Line 6) and each specific day \( j \) (the last day of time window \( R \) of Line 6), we solve the sub-problem (Line 7) with this specific \( R = \{i, i + 1, \cdots, j\} \) using Relaxed-GraphMP algorithm which will be elaborated later.

For each connected subset of keywords \( S \) returned by Relaxed-GraphMP, its p-value is estimated by randomization test procedure [47](Line 8). The pair \((S, R)\) will be added into the result set \( \mathcal{O} \) (Line 9) if its empirical p-value is less than a predefined significance level \( \alpha \) (e.g., 0.05). The procedure getPValue in Line 8 refers to a randomization testing procedure based on the input graph \( G \) to calculate the empirical p-value of the pair \((S, R)\) [47]. Finally, we return the set \( \mathcal{O} \) of significant clusters as indicators of censorship events in the news data set.

Line 7 in Algorithm 1 aims to solve an instance of Problem 1 given a specific time window \( R \), which is a set optimization problem subject to a connectivity constraint. Tung-Wei et. al. [35] proposed an approach for maximizing submodular set function subject to a connectivity constraint on graphs. However, our objective function \( F(S, R) \) is non-submodular as shown
Algorithm 1 GraphDPD

1: **Input**: Graph Instance $G$ and significant level $\alpha$;
2: **Output**: set of anomalous connected subgraphs $O$;
3: $O \leftarrow \emptyset$;
4: for $i \in \{1, \cdots, T\}$ do
5: for $j \in \{i + 1, \cdots, T\}$ do
6: \(R \leftarrow \{i, i + 1, \cdots, j\} ; // \) time window $R$
7: \(S \leftarrow \text{RELAXED-GRAPHMP}(G, R)\);
8: if $\text{getPValue}(G, S, R) \leq \alpha$ then
9: \(O \leftarrow O \cup (S, R)\);
10: end if
11: end for
12: end for
13: return $O$;

in Theorem 1 and this approach is not applicable here.

**Theorem 1.** Given a specific window $R$, our objective function $F(S, R)$ defined in (3.2) is non-submodular.

We propose a novel algorithm named RELAXED-GRAPHMP to approximately solve Problem 1 in nearly linear time with respect to the total number of nodes in the graph. We first transform the GLRT statistic in Equation (3.2) to a vector form. Let $x$ be an $n$-dimensional vector \((x_1, x_2, \cdots, x_n)^T\), where $x_i \in \{0, 1\}$ and $x_i = 1$ if $i \in S$, $x_i = 0$ otherwise. We define $\mathcal{P}, \mathcal{Q}, \Lambda_a, \Lambda_b$ as follows:

\[
\mathcal{P} = \left[ \sum_{t \in R} a'_(1), \cdots, \sum_{t \in R} a'_(n) \right]^T, \Lambda_a = [\lambda_a(1), \cdots, \lambda_a(n)]^T, \\
\mathcal{Q} = \left[ \sum_{t \in R} b'_(1), \cdots, \sum_{t \in R} b'_(n) \right]^T, \Lambda_b = [\lambda_b(1), \cdots, \lambda_b(n)]^T.
\]
Therefore, $C_a$, $C_b$, $B_a$, and $B_b$ in Equation (3.2) can be reformulated as follows:

$$
C_a = P^T x, \quad C_b = Q^T x, \quad B_a = |R|\Lambda_a^T x, \quad B_b = |R|\Lambda_b^T x
$$

Hence, $F$ can be reformulated as a relaxed function $\hat{F}$:

$$
\hat{F}(x, R) = P^T x \log \frac{P^T x}{|R|\Lambda_a^T x} + |R|\Lambda_a^T x - P^T x + Q^T x \log \frac{Q^T x}{|R|\Lambda_b^T x} + |R|\Lambda_b^T x - Q^T x
$$

(3.4)

We relax the discrete domain $\{0, 1\}^n$ of $S$ to the continuous domain $[0, 1]^n$ of $x$, and obtain the relaxed version of Problem 1 as described in Problem 2.

**Problem 2 (Relaxed GLRT Optimization Problem).** Let $\hat{F}$ be a continuous surrogate function of $F$ that is defined on the relaxed domain $[0, 1]^n$ and is identical to $F(S, R)$ on the discrete domain $\{0, 1\}^n$. The relaxed form of GLRT Optimization Problem is defined the same as the GLRT optimization problem, except that, for each pair $(S_i, R_i)$ in $\mathcal{O}$, the subset of keywords $S_i$ is identified by solving the following problem with $S_i = \text{supp}(\hat{x})$:

$$
\hat{x} = \arg \max_{x \in [0,1]^n} \hat{F}(x, R_i) \quad \text{s.t.} \quad \text{supp}(x) \text{ is connected.}
$$

where $\text{supp}(x) = \{i | x_i \neq 0\}$ is the support of $x$. The gradient of $\hat{F}(x, R)$ has the form:

$$
\frac{\partial \hat{F}(x, R)}{\partial x} = \log \frac{P^T x}{|R|\Lambda_a^T x} P + \left( |R| - \frac{P^T x}{\Lambda_a^T x} \right) \Lambda_a + \log \frac{Q^T x}{|R|\Lambda_b^T x} Q + \left( |R| - \frac{Q^T x}{\Lambda_b^T x} \right) \Lambda_b
$$

(3.5)
Algorithm 2 Relaxed-GraphMP

1: **Input**: Graph instance $G$, continuous time window $R$;
2: **Output**: the co-occurrence subgraph $G_S$;
3: $i \leftarrow 0$; $x^i \leftarrow$ an initial vector;
4: repeat
5: $\nabla \hat{F}(x^i, R) \leftarrow \frac{\partial F(x^i, R)}{\partial x^i}$ by Equation (3.5);
6: $g \leftarrow \text{Head}(\nabla \hat{F}(x^i, R), G)$; // Head projection step
7: $\Omega \leftarrow \text{supp}(g) \cup \text{supp}(x^i)$;
8: $b \leftarrow \arg \max_{x \in [0,1]^n} \hat{F}(x, R) \ s.t. \ \text{supp}(x) \subseteq \Omega$;
9: $x^{i+1} \leftarrow \text{Tail}(b, G)$; // Tail projection step
10: $i \leftarrow i + 1$, $S \leftarrow \text{supp}(x^i)$;
11: until halting condition holds;
12: return $G_S$;

Our proposed algorithm Relaxed-GraphMP decomposes Problem 2 into two sub-problems that are easier to solve: 1) a single utility maximization problem that is independent of the connectivity constraint; and 2) head projection and tail projection problems [29] subject to connectivity constraints. We call our method Relaxed-GraphMP which is analogous to GraphMP proposed by Chen et al. [11]. The high level of Relaxed-GraphMP is shown in Algorithm 2. It contains 4 main steps as described below.

- **Step 1**: Compute the gradient of relaxed GLRT problem (Line 5). The calculated gradient is $\nabla \hat{F}(x^i, R)$. Intuitively, it maximizes this gradient with connectivity constraint that will be solved in next step.

- **Step 2**: Compute the head projection (Line 6). This step is to find a vector $g$ so that the corresponding subset $\text{supp}(g)$ can maximize the norm of the projection of gradient $\nabla \hat{F}(x^i, R)$ (See details in [29]).

- **Step 3**: Solve the maximization problem without connectivity constraint. This step (Line 7,8) solves the maximization problem subject to the $\text{supp}(x) \subseteq \Omega$, where $\Omega$ is the union of the support of the previous solution $\text{supp}(x^i)$ with the result of head
projection supp(g) (Line 7). A gradient ascent based method is proposed to solve this problem. Details is not shown here due to space limit.

- **Step 4:** Compute the tail projection (Line 9). This final step is to find a subgraph $G_S$ so that $b_S$ is close to $b$ but with connectivity constraint. This tail projection guarantees to find a subgraph $G_S$ with constant approximation guarantee (See details in [29]).

- **Halting:** The algorithm terminates when the condition holds. Our algorithm returns a connected subgraph $G_S$ where the connectivity of $G_S$ is guaranteed by Step 4.

**Time Complexity Analysis:** The GraphDPD algorithm is efficient as its time complexity is proportional to the total number of continuous time windows $T^2$. Therefore, the time complexity of GraphDPD is mainly dependent on the run time of Relaxed-GraphMP. We give the detailed time complexity analysis in Theorem 2.

**Theorem 2.** GraphDPD runs in $O(T^2 \cdot t(nT + nl + |E|\log^3 n))$ time, where $T$ is the maximal time window size, $nT$ is the time complexity of Line 5 in Algorithm 1, $nl$ is the run time of Line 8 using gradient ascent, $|E|\log^3 n$ is the total run time of head projection and tail projection algorithms, and $t$ is the total number of iterations needed in Algorithm 2.

### 3.4 Experiments

Through experiments conducted on semi-synthetic data and real data, we evaluated the performance of our proposed approach in censorship pattern detection compared with alternative methods. The results of these experiments show the superiority of GraphDPD algorithm for detecting likely patterns of media self-censorship over other state-of-the-art options.
3.4.1 Experimental Design

**Real world datasets:** Table 3.4 gives a detailed description of real-world datasets we used in this work. Details of Twitter and news data access have been provided in Section 3. Daily keyword co-occurrence graphs, which integrate *News* with *Twitter*, are generated as described in Section 3.2.1.

Table 3.4: Real-world dataset used in this work. Tweets: average number of daily tweets. News Articles: average number of daily local news articles. News Outlets: number of news outlets used in this study.

<table>
<thead>
<tr>
<th>Country</th>
<th>Daily Tweets</th>
<th>Daily News Articles</th>
<th># of News Outlets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>84556</td>
<td>444</td>
<td>13</td>
</tr>
<tr>
<td>Venezuela</td>
<td>65916</td>
<td>91</td>
<td>6</td>
</tr>
</tbody>
</table>

**Data Preprocessing:** The preprocessing of the real world datasets has been discussed in detail in Section 3.2.1. In particular, we considered keywords whose day-by-day frequencies in news media and Twitter data have linear correlations above 0.15, in order to filter noisy keywords.

**Semi-synthetic datasets:** We create semi-synthetic datasets by using the coordinates from real-world datasets and injecting anomalies [68].

**Our proposed Graph-DPD and baseline methods:** We compare our proposed method with one baseline method LTSS [48], which finds anomalous but not necessarily connected subsets of data records by maximizing a score function. We also compare our proposed
method with two state-of-art baseline methods designed specifically for connected anomalous subgraph detection, namely, EventTree [60] and NPHGS [10].

**Performance Metrics:** The performance metrics include: (1) precision (Pre), (2) recall (Rec), and (3) f-measure (F-score). Given the returned subset of nodes $S$ and the corresponding true subset of anomalies $S^*$, we can calculate these metrics as follows:

\[
\text{Pre} = \frac{|S \cap S^*|}{|S|}, \quad \text{Rec} = \frac{|S \cap S^*|}{|S^*|}, \quad \text{F-score} = \frac{2|S \cap S^*|}{|S^*| + |S|}
\]

**Collecting labels for real data:** We collect labels for real-world instances of censorship from all abnormal absence patterns identified in *News* by all baseline methods. For each abnormal absence pattern in *News*, we need to first identify if there are any relevant events of interest taking place around the associated time region. Although there is no publicly available gathered information of all existing censorship, we can validate the correctness of the detected self-censorship through evidences in news reports. For example, news articles\(^2\)\(^3\) verified self-censorship in Ultimas Noticias, the largest daily in Venezuela, about the massive protests in February 2014. An indicator of censorship pattern is also considered as valid if we can find the event of interest is: 1) not reported in some local news outlets while reported in some different local news outlets, 2) reported in influential international news outlets, and 3) reported of censorship activity in local news media from other news outlets during the associated time window. The evaluation process is analyzed with the inner-annotator agreement by multiple independent annotators.

\(^2\)http://www.nybooks.com/daily/2014/04/09/venezuela-protests-censorship/
3.4.2 Semi-synthetic Data Evaluation

**Semi-synthetic datasets:** We create semi-synthetic datasets by using the coordinates from real-world datasets and injecting anomalies.

Ten daily keyword co-occurrence graphs are randomly selected to inject with random true anomaly connected subgraphs using a random walk algorithm [68] with a restart probability of 0.1. The number of nodes in the true anomaly subgraph is \( x \) percentage of the number of nodes in the daily co-occurrence graph, where \( x \in \{0.05, 0.1, 0.15\} \). For convenience but without loss of generality, we specified \( q_t \cdot q_n = 1.0 \), where \( q_t \) controls the scale of anomaly in tweets and \( q_n \) controls scale of anomaly in local news articles. In our experiments, we set \( q_t = \{1.0, 2.0, \ldots, 10.0, 15.0, \ldots, 35.0\} \), and set \( q_n = 1/q_t \) correspondingly.

**Settings for Graph-DPD and baseline methods:** The maximal window size \( T \) and result threshold \( \alpha \) in Graph-DPD are set as 7 and 0.05 respectively. However, our algorithm is not sensitive to the setting of \( T \) and \( \alpha \). We compare our proposed method with one baseline method LTSS [48], which finds anomalous but not necessarily connected subsets of data records by maximizing a score function. We also compare our proposed method with two state-of-art baseline methods designed specifically for connected anomalous subgraph detection, namely, EventTree [60] and NPHGS [10]. Model parameters are tuned following the original papers. Specifically, for EventTree we tested \( \lambda = \{0.0001, 0.0006, \ldots, 0.001, 0.006, \ldots, 0.010, 0.015, \ldots, 0.1, 0.5, 1.0, \ldots, 20.0\} \). For NPHGS, we set the number of seed entities \( K = 400 \) and typical significance levels \( \alpha_{\text{max}} = 0.15 \) as the authors suggested. Since the baseline methods are designed to detect anomalies on one data source at one time, they are tested separately on Twitter and News, which are labeled as LTSS\(_{\text{News}}\), LTSS\(_{\text{Twitter}}\), EventTree\(_{\text{News}}\), EventTree\(_{\text{Twitter}}\), NPHGS\(_{\text{News}}\) and NPHGS\(_{\text{Twitter}}\). Specifically, LTSS\(_{\text{Twitter}}\), EventTree\(_{\text{Twitter}}\) and NPHGS\(_{\text{Twitter}}\) are burst detection baseline methods while LTSS\(_{\text{News}}\),
EventTreeNews, and NPHGSNews are absenteeism detection baseline methods by some transformations on attributes.

**Performance Metrics:** The performance metrics include: (1) precision (\(\text{Pre}\)), (2) recall (\(\text{Rec}\)), and (3) f-measure (F-score). Given the returned subset of nodes \(S\) and the corresponding true subset of anomalies \(S^*\), we can calculate these metrics as follows:

\[
\text{Pre} = \frac{|S \cap S^*|}{|S|}, \quad \text{Rec} = \frac{|S \cap S^*|}{|S^*|}, \quad \text{F-score} = \frac{2|S \cap S^*|}{|S^*| + |S|}
\]

We evaluate the accuracy of our approach to detect the disrupted ground truth anomalies. Figure 3.2 shows the average precision, recall, and F-measure in detecting the injected anomalies using the semi-synthetic data. We find that overall our approach consistently outperforms all other baseline methods.

**Detection power.** (1) **Our approach.** Our approach outperforms baseline methods especially at low perturbation intensities where the detection is harder to carry out, and the performance increases gradually with the increase of perturbation intensity. In particular, it has a high accuracy of detecting injected anomalies when \(q_t \geq 10\) regardless of the size of injected anomalies. Measures of recall using NPHGSTwitter are as good as our approach while the other baseline methods are significantly worse than our approach especially when the size of disrupted ground truth anomalies is small and perturbation intensity is low. However, the measures of precision using NPHGSTwitter are much worse than our approach. Considering overall F-score, NPHGSNews and NPHGSTwitter look similar to our approach when perturbation intensity is low while much worse than our approach when perturbation intensity is high. When we increase \(q_t\), EventTree based methods perform worse than our approach, especially when the size of ground truth anomalies is small. (2) **NPHGS.** When
$q_t \in \{1.0, 2.0\}$ and true ratio $x \in \{0.05, 0.10\}$, the precision of $\text{NPHGS}_{\text{News}}$ is better than our method. However, when $x = 0.15$, the recall of $\text{NPHGS}_{\text{News}}$ becomes quite low, which indicates its poor behavior when true subgraph is relatively large. (3) **EventTree.** The recall of $\text{EventTree}_{\text{News}}$ and $\text{EventTree}_{\text{Twitter}}$ is among the best when $q_t$ is less than 2.0. The reason is that results of EventTree are easier affected by noise nodes. (4) **LTSS.** In general, LTSS did well in average recall but poorly in average precision as the size of anomalous graph increases. Hence, our approach outperforms the baseline by detecting connected clusters of keywords.
Figure 3.1: Example TSDF in News vs. TSDF in Twitter for a set of connected keywords. These keywords are relevant to the 43 missing students from Ayotzinapa Rural Teachers’ College on Sep 26th, 2014 in Mexico. We can find consistent censorship pattern in El Mexicano Gran Diario Regional (el-mexicano.com.mx) shortly after the students are missing.
Figure 3.2: Anomaly detection results (mean precision (left), recall (center), and F-measure (right) vs. perturbation intensity) for different anomaly subgraph sizes (increased size from top to bottom) in semi-synthetic data. X-axis represents $q_t$, which implies the scale of anomaly injected in Twitter. $q_n$, which implies the scale of anomaly injected in News, is varied following $q_t \times q_n = 1.0$. 
Table 3.5: Example indicators of censorship identified by our approach in Mexico and Venezuela during Year 2014 (with significance level $\leq 0.05$)

<table>
<thead>
<tr>
<th>Date</th>
<th>Example censored keywords</th>
<th>Example local news media detected with censorship</th>
<th>Reasons for censorship in news media</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-05-01</td>
<td><code>reforma(reform), gasolina(petrol), educación(education)</code></td>
<td>Noroeste</td>
<td>Tens of thousands of people marched in Mexico City on Labor Day to protest the new laws, which target at Mexico’s education system and opening up the state controlled oil industry to foreign investors.</td>
</tr>
<tr>
<td>2014-09-27</td>
<td><code>ayotzinapa, iguala, normalistas, desaparecidos(missing), detenidos(detained), protesta(protest)</code></td>
<td>El Mexicano Gran Diario Regional</td>
<td>43 students from the Ayotzinapa Rural Teachers’ College went missing and kidnapped in Iguala on September 26, 2014.</td>
</tr>
<tr>
<td>2014-11-10</td>
<td><code>ayotzinapa, estudiantes(students), normalistas, desaparecidos(missing), protesta(protest), militares(military), iguala</code></td>
<td>El Mexicano Gran Diario Regional</td>
<td>Protests in Mexico City demanding the return of the missing students, who came from Ayotzinapa Rural Teachers’ College and went missing in Iguala on September 26, 2014, turned violent for the first time.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014-02-18</td>
<td><code>represión(repression), disparó(shooting), marchamos(march), heridos(wounded), nicolasmaduro, armados(armed), leopoldolopez, apresar(arrest)</code></td>
<td>Ultimas Noticias</td>
<td>Mass protests led by opposition leaders, including Leopoldo López, occurred in 38 cities across Venezuela asking for the release of the arrested students.</td>
</tr>
<tr>
<td>2014-05-01</td>
<td><code>muertes(deaths), cambio(change), caracas, presidente(president), labor</code></td>
<td>El Tiempo in Anzoategui</td>
<td>Thousands of Venezuelans demonstrated in Caracas to commemorate Labor Day and denounce shortages.</td>
</tr>
<tr>
<td>2014-08-12</td>
<td><code>gubernamental(government), anticontrabando, contrabando, ébola, muerte(death)</code></td>
<td>El Nacional</td>
<td>Venezuela is the only country in Latin America with increasing number of malaria. With the spreading of Ebola virus, Venezuela is one of the most vulnerable countries in Latin America.</td>
</tr>
</tbody>
</table>
3.4.3 Real Data Evaluation

As before, we applied three anomaly detection baseline methods, LTSS, NPHGS, and Event-Tree, to detect anomalies in News on graphs with all possible time windows from 3 days to 7 days during Year 2014. The baseline methods can find anomalous subgraphs according to their own score functions; however, they are not able to evaluate the significance level of each subgraph. For the purpose of comparison, we remove duplicate subgraphs with overlapping time regions in the same manner as our method. The remaining subgraphs are ranked from the best to the worst according to their function values and the same number of subgraphs are selected from the top to compare with subgraphs detected by our method.

Table 3.6: Comparison of false positive rates in censorship detection between GraphDPD and three baseline methods: LTSS, NPHGS, and EventTree on real data of Mexico and Venezuela during year 2014.

<table>
<thead>
<tr>
<th>Country</th>
<th>LTSS</th>
<th>NPHGS</th>
<th>EventTree</th>
<th>GraphDPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>0.722</td>
<td>0.667</td>
<td>0.556</td>
<td>0.278</td>
</tr>
<tr>
<td>Venezuela</td>
<td>0.714</td>
<td>0.786</td>
<td>0.643</td>
<td>0.357</td>
</tr>
</tbody>
</table>

Table 6 summarizes the comparison of false positive rates in censorship detection and our method outperforms LTSS, NPHGS, and EventTree. The baseline methods, which are designed for event detection instead of censorship detection, will capture all falling patterns in News. In particular, the baseline methods are not able to successfully differentiate censored events from non-censored events, e.g., the normal end of attention paid to breaking events.

We also looked at whether our method performed well in avoiding detection of events that are only of interest to social media. If it were detecting topics only of interest to social
media, we would expect a large number of detected censorship events to be covered by no
news sources. Figure 3.3 shows a density plot of the proportion of sources reporting the
detected events in our real-world data. It clearly shows that few of the events detected by
the algorithm are completely missing from news media, i.e. that the proportion of sources
reporting the event is zero.

3.4.4 Case Studies

This section provides some more detailed case studies to illustrate both the method and the
types of cases that were self-censored. Table 3.5 summarizes a list of example instances of
censorship identified by our approach in Mexico and Venezuela with significance level \( \leq 0.05 \).
The rest of this section will evaluate a couple of these instances in greater detail.

**Mexico May 2014.** In December 2013, Mexican president and Congress amended the
Constitution, opening up the state controlled oil industry to foreign investors. Tens of
thousands of protesters demonstrated in Mexico City on Labor Day (May 1) to protest
against the energy reform. In additions, protesters were also unsatisfied with the 2013
reforms of the educational sector. However, this incident was not reported in a number
of influential newspapers in Mexico, which is an indicator of censorship. Fig. 3.4a shows
a cluster of censored keywords detected by our method around May 1, 2014 in Mexico.
Our approach has successfully captured consistent censorship patterns among a collection of
relevant keywords, which well describe the topics around which the May 1 demonstrations
were organized (reforma, gasolina, dinero, educación, escuela).

**Venezuela February 2014.** As a result of the collapse of the price of oil, Venezuela
suffered from inflation, shortages of basic foodstuffs and other necessities. Mass opposition
protests led by opposition leaders demanding the release of the students occurred in 38 cities
Figure 3.3: The density of the proportion of sources reporting events that are detected as censored by our model. The results suggest that we are capturing very few events that are only of interest to social media.
across Venezuela on February 12, 2014. While this incident was reported by a number of major international newspapers, there was significant censorship in the country’s largest daily Ultimas Notícias, an event reported by a number of international news outlets. Fig. 3.4b shows a cluster of censored keywords detected by our method around February 18, 2014 in Venezuela, which well describes the populations involved (estudiante, chavistas, opositores, leopoldolopez) and the target of the demonstrations (nicolasmaduro).
Figure 3.4: Word cloud representing censored keywords in News identified by our method.
Chapter 4

Predicting Information Credibility in Social Media

4.1 Introduction

Nowadays, general public are exposed to large amount of data through fast-growing social media platform and online news media. While the internet users are gaining benefits from the rapid information flow on various news topics, users are experiencing questionable information credibility at the same time. One of the solutions to verify the correctness of information is referring to the fact-checking websites, for example Snopes and PolitiFact, which accept questions about whether an event or a claim is trustworthy from users. The fact-checking websites then rely on domain experts to verify the validity of the information through research materials. Hence, this fact-checking process is very expensive and can not be generalized to catch up with the pace of upcoming information.
The difficulties existed in fact checking through human efforts show the necessity of designing algorithms to predict the credibility of information automatically. A number of studies have been viewing this problem as a binary classification problem and have proposed sets of features to improve the prediction accuracy. With different data source of interest, [9] and [61] extracted features from headlines and news contents to detect misinformation in online news media while [21, 57, 78, 3, 28, 37, 27, 8] extracted features from topic, user, and propagation aspect to detect misinformation in social media. However, most of these studies focus on classifying the whole news topic instead of single piece of information. It is noticeable that pieces of information can have quite different credibility level with different opinions or sentiment even they are discussing about the same news topic.

In this chapter, we present a cluster-based credibility framework, where we improve the accuracy of instance-level credibility prediction through grouping instances with potentially similar credibility level into clusters and estimating the credibility of clustered instances with aggregated features. We also formulate a new benchmark dataset for future fake news research as the publicly available datasets for fake news study have issues on either the topics included or the ground-truth labels.

The main contributions of our study are summarized as follows:

- **New Benchmark Dataset for fake news research**: Existing research on fake news have made a few datasets publicly available, however, there are no agreed upon benchmark datasets. The reason behind this issue is mainly because of the topics included in the datasets or the way that ground-truth labels are collected. We build a new benchmark dataset on a wide range of topics with ground truth labels collected from fact-checking website Snopes.
• **Formulation of a cluster-based credibility prediction framework**: We propose a cluster-based framework to predict information credibility at the instance level by grouping instances with potentially similar credibility level into clusters. The instances are clustered based on their linguistic and propagation network.

• **Extensive experiments to validate the proposed techniques**: We conduct comprehensive experiments on real-world Twitter data to evaluate our proposed approach. The results demonstrate that our proposed approach outperforms baseline methods in the accuracy of predicting information credibility at the instance level.

The remainder of this chapter is organized as follows. Section 4.2 describes the dataset we used in this work, features extracted at both instance level and topic level, and our proposed framework for information credibility prediction. The experimental results on real world data are presented in Section 4.3.

### 4.2 Methodology

In this section, we present a new dataset including Twitter posts and news articles created for fake news study. In addition, we propose one set of topic-level features and another set of instance-level features to characterize our dataset. Furthermore, we develop a framework utilizing cluster-based information to improve instance-level credibility prediction.

#### 4.2.1 Dataset

As summarized in [65], there are a few publicly available datasets for fake news research. However, there are no agreed upon benchmark datasets due to the contents included in the datasets or the way that the datasets are collected. Dataset *BuzzFeedNews* comprises news
published in Facebook from 9 news agencies from September 19 to 23 and September 26 and 27 during the 2016 U.S election. Dataset *LIAR* collects 12,836 short statements, rather than the entire news contents, from fact-checking website PolitiFact. Dataset *BS Detector* is collected from a browser extension and the labels are outputs of *BS Detector* instead of ground truth annotated by human experts. Dataset *CREDBANK* includes a large scale of tweets over a 3-month time period. The tweets are grouped to over 1,000 news events and each news event is labeled with a list of credibility scores given by 30 annotators from Amazon Mechanical Turk. However, the list of credibility scores given to the news event might cover a wide range of values and the lacking of expert annotators might lead to difficulties in determining the credibility of the news event.

Issues associated the existing publicly available datasets for fake news research give us insights to the necessity of building a new dataset on a wide range of topics with ground truth labels annotated by human experts. We utilize the widely used fact-checking website Snopes to collect ground truth labels of news topics. The labels given to the news topics on Snopes include False, Mostly False, Mixture, Mostly True, and True. For the accuracy of our dataset, we only consider news topics labeled as False and True as the news topics of interest. The relevant Twitter posts and news articles are collected upon keyword matching criteria and the sets of keywords used to represent the news topics are generated automatically based on the URLs and the titles of reportings on Snopes. The URL of reporting on Snopes is a good starting point for composing the set of keywords to represent the news topic as the URL includes a set of keywords that briefly summarize the news reporting. In addition to the set of keywords embedded in the URL, we also add the nouns and verbs in the title of news reporting to the set of keywords representing the news topic.

Suppose a news topic $\mathbb{T}$ is represented with a set of keywords $t_1, t_2, \cdots, t_n$, a Twitter post is relevant to news topic $\mathbb{T}$ if it contains at least $n/2 + 1$ keywords in the set $t_1, t_2, \cdots, t_n$. The
EMBERS project [58] utilized Gnip Decahose Twitter data, which provided a 10% random sample of the realtime Twitter Firehose through a streaming connection. Based on the Gnip Decahose Twitter data, we filtered relevant Twitter posts with respect to the news topics of interest. News topics having less than 100 Twitter posts are removed as the topics are not receiving much attention in the social media. Furthermore, we collected relevant news articles by querying top 20,000 search results in Bing Search Engine for each news topic of interest. The following attributes of the news articles are recorded in our dataset: URL, title, body, date and time of publication, and publisher. Table 4.1 gives a detailed description of the dataset we used in this work, including number of true and false news topics and average volume of tweets per topic.

<table>
<thead>
<tr>
<th>Total number of topics</th>
<th>701</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of false topics</td>
<td>311</td>
</tr>
<tr>
<td>Number of true topics</td>
<td>390</td>
</tr>
<tr>
<td>Average number of tweets per topic</td>
<td>2561</td>
</tr>
</tbody>
</table>

Table 4.1: Description of the dataset we used in this work.

4.2.2 Feature Extraction

Existing studies have been exploring features for predicting information credibility in social media from a variety of categories. Here we follow the categories and propose a set of features can be extracted from our dataset. There are two strategies to extract the features: topic-level based and instance-level based. The differences between these two strategies can be concluded in two aspects: 1) topic-level based features represent aggregated information
within a collection of tweets while instance-level based features represent information from every single tweet, and 2) instance-level based features lose information of interaction among tweets, e.g. retweet information.

Table 4.2 lists a set of features extracted at topic level and Table 4.3 lists a set of features extracted at instance level. The instance-level features are prepared for the baseline methods of instance-level credibility predication while our proposed approach utilizes topic-level features to improve instance-level credibility prediction. The topic-level features can be grouped into three categories: message-based, user-based, and propagation-based while instance-level features can be grouped into two categories: message-based and user-based. Note that all of these features are independent of the news topics and hence, models trained with our dataset can be generalized to new topics as well.

**Message-based features** consider characteristics of messages, where some of the features can be generalized to other social media and some of the features are only specific to Twitter. The sentiment score is estimated with the Sentiment Annotator from Stanford CoreNLP and given one of three values representing positive sentiment, neutral sentiment, and negative sentiment. Aggregation among a collection of tweets is performed through two functions: sum and average.

**User-based features** consider characteristics of users who post the tweets. Aggregation among a collection of tweets is performed through the averaging function.

**Propagation-based features** consider characteristics of networks built upon retweet information. Each node in the network represents a single tweet and each edge represents a retweet relation. Hence, this category is not applicable to instance-level features.
<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message</td>
<td>Average Retweet Count</td>
<td>Average number of retweets</td>
</tr>
<tr>
<td></td>
<td>Average Favorite Count</td>
<td>Average number of favorites</td>
</tr>
<tr>
<td></td>
<td>Average Sentiment Score</td>
<td>Average score of sentiment analysis</td>
</tr>
<tr>
<td></td>
<td>Number of Tweets</td>
<td>Total number of tweets</td>
</tr>
<tr>
<td></td>
<td>Average length</td>
<td>Average length of tweets, in characters</td>
</tr>
<tr>
<td></td>
<td>Fraction of Question Mark</td>
<td>Fraction of tweets containing ‘?’</td>
</tr>
<tr>
<td></td>
<td>Fraction of Exclamation Mark</td>
<td>Fraction of tweets containing ‘!’</td>
</tr>
<tr>
<td></td>
<td>Fraction of URL</td>
<td>Fraction of tweets containing a URL</td>
</tr>
<tr>
<td></td>
<td>Fraction of User Mention</td>
<td>Fraction of tweets containing user mentions</td>
</tr>
<tr>
<td></td>
<td>Fraction of Hashtag</td>
<td>Fraction of tweets containing hashtags</td>
</tr>
<tr>
<td></td>
<td>Fraction of IsRetweet</td>
<td>Fraction of tweets containing ‘RT ’</td>
</tr>
<tr>
<td></td>
<td>Number of Distinct Short URLs</td>
<td>Total number of distinct short URLs</td>
</tr>
<tr>
<td></td>
<td>Number of Distinct Hashtags</td>
<td>Total number of distinct hashtags</td>
</tr>
<tr>
<td></td>
<td>Number of Distinct Long URLs</td>
<td>Total number of distinct expanded URLs</td>
</tr>
<tr>
<td></td>
<td>Number of Distinct User Mentions</td>
<td>Total number of distinct user mentions</td>
</tr>
<tr>
<td></td>
<td>Number of Distinct Authors</td>
<td>Total number of distinct users</td>
</tr>
<tr>
<td>User</td>
<td>Fraction Verified</td>
<td>Fraction of verified user</td>
</tr>
<tr>
<td></td>
<td>Average Statuses Count</td>
<td>Average number of statuses</td>
</tr>
<tr>
<td></td>
<td>Average Followers Count</td>
<td>Average number of followers</td>
</tr>
<tr>
<td></td>
<td>Average Favorites Count</td>
<td>Average number of favorites</td>
</tr>
<tr>
<td></td>
<td>Average Friends Count</td>
<td>Average number of friends</td>
</tr>
<tr>
<td></td>
<td>Average Listed Count</td>
<td>Average number of listed</td>
</tr>
<tr>
<td></td>
<td>Average Account Length</td>
<td>Average length of user account</td>
</tr>
<tr>
<td>Propagation</td>
<td>Number of Nodes</td>
<td>Total number of nodes</td>
</tr>
<tr>
<td></td>
<td>Number of Isolated Nodes</td>
<td>Total number of isolated nodes</td>
</tr>
<tr>
<td></td>
<td>Fraction of Isolated Nodes</td>
<td>Fraction of isolated nodes</td>
</tr>
<tr>
<td></td>
<td>Diameter</td>
<td>Diameter of the propagation network</td>
</tr>
<tr>
<td></td>
<td>Radius</td>
<td>Radius of the propagation network</td>
</tr>
<tr>
<td></td>
<td>Max Degree</td>
<td>Maximum degree of node</td>
</tr>
<tr>
<td></td>
<td>Average Degree</td>
<td>Average degree of node</td>
</tr>
<tr>
<td></td>
<td>Number of Connected Components</td>
<td>Number of connected components</td>
</tr>
<tr>
<td></td>
<td>Max Connected Size</td>
<td>Size of the largest connected component</td>
</tr>
</tbody>
</table>

Table 4.2: Features extracted at topic level and grouped into three categories: Message, User, and Propagation.
<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message</td>
<td>Posting Platform</td>
<td>Mobile or Web</td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>Length of tweet, in characters</td>
</tr>
<tr>
<td></td>
<td>Number of Words</td>
<td>Number of words</td>
</tr>
<tr>
<td></td>
<td>Number of URLs</td>
<td>Number of URLs</td>
</tr>
<tr>
<td></td>
<td>Number of Hashtags</td>
<td>Number of hashtags</td>
</tr>
<tr>
<td></td>
<td>Number of Unique Characters</td>
<td>Number of unique characters</td>
</tr>
<tr>
<td></td>
<td>Number of Stock Symbol</td>
<td>Number of ‘$’</td>
</tr>
<tr>
<td></td>
<td>Number of Positive Emoticons</td>
<td>Number of ‘:-)’ and ‘;-)’</td>
</tr>
<tr>
<td></td>
<td>Number of Negative Emoticons</td>
<td>Number of ‘:-(' and ‘;-('</td>
</tr>
<tr>
<td></td>
<td>Number of ‘via’</td>
<td>Number of ‘via’</td>
</tr>
<tr>
<td></td>
<td>Number of Colon Symbol</td>
<td>Number of ‘:’</td>
</tr>
<tr>
<td></td>
<td>Number of Swear Words</td>
<td>Number of swear words</td>
</tr>
<tr>
<td></td>
<td>Number of Pronouns</td>
<td>Number of pronouns</td>
</tr>
<tr>
<td></td>
<td>Number of Self Mentions</td>
<td>Number of ‘I’, ‘my’, and ‘mine’</td>
</tr>
<tr>
<td></td>
<td>Retweet Count</td>
<td>Retweet count</td>
</tr>
<tr>
<td></td>
<td>Favorites Count</td>
<td>Favorites count</td>
</tr>
<tr>
<td></td>
<td>User Mention Count</td>
<td>User Mention Count</td>
</tr>
<tr>
<td></td>
<td>IsRetweet</td>
<td>Is retweet</td>
</tr>
<tr>
<td></td>
<td>IsReply</td>
<td>Is reply</td>
</tr>
<tr>
<td>User</td>
<td>Followers Count</td>
<td>Followers count</td>
</tr>
<tr>
<td></td>
<td>Friends Count</td>
<td>Friends count</td>
</tr>
<tr>
<td></td>
<td>Length of Account</td>
<td>Length of account, in days</td>
</tr>
<tr>
<td></td>
<td>User Location</td>
<td>User has location in profile</td>
</tr>
<tr>
<td></td>
<td>Ratio of Statuses to Followers</td>
<td>Ratio of statuses to followers</td>
</tr>
<tr>
<td></td>
<td>Ratio of Friends to Followers</td>
<td>Ratio of friends to followers</td>
</tr>
<tr>
<td></td>
<td>Statuses Count</td>
<td>Statuses count</td>
</tr>
<tr>
<td></td>
<td>Verified</td>
<td>User is verified</td>
</tr>
<tr>
<td></td>
<td>Length of Summary</td>
<td>Length of summary</td>
</tr>
<tr>
<td></td>
<td>Length of Screen Name</td>
<td>Length of screen name</td>
</tr>
<tr>
<td></td>
<td>Has URL</td>
<td>User has URL in profile</td>
</tr>
<tr>
<td></td>
<td>Favorites Count</td>
<td>Favorites count</td>
</tr>
<tr>
<td></td>
<td>Listed Count</td>
<td>Listed count</td>
</tr>
</tbody>
</table>

Table 4.3: Features extracted at instance level and grouped into two categories: Message and User.
4.2.3 Cluster-based Prediction

Instance-level credibility prediction is generally more complicated than topic-level credibility prediction due to the fact that less useful features extracted potentially and features might be subject to higher noise level. Predicting instance-level credibility in our dataset is even more sophisticated. In addition to identifying different credibility level between false instances within false topics and true instances within true topics, we are also interested in identifying different credibility level between (1) true instances and false instances within false topics, and (2) true instances and false instances within true topics. It’s not trivial as instances within the same topic can be quite similar in many aspects and different viewpoints can lead to quite different credibility level.

Here we propose a cluster-based prediction framework, where instances with potentially similar credibility level are grouped into clusters. The credibility of the clusters are estimated with topic-level features and instances within the same cluster share the same credibility level. One of the common approaches for clustering text instances is representing text in vectors with TF-IDF or doc2vec and calculating the distance between the vectors. However, this approach does not work well on our data as instances of different credibility level within the same topic can be very similar in the vector and easily grouped into the same cluster. We notice that the propagation network is a good starting point for grouping tweets in the same thread into the same cluster.

Algorithm 3 gives a description of our proposed approach. Given a collection of tweets, the algorithm outputs a set of clusters which instances in the same cluster will share the same credibility level. The algorithm can be divided into two major steps: set up the propagation networks and merge similar propagation networks into set of clusters.

The propagation networks are built upon retweeting relation and the edges in the propagation networks are undirected as the set of features do not characterize the information
flow. The clusters are further merged if the root tweets in the clusters have an overlapping of greater than 80% in tokens. Here the root tweet in a cluster is defined as the tweet with the earliest time stamp, which is also treated as the origin of information in the cluster. The procedure terminates if the clusters cannot be merged any further. The credibility level of a cluster is estimated with the aggregated topic-level features and the instances in the cluster is assigned with the same credibility level as the cluster.

Algorithm 3 Cluster-based Credibility Prediction in Social Media

1: Input: Tweet Instances $T$
2: Output: set of clusters which instances in the same cluster will share the same credibility level $O$
3: $O \leftarrow \emptyset$
4: for $i \in \{1, \ldots, n\}$ do
5: for $j \in \{i + 1, \ldots, n\}$ do
6: if $T_i$ retweeted $T_j$ or $T_j$ retweeted $T_i$ then
7: add edge $(T_i, T_j)$ to $O$
8: end if
9: end for
10: end for
11: while $C \leftarrow O$
12: do
13: for pair of cluster $(C_i, C_j)$ in $C$ without repeating do
14: $p \leftarrow$ tokens of tweet with the earliest time stamp in $C_i$
15: $q \leftarrow$ tokens of tweet with the earliest time stamp in $C_j$
16: if $\text{Intersection}(p, q)/\text{Union}(p, q) \geq 0.8$ then
17: add connected $C_i, C_j$ to $C$ and delete $C_i, C_j$ from $C$
18: end if
19: end for
20: if $C = \emptyset$ then break
21: end if
22: $O \leftarrow C$
23: end while
24: return $O$;
4.3 Experiments

In this section, we evaluate the performance of our proposed approach in instance-level credibility prediction compared with the baseline methods. The results of these experiments show the superiority of our proposed approach for predicting information credibility at the instance level over the baseline methods.

4.3.1 Quantitatively Evaluation

In order to find out the classifier that can achieve the best performance, we tested on dataset for a 5-fold cross validation with six widely used classifiers: k nearest neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (RT). The classifiers are tested on topic-level features as well as instance-level features.

In addition, MI-SVM [4] is also a good fit for our problem if we assume: (1) false news topics include true tweets and at least one false tweet, and (2) false tweets in true news topics can be ignored. Here each news topic represent a bag and relevant tweets represent instances in the bag. Notice that our dataset only has ground-truth labels for the news topics, which can be used as labels for MI-SVM and topic-level classification. However, ground-truth labels for every single tweet are very expensive to collect due to the size of the dataset. As an assumption, every single tweet inherit the label of its relevant news topic for the purpose of selecting the classifier with the best performance.

Table 4.4 summarizes the performance of the classifiers including accuracy (Acc), recall (Rec), precision (Pre), and F-1 score. As shown in Table 4.4, SVM outperforms the other classifiers in both topic-level and instance-level experiments. Therefore, we utilize SVM to
compare performance of our proposed approach in credibility prediction with the baseline methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Classifier</th>
<th>Acc</th>
<th>Rec</th>
<th>Pre</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instance-level classification</td>
<td>kNN</td>
<td>0.5937 (0.0112)</td>
<td>0.4873 (0.0232)</td>
<td>0.5470 (0.0130)</td>
<td>0.5153 (0.0186)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td><strong>0.6442 (0.0174)</strong></td>
<td>0.4780 (0.0267)</td>
<td><strong>0.6305 (0.0253)</strong></td>
<td><strong>0.5436 (0.0255)</strong></td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.5629 (0.0395)</td>
<td>0.4930 (0.3067)</td>
<td>0.5677 (0.0950)</td>
<td>0.4420 (0.1915)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.6296 (0.0188)</td>
<td>0.4316 (0.0293)</td>
<td>0.6178 (0.0289)</td>
<td>0.3047 (0.0294)</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.6275 (0.0204)</td>
<td>0.4767 (0.0334)</td>
<td>0.6006 (0.0276)</td>
<td>0.5314 (0.0310)</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.5781 (0.0114)</td>
<td><strong>0.4990 (0.0233)</strong></td>
<td>0.5258 (0.0135)</td>
<td>0.5120 (0.0170)</td>
</tr>
<tr>
<td></td>
<td>MI-SVM</td>
<td>0.5550 (0.0029)</td>
<td>0.0065 (0.0080)</td>
<td>0.1667 (0.2108)</td>
<td>0.0124 (0.0152)</td>
</tr>
<tr>
<td>Topic-level classification</td>
<td>kNN</td>
<td>0.8773 (0.0084)</td>
<td>0.8199 (0.0194)</td>
<td>0.8960 (0.0273)</td>
<td>0.8557 (0.0087)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td><strong>0.9120 (0.0295)</strong></td>
<td><strong>0.9132 (0.0261)</strong></td>
<td>0.8926 (0.0459)</td>
<td><strong>0.9022 (0.0306)</strong></td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>0.6719 (0.0728)</td>
<td>0.3347 (0.2210)</td>
<td>0.8273 (0.0914)</td>
<td>0.4388 (0.1872)</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.8916 (0.0170)</td>
<td>0.8488 (0.0245)</td>
<td><strong>0.9024 (0.0319)</strong></td>
<td>0.8743 (0.0191)</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.8860 (0.0351)</td>
<td>0.8650 (0.0632)</td>
<td>0.8761 (0.0268)</td>
<td>0.8697 (0.0427)</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.8247 (0.0473)</td>
<td>0.8072 (0.0495)</td>
<td>0.8038 (0.0643)</td>
<td>0.8041 (0.0483)</td>
</tr>
</tbody>
</table>

Table 4.4: Performance of our proposed approach with baseline methods on a 5-fold cross validation.
Although there is no available ground-truth labels at the instance level, we manually labeled 200 tweets to evaluate the performance of our proposed framework. We randomly selected 25 false news topics and 25 true news topics from our dataset, where for each news topic four tweets were selected from different clusters. The testing data was designed in the manner of selecting tweets with potential different credibility level in each news topic.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>Rec</th>
<th>Pre</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.675</td>
<td>0.511</td>
<td>0.672</td>
<td>0.581</td>
</tr>
<tr>
<td>MI-SVM</td>
<td>0.56</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>CTP</td>
<td>0.695</td>
<td>0.932</td>
<td>0.599</td>
<td>0.729</td>
</tr>
</tbody>
</table>

Table 4.5: Performance of our proposed approach with baseline methods on instance-level credibility prediction.
Figure 4.1: ROC curve of our proposed approach with baseline methods SVM and MISVM.
As shown in Table 4.5, our proposed approach outperforms the baseline methods significantly. In particular, MISVM does not work well for this problem as it predicts every instance as negative. In order to perform a comprehensive comparison, we set different thresholds to determine positive and negative classes. Through different thresholds, we can get different sensitivities and specificities. The ROC curve of our proposed approach, labeled as CTP, with two baseline methods are shown in Fig. 4.1. The performance are compared based on the area under the curve (AUC). Our proposed approach significantly achieved a better performance with an AUC score of 0.8132 while SVM achieved an AUC score of 0.6504 and MISVM achieved an AUC score of 0.4387. The performance of MISVM is even worse than random guess as its AUC score is less than 0.5. We also noticed that our approach can achieve sensitivity greater than 0.8 and specificity greater than 0.7 simultaneously.

4.3.2 Case Studies

In addition to the performance metrics, we also provide two example topics in Fig. 4.2 to show that our proposed approach is capable to distinguish tweets with different credibility level in the same news topic. The news topic is a claim that the darkest night in 500 years will take place on 20 December 2016 due to a lunar eclipse. This topic has been verified as fake on Snopes. From our dataset, we can observe different opinions propagating in the social media and not surprisingly, some of them are spreading the fake news while some of them are clarifying the truth. Although it’s not easy for the crowd to tell which opinions are trustworthy without specific domain knowledge, our proposed approach is capable to effectively predict the credibility level at the instance level.

1https://www.snopes.com/fact-check/will-tonight-be-the-darkest-night-in-500-years/
(a) Example tweet predicted with low credibility
Shortest day of the year plus
Tonight will be the darkest night of the past 500 years io9.com/5715413/tonigh...

(b) Example tweet predicted with high credibility
Shedding some light on the subject
No, tonight won’t be the darkest night in 500 years @CNN
 cnn.it/2hrTBvf

Figure 4.2: Example tweets of different credibility level in the same news topic.
Chapter 5

Detecting Propaganda in News Media

5.1 Methodology

In this section, we present a novel non-parametric scan statistics approach to characterize the media propaganda pattern and an efficient algorithm to automatically detect such patterns in nearly linear time.

5.1.1 Pattern Analysis

If we recall the formulation of censorship pattern in Chapter 3, we are able to observe similarities between censorship pattern and propaganda pattern. Instead of detecting absenteeism in time series of daily frequency in news articles as censorship pattern, we are interested in detecting abnormally spike in time series of daily frequency in news articles as propaganda pattern. Therefore, we define that a propaganda pattern exists if, for a cluster of connected keywords:
1. Their normalized time series of daily frequency in at least one local news outlet are consistently different from their normalized time series of daily frequency in Twitter during a time period,

2. Their normalized time series of daily frequency in local news outlets are consistently well correlated to their normalized time series of daily frequency in Twitter before the time period,

3. Their normalized time series of daily frequency in at least one different local news outlet are well correlated to their normalized time series of daily frequency in Twitter during the time period.

5.1.2 Problem Formulation

Hypothesis-testing Framework

From the definition of propaganda pattern, we can detect such pattern by comparing pairs of news sources to identify convergence and divergence and concluding topics on which the news sources diverge. The proposed hypothesis-testing framework for censorship pattern detection in Chapter 3 can also be adapted to tackle this problem. Suppose we have news reports represented in an undirected keyword co-occurrence graph $G = (V, E)$ and each node $v \in V$ is associated with four attributes: time series of daily frequency of node $v$ in one news outlet represented as $\{a^t(v)\}_{t=1}^T$, expected daily frequency of node $v$ in one news outlet represented as $\lambda^a(v)$, time series of daily frequency of node $v$ in another news outlet represented as $\{b^t(v)\}_{t=1}^T$, and expected daily frequency of node $v$ in another news outlet represented as $\lambda^b(v)$. Our goal is to detect a cluster (subset) of co-occurred keywords and a time window as an indicator of partisanship, such that the distribution of frequencies of these
keywords in one news dataset is significantly different from that in the other news dataset. We consider two hypotheses: under the null \((H_0)\), the daily frequencies of each keyword \(v\) in the news datasets follow two different Poisson distributions with the mean parameters \(\lambda_a(v)\) and \(\lambda_b(v)\), respectively; under the alternative \((H_1(S, R))\), there is a connected cluster \(S\) of keywords and a continuous time window \(R \subseteq \{1, \cdots, T\}\), in which the daily frequencies of each keyword \(v\) in one news dataset follow a Poisson with an elevated mean parameter \(q_a \cdot \lambda_a(v)\), and those in the other news dataset follows a Poisson with another elevated mean parameter \(q_b \cdot \lambda_b(v)\).

Formally, they can be defined as follows:

- **Null hypothesis** \(H_0\):

  \[
  a^t(v) \sim \text{Pos}(\lambda_a(v)), \forall v \in \mathbb{V}, t \in \{1, \cdots, T\}
  \]
  \[
  b^t(v) \sim \text{Pos}(\lambda_b(v)), \forall v \in \mathbb{V}, t \in \{1, \cdots, T\}
  \]

- **Alternative hypothesis** \(H_1(S, R)\):

  \[
  a^t(v) \sim \text{Pos}(q_a \cdot \lambda_a(v)), \forall v \in \mathbb{V}, t \in \{1, \cdots, T\}
  \]
  \[
  b^t(v) \sim \text{Pos}(q_b \cdot \lambda_b(v)), \forall v \in S, t \in R
  \]
  \[
  a^t(v) \sim \text{Pos}(\lambda_a(v)), \quad b^t(v) \sim \text{Pos}(\lambda_b(v)), \forall v \notin S \text{ or } t \notin R
  \]

where \(q_a > q_b > 1\), \(S \subseteq \mathbb{V}\).

Similarly, a generalized log likelihood ratio test (GLRT) statistic can be defined to compare these two hypotheses and we can obtain the parametric form of the GLRT statistic. The problem of propaganda detection can then be reformulated as highest scoring clusters detection and statistical significance analysis as we proposed in \textsc{GraphDpd} algorithm.

However, this framework is only capable to compare time series of daily frequency from
two different data sources at one time. For each pair of local news outlets, our approach detected the connected cluster of keywords that maximizes the objective function. As a post-processing step, we aggregated indicators of propaganda patterns across all news outlets and merge similar indicators of propaganda pattern in different news outlets. However, this post-processing step might introduce noises to the results as we need to set a list of parameters on the thresholds. Hence, we are further interested in developing a novel algorithm, which can compare the time series of daily frequency from all local news articles at the same time and identify the subset of local news outlets with indicators of propaganda pattern directly.

**Non-parametric Scan Statistics**

Given an undirected keyword co-occurrence graph $G = (V, E)$, where $V = \{1, 2, \cdots, n\}$, $n$ refers to the total number of nodes, and $E \subseteq V \times V$ is a set of edges. An edge $(i, j)$ indicates the co-occurrence of keywords $i$ and $j$ in at least local news articles or Twitter posts. From 4 International Media & Newspapers, we collected a list of domain names of top newspapers for each country of interest. Each node $v$ is associated with time series of daily frequency, including time series of daily frequency in Twitter posts $\{a^t(v)\}_{t=1}^T$ and time series of daily frequency in each local newspaper $\{b^t_i(v)\}_{t=1}^T$, where $i = \{1, 2, \cdots, m\}$ and $m$ is the total number of local newspapers in the target country.

For each observation $a^t(v)$, we measure the significance of observing this value, which is also known as statistical p-value and denoted as $p^{a,t}(v)$, as the number of historical observations of daily term frequency of keyword $v$ in Twitter posts which have a higher or equal value divided by the total number of historical observations of keyword $v$ in Twitter posts. Similarly, for each observation $b^t_i(v)$, we measure the significance of observing this value, which is denoted as $p^{b,t}_i(v)$, as the number of historical observations of daily term frequency of
keywords \( v \) in local newspaper \( i \) which have a higher or equal value divided by the total number of historical observations of keyword \( v \) in local newspaper \( i \).

Non-parametric scan statistics, which measures the joint significance of multiple p-values in the subset of anomaly \( S \), has the following general form:

\[
F(S) = \phi(\alpha, N_\alpha(S), N(S)),
\]

(5.1)

where \( \alpha \) is a predefined significance level, \( N(S) \) is the size of \( S \), and \( N_\alpha(S) \) is the number of p-values in \( S \) that are significant at level \( \alpha \). In this proposal, we apply the Berk-Jones (BJ) statistic [7] as our non-parametric scan statistic and the scoring function can be reformulated as:

\[
F_{BJ}(S) = N(S) \times KL\left( \frac{N_\alpha(S)}{N(S)}, \alpha \right),
\]

(5.2)

where \( KL \) is the Kullback-Liebler divergence and is defined as:

\[
KL(a, b) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}.
\]

(5.3)

The effectiveness of BJ statistic for detecting anomalous patterns has been demonstrated in many studies [41], comparing to other non-parametric scan statistic such as the Higher Criticism. From Eqn. 5.2, we define the scoring function for anomaly in Twitter posts as:

\[
F^{Twitter}_{BJ}(S) = N(S) \times KL\left( \frac{N_\alpha(S)}{N(S)}, \beta \right),
\]

(5.4)
and the scoring function for anomaly in local newspapers as:

\[ F_{BJ}^{\text{Newspapers}}(S, F) = N(S) \cdot N(F) \times KL \left( \frac{\varphi(S, F, \alpha)}{N(S) \cdot N(F)}, \alpha \right), \]  

where \( \alpha \) is the predefined significance level for local newspapers and \( \beta \) is the predefined significance level for Twitter posts; \( S \subseteq \mathcal{V} \) refers to a subset of vertices/keywords and \( N(S) \) refers to the size of \( S \); \( F \subseteq \{1, 2, \cdots, m\} \) refers to a subset of newspapers and \( N(F) \) refers to the size of \( F \); \( \varphi(S, F, \alpha) \) refers to the number of p-values of keyword \( v \) in local newspaper \( i \) for all \( v \in S \) and \( i \in F \) that are less than or equal to \( \alpha \). Given Eqn. 5.4 and Eqn. 5.5, we define the non-parametric scan statistic as the sum of Eqn. 5.4 and Eqn. 5.5, and has the following form:

\[ F_{BJ}(S, F) = N(S) \cdot N(F) \times KL \left( \frac{\varphi(S, F, \alpha)}{N(S) \cdot N(F)}, \alpha \right) + N(S) \times KL \left( \frac{N^\alpha(S)}{N(S)}, \beta \right). \]  

Given the non-parametric scan statistics, the problem of propaganda detection can be reformulated as composed of two major components: 1) \textbf{Highest scoring clusters detection}. The highest scoring clusters are identified by maximizing the non-parametric scan statistics over all possible clusters of keywords and time windows; 2) \textbf{Statistical significance analysis}. The empirical p-values of the identified clusters are estimated via a randomization testing procedure. The clusters are significant indicators of propaganda patterns if their p-values are below a predefined significance level (e.g., 0.05).
5.1.3 GraphMP-FS Algorithm

Our proposed GRAPHDPD algorithm can be extended to detect propaganda patterns. However, the proposed RELAXED-GRAPHMP algorithm is not able to adapt to this problem as it can only maximize the scoring function over clusters of keywords subject to a connectivity constraint on graphs. Hence, we need to propose another algorithm to maximize the scoring function over clusters of keywords and sets of news outlets at the same time subject to a connectivity constraint on graphs. We propose a novel algorithm named GRAPHMP-FS to approximately solve this problem in nearly linear time with respect to the total number of nodes in the graph. We first transform the objective function in Eqn. 5.6 to a vector form. Let \( x \) be an \( n \)-dimensional vector \( (x_1, x_2, \cdots, x_n)^T \), where \( x_i \in \{0, 1\} \) and \( x_i = 1 \) if \( i \in S \), \( x_i = 0 \) otherwise. Similarly, let \( y \) be an \( m \)-dimensional vector \( (y_1, y_2, \cdots, y_m)^T \), where \( y_i \in \{0, 1\} \) and \( y_i = 1 \) if \( i \in F \), \( y_i = 0 \) otherwise. Let \( U \) be an \( n \times m \) dimensional matrix, where \( u_{i,j} \in \{0, 1\} \) and \( u_{i,j} = 1 \) if \( p_j(i) \leq \alpha \), \( u_{i,j} = 0 \) otherwise, for \( i \in \{1, 2, \cdots, n\} \) and \( j \in \{1, 2, \cdots, m\} \). Similarly, let \( W \) be an \( n \)-dimensional vector \( (w_1, w_2, \cdots, w_n)^T \), where \( w_i \in \{0, 1\} \) and \( w_i = 1 \) if \( p(i) \leq \beta \), \( w_i = 0 \) otherwise. Therefore, \( F(S, F, R) \) can be reformulated as:

\[
F_{BJ}(x, y, R) = 1^T X \cdot 1^T Y \cdot KL\left(\frac{X^T U Y}{1^T X \cdot 1^T Y}, \alpha\right) \\
+ 1^T X \cdot KL\left(\frac{W^T X}{1^T X}, \beta\right).
\]

(5.7)

Suppose we define \( P \) and \( Q \) as follows:

\[
P = \frac{x^T W}{1^T X}, \quad Q = \frac{x^T U y}{1^T X \cdot 1^T Y},
\]
the gradient of \( F_{BJ}(x, y, R) \) with respect to \( x \) has the form:

\[
\frac{\partial F}{\partial X} = 1 \cdot (1^T Y) \cdot (Q \log \frac{Q}{\alpha} + (1 - Q) \log \frac{1 - Q}{1 - \alpha}) \\
+ (1^T X \cdot 1^T Y) \cdot ((\frac{\partial Q}{\partial X} \cdot \log \frac{Q}{\alpha}) - (\frac{\partial Q}{\partial X} \cdot \log \frac{1 - Q}{1 - \alpha})) \\
+ 1 \cdot (P \log \frac{P}{\beta} + (1 - P) \log \frac{1 - P}{1 - \beta}) \\
+ (1^T X) \cdot ((\frac{\partial P}{\partial X} \cdot \log \frac{P}{\beta}) - (\frac{\partial P}{\partial X} \cdot \log \frac{1 - P}{1 - \beta})),
\]

(5.8)

and the gradient of \( F_{BJ}(x, y, R) \) with respect to \( y \) has the form:

\[
\frac{\partial F}{\partial Y} = 1 \cdot (1^T X) \cdot (Q \log \frac{Q}{\alpha} + (1 - Q) \log \frac{1 - Q}{1 - \alpha}) \\
+ (1^T X \cdot 1^T Y) \cdot ((\frac{\partial Q}{\partial Y} \cdot \log \frac{Q}{\alpha}) - (\frac{\partial Q}{\partial Y} \cdot \log \frac{1 - Q}{1 - \alpha})).
\]

(5.9)

**Algorithm 4** **GraphMP-FS**

1: **Input**: Network instance \( G \), continuous time window \( R \);
2: **Output**: The co-occurrence subgraph \( G_s \) and the subset of attributes \( F \);
3: \( i \leftarrow 0; x^i, y^i \leftarrow \) initial vectors;
4: **repeat**
5: \( g_x = H_x(\nabla_x f(x^i, y^i)) \);
6: \( \Omega_x = \text{supp}(g_x) \cup \text{supp}(x^i) \);
7: \( (b_x, b_y) = \arg \max_{x \in [0, 1]^n, y \in [0, 1]^m} \hat{f}(x, y) \quad \text{s.t.} \quad \text{supp}(x) \in \Omega_x \)
8: \( x^{i+1} = T_x(b_x); y^{i+1} = b_y \)
9: \( i \leftarrow i + 1; F \leftarrow \text{supp}(y^i), S \leftarrow \text{supp}(x^i) \)
10: **until** halting condition holds
11: **return** \( (S, F, R) \);
We propose a novel algorithm called \textsc{GraphMP-FS}, which decomposes Problem into two components: 1) maximize the object function of two variables without the constraint on connectivity; 2) apply head projection and tail projection to fulfill the constraint on connectivity \cite{29}. The brief procedure of \textsc{GraphMP-FS} is shown in Algorithm 4 and the details are summarized as follows.

- **Maximize the object function of two variables without the constraint on connectivity.** This step (Line 7) maximize the object function with respect to variable $x$ and $y$, where $x$ refers to a subset of vertices (keywords) and $y$ refers to a subset of attributes (local publishers). As we have connectivity constraint on the subset of vertices while no constraint on the subset of attributes, the object function is maximized subject to the $\text{supp}(x) \in \Omega_x$, where $\Omega_x$ is calculated as shown in Line 6. A two-step approach is applied to approximately maximize the object function since we cannot maximize the object function with respect to two variables simultaneously. The first step, as shown in Eqn. 5.10, is to maximize the object function with respect to variable $x$ using the estimator $b_y$ in Eqn. 5.11 subject to the constraint on $\text{supp}(x)$. The second step, as shown in Eqn. 5.11, is to maximize the object function with respect to variable $y$ using the estimator $b_x$ in Eqn. 5.10. The initial values of $b_x$ and $b_y$ are set to the values of $x$ and $y$ after Line 6. Eqn. 5.10 and Eqn. 5.11 are typical maximization problems of one variable, which are solved using a gradient ascent optimization algorithm. Eqn. 5.10 and Eqn. 5.11 are repeated until the pair value of $(b_x, b_y)$ converges, where it usually converges within five iterations in our experiments.

$$
    b_x = \arg \max_{x \in [0,1]^n} \hat{f}(x, b_y) s.t. \text{supp}(x) \in \Omega_x
$$

(5.10)
$$b_y = \arg \max_{y \in [0,1]^m} \hat{f}(b_x, y)$$ (5.11)

• Apply head projection and tail projection to fulfill the constraint on connectivity. Line 5 applies head projection on $x$ and Line 8 applies tail projection on $x$ to guarantee the connectivity constraint on $x$. Because we do not have any constraint on $y$, we do not need to do similar things to $y$.

• Halting.

5.2 Experiments

In this section, we evaluated the performance of our proposed approach in propaganda pattern detection compared with alternative methods on real-world data. The results of these experiments show our proposed algorithm for detecting likely patterns of media propaganda outperforms other state-of-the-art options.

5.2.1 Data

Datasets: In this work, we evaluate the performance of our proposed approach with real-world Twitter and news articles datasets of Mexico and Venezuela from Jan 1, 2014 to Dec 31, 2014. The Twitter posts are collected as 10% random samples and the news articles are collected in the EMBERS project, which are sourced from around 6,000 news agencies. For each country of interest, we retrieved the most popular newspapers from "4 International Media & Newspapers" http://www.4imn.com/.

Post-processing: For every continuous time window from 3 to 7 days starting from Jan 1, 2014 to Dec 31, 2014, our approach is capable to detect a connected cluster of keywords
and a set of news outlets with the highest score according to the scoring function. Since we do not have ground truth about whether a score is significant or not, we perform 5,000 random permutations and record the function scores of each optimal connected cluster of keywords and sets of news outlets that maximizes the scoring function. Every detection of a connected cluster of keywords and a set of news outlets is significant only if its p-value is less than a predefined significance level (0.05). The overlapped time windows are eliminated by merging connected clusters of keywords and sets of news outlets within 5 days. Each distinct combination of connected cluster of keywords, set of news outlets, and associated time window can be treated as an indicator of propaganda pattern. If an indicator of propaganda pattern is detected in all newspapers in the country, it’s eliminated as it might be caused by different topics of interest in various data sources.

5.2.2 Evaluation

We compared the performance of our proposed approach with three state-of-the-art methods for anomaly detection: LTSS, NPHGS, and EventTree. Although each baseline method has its own scoring function to detect the optimal indicator, the baseline methods are not able to determine whether the indicator is significant or not. In order to compare fairly, we remove duplicate indicators falling in overlapping time windows in the same way as the post-processing of our approach and the same number of indicators with top function scores from each country of interest are selected to compare with our approach.

Table 5.1 summarizes the false positive rates in detecting indicators of propaganda pattern between our proposed approach GraphMP-FS and three baseline methods: LTSS, NPHGS, and EventTree. It is not surprisingly that the baseline methods are not doing well in propaganda detection as they are formulated for event detection. In another word, the baseline
methods are likely to capture various abnormal patterns in News and hard to tell whether an anomaly is real propaganda or normal spike of occurring event.

<table>
<thead>
<tr>
<th>Country</th>
<th>LTSS</th>
<th>NPHGS</th>
<th>EventTree</th>
<th>GRAPHMP-FS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>0.783</td>
<td>0.696</td>
<td>0.609</td>
<td>0.304</td>
</tr>
<tr>
<td>Venezuela</td>
<td>0.765</td>
<td>0.824</td>
<td>0.706</td>
<td>0.412</td>
</tr>
</tbody>
</table>

Table 5.1: Comparison of false positive rates in propaganda detection between GRAPHMP-FS and three baseline methods: LTSS, NPHGS, and EventTree on real data of Mexico and Venezuela during year 2014.
Chapter 6

Conclusion and Future Work

In this chapter, we conclude the problems that have been explored in this dissertation and propose the future directions of research in this domain.

6.1 Conclusion

The proposed research aims to detect three types of anomalous information patterns in social media and news media. The first anomalous information pattern is self-censorship in news media, where the information is missing. The second anomalous information pattern is fake news in social and news media, where low credibility information are propagated over the media network. The third anomalous information pattern is propaganda in news media, where news sources converge and diverge on certain news topics.

In this dissertation, we have presented a novel unsupervised approach to identify censorship patterns in domestic news media using social media as a sensor. This approach has demonstrated great promise in detecting self-censorship as compared to current event-detection
technologies. It also has demonstrated utility for providing an assessment of freedom of the press in countries with active social media populations, and for understanding the patterns of self-censorship within countries.

This method may also have applications in other areas. Indeed, the GraphDPD methodology should be useful for many applications where text is compared between sources over time. In future work, we intend to explore these themes further, including developing strategies for forecasting censorship.

In this dissertation, we have presented a cluster-based approach to predict the credibility of information at the instance level. This approach has demonstrated it capability in predicting credibility as compared to traditional classification methods. It has also demonstrated utility for understanding how the information become viral in the propagation process. In future work, we intend to improve the clustering process and also, possibly design a strategy to further dig into different credibility level inside the same cluster. We can also extend this work to other social media platforms to explore additional information for credibility analysis.

6.2 Future Work

Here we discuss future work that could be explored on these problems later.

6.2.1 Media Censorship Forecasting

In our current work, we propose a hypothesis-testing framework to detect censorship pattern in news media. A future direction on this topic is trying to predict occurrences of media
censorship by leveraging information pattern in time series.

### 6.2.2 Predicting Information Credibility in News Media

In our current work, we propose a cluster-based approach to predict information credibility at the instance level in social media. We notice that existing studies on misinformation detection in news media mostly focus on the article level, which utilized features extracted from headlines and news contents to predict credibility of news articles. However, information credibility prediction at the sentence level has received not as much attention mainly due to the lack of features can be extracted at the sentence level.

This problem is of interest as sentences talking about different claims or opinions can have quite different credibility level. For example, Snopes\(^1\) reported a fake news on a claim that Corona beer founder Antonino Fernandez made everyone in his village a millionaire after his death. Claims of death of the founder and millionaire can be reported in the same news article. Even if the news article is identified as a fake news, readers can still be confused about whether death of the founder is trustworthy. Hence, a future direction of this work is to develop a novel framework to predict information credibility at the sentence level in news articles.

Existing studies on misinformation detection in online news media viewed this problem as a binary classification problem and extracted features from headlines and news contents [9, 61]. However, such approach does not work well at the sentence level prediction due to the lack of useful features. Another approach for this issue is to determine the truthfulness of a statement with knowledge graphs [14, 64].

A statement of fact can be represented in a (subject, predicate, object) triple, for example, (“Obama”, ”is”, ’president’). A knowledge graph (KG) is built upon a set of such triples,

---

\(^1\)https://www.snopes.com/fact-check/corona-founder-made-every-person-in-his-village-a-millionaire/
where nodes represent entities and edges represent predicates. Given a new statement, we can predict its credibility through checking whether it exists as an edge of the knowledge graph or if we can find a path linking its subject to its object within the knowledge graph. However, this approach cannot be generalized to upcoming news topics as the knowledge graph is normally built upon the DBpedia database, which consists of factual statements extracted from Wikipedia’s infoboxes. Although the database includes a large amount of relations, the type of relations are still limited and trivial. Furthermore, it is not easy for the database to catch up with the fast-growing information.

One direction of future work is to predict information credibility at the sentence level in news media through a new type of knowledge graph, built upon relations extracted from Twitter posts. Given the tweets associated with credibility level, we can extract relations from the tweets using open information extraction (openIE) provided by Stanford CoreNLP. In case of the same relation extracted from multiple tweets with different credibility prediction, it is necessary to formulate an aggregate function to assign a single credibility prediction to the relation.

Given a sentence in a news article, its credibility level is aggregated among the credibility level of each relation extracted from the sentence. The truthfulness of a relation is estimated through checking whether it exists as an edge of the knowledge graph or if we can find a path linking its subject to its object within the knowledge graph. Under the simple situation, the credibility of a relation is equivalent to the credibility of an edge if it exists as an edge of the knowledge graph, which means we can find an exact match of the relation in the social media. In case of a more complicated situation where we cannot find an exact match, we need to develop a novel algorithm to find a path linking subject to object within the knowledge graph and estimate the credibility level.

Finally, we can explore the usage of deep learning and GPU [17, 18] to accelerate the model.

training and improve the training performance.

6.2.3 Study of Social Bots

In our current work, we propose a non-parametric scan statistic approach to detect propaganda pattern in news media, which discover topics where different news outlets diverge. In addition to detecting propaganda in news media, a future direction is to study propaganda in social media. One potential reason of propaganda in social media is amplification and dampening of social bots, which could be tackled by comparing tweets posted by social bots with tweets posted by human users to identify (1) amplification by social bots: topics which news and general tweets have normal spike while tweets posted by social bots have an abnormally strong pike, and (2) dampening by social bots: topics which news and general tweets have normal spike while tweets posted by social bots are silent.

One important issue in this problem is to distinguish social bot users from human users in social media. [20] has introduced state-of-art algorithms in identifying social bots on Twitter. In particular, BotOrNot developed a publicly available service with more than 1,000 features to evaluate whether a Twitter account behaves similarly to social bots. The features exposed through this dataset can be fruitfully combined with our methods to develop more advanced techniques.
Bibliography


