Detecting Media Self-Censorship without Explicit Training Data

Chapter · January 2020

DOI: 10.1137/1.9781611976236.62

7 authors, including:

Baojian Zhou
State University of New York
19 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Feng Chen
University at Albany, The State University of New York
68 PUBLICATIONS **833** CITATIONS
SEE PROFILE

David Mares
University of California, San Diego
62 PUBLICATIONS **643** CITATIONS
SEE PROFILE

Patrick Butler
University of Southern Denmark
27 PUBLICATIONS **490** CITATIONS
SEE PROFILE

Detecting Media Self-Censorship without Explicit Training Data

Rongrong Tao * Baojian Zhou † Feng Chen ‡ David Mares § Patrick Butler *
Naren Ramakrishnan * Ryan Kennedy ¶

Abstract

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 to identify news media self-censorship by using social media as a sensor. 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, versus other state-of-the-art algorithms, using both semi-synthetic and real-world data from Mexico and Venezuela during Year 2014. 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 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 [1].

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 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 paper, 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 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 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

[1]Virginia Tech, VA, USA {rrtao@vt.edu, pabutler@vt.edu, naren@cs.vt.edu}
[2]University at Albany, SUNY, Albany, NY, USA {bzhou6@albany.edu}
[3]University of Texas at Dallas, Richardson, TX, USA {feng.chen@utdallas.edu}
[4]University of California at San Diego, San Diego, CA, USA {dmares@ucsd.edu}
[5]University of Houston, Houston, TX, USA {rkennedy@uh.edu}
[1]https://freedomhouse.org/event/new-challenges-freedom-expression-latin-america

Table 1: Summary of Twitter Transparency Report for Year 2014 on selected countries.

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

to detect news self censorship events in Latin America. The main contributions of this paper 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 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 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.

2 Related Work

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 [3] 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 [14] 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. [16], 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. [12] presented the first open-domain system for event extraction and an approach to classify events based on latent variable models. [13] 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, [4] 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. [10] 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, [9] observed both similarities and differences of coverage of events between social media and news media.

3 Data Analysis

Table 3 summarizes the notation used in this work. The EMBERS project [11] provided a collection of Latin American news articles and Twitter posts. The news dataset was sourced from around 6000 news agencies during 2014 across the world. From "4 International Media & Newspapers", we retrieved a list of top news-

papers 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 1) but severe level of censorship in news media according to the Freedom of the Press Report.

3.1 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 $\Gamma = 10$ 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 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 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 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. 1 compares TSDF in El Mexicano Gran Diario Regional (elmexicano.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. 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. 1a to Fig. 1d are normalized to [0, 500] for visualization. Fig. 1a to Fig. 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. 1a to Fig. 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 2: Different patterns of co-occurrence observed between social media and news media sources.

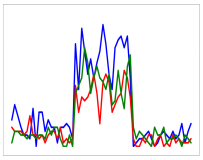
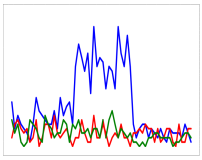
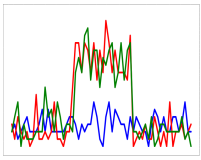
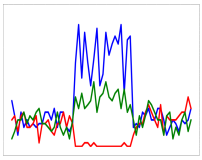
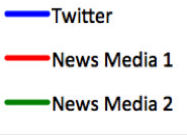
Topic is of interest in both social media and news media.	Topic is of interest in social media but not in news media.	Topic is of interest in news media but not in social media.	Censorship in one news media source.	
				
Example: In early March 2014, Malaysia Airlines Flight MH 370 went missing.	Example: 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.	

Table 3: Description of major notation.

Variable	Meaning
$\{a^t(v)\}_{t=1}^T$	time series of daily frequency of node v in uncensored Twitter dataset
$\lambda_a(v)$	expected daily frequency of node v in the Twitter dataset.
$\{b^t(v)\}_{t=1}^T$	time series of daily frequency of node v in the censored news dataset
$\lambda_b(v)$	expected daily frequency of node v in data source b
TSDF	time series of daily frequency

4 Methodology

4.1 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 $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, where $\mathbb{V} = \{1, 2, \cdots, n\}$ refers to the ground set of nodes/keywords, n refers to the total number of nodes, and $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ 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 $v \in \mathbb{V}$ 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. 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:

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

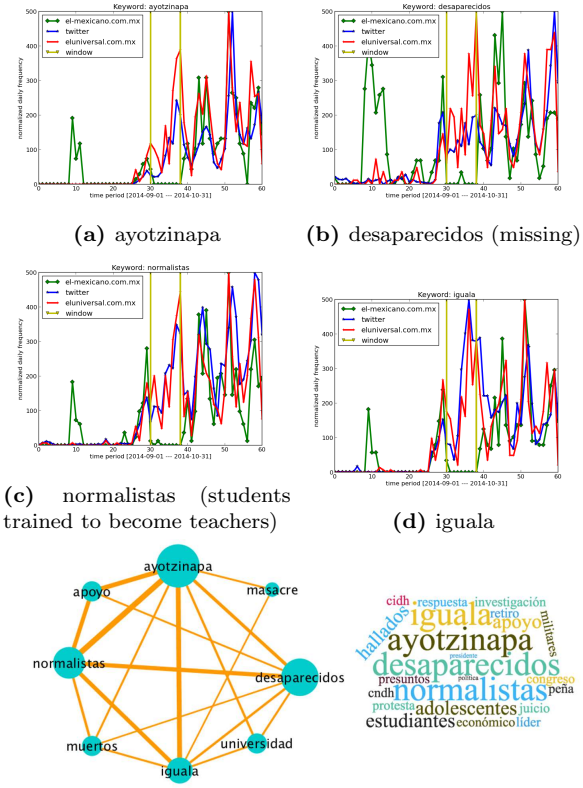
- Alternative hypothesis $H_1(S, R)$:

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

where $q_a > 1, q_b < 1$, $S \subseteq \mathbb{V}$, the subgraph induced by S (denoted as $\mathbb{G}_S$) 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 $\{i, i+1, \cdots, j\}, 1 \le i \le j \le T$. 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:

$$F(S, R) = \log \frac{\max_{q_a > 1} \prod_{t \in R} \prod_{v \in S} p(a^t(v); q_a \lambda_a(v))}{\prod_{t \in R} \prod_{v \in S} p(a^t(v); \lambda_a(v))}$$
$$+ \log \frac{\max_{q_b < 1} \prod_{t \in R} \prod_{v \in S} p(b^t(v); q_b \lambda_b(v))}{\prod_{t \in R} \prod_{v \in S} p(b^t(v); \lambda_b(v))}. \tag{4.1}$$

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



(a) ayotzinapa **(b)** desaparecidos (missing)

(c) normalistas (students trained to become teachers) **(d)** iguala

(e) Left: The strong connectivity of these keywords indicates their frequent co-occurrence in *Twitter* and *News*. A larger size of node indicates higher keyword frequency and a larger width of edge indicates more frequently co-occurrence; Right : word cloud representing censored keywords in *News* around 2014-09-26 in Mexico

Figure 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.

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:

(4.2)
$$F(S,R) = \left(C_a \log \frac{C_a}{B_a} + B_a - C_a\right) + \left(C_b \log \frac{C_b}{B_b} + B_b - C_b\right)$$

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 [7], 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 $\mathbb{G}(\mathbb{V},\mathbb{E})$ and a predefined significance level α, the* GLRT *optimization problem is to find the set of highest scoring and significant clusters $\mathbb{O}$. Each cluster in $\mathbb{O}$ is denoted as a specific pair of connected subset of keywords ($S_i \subseteq \mathbb{V}$) 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:*

(4.3) $$\max_{S\subseteq \mathbb{V}} F(S,R_i) \text{ s.t. } S \text{ is connected,}$$

and is significant w.r.t the significance level α.

4.2 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 4.1. 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-GRAPMP 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 [7](Line 8). The pair (S,R) will be added into the result set $\mathbb{O}$ (Line 9) if its empirical p-value is less than a predefined significance level α (e.g., 0.05). The procedure getPValue in Line 8 refers to a randomization testing procedure based on the input graph $\mathbb{G}$ to calculate the empirical p-value of the pair (S,R) [7]. Finally, we return the set $\mathbb{O}$ of signifiant clusters as indicators of censorship events in the news data set.

ALGORITHM 4.1. (GRAPHDPD)
1: **Input**: Graph Instance $\mathbb{G}$ and significant level α;
2: **Output**: set of anomalous connected subgraphs $\mathbb{O}$;
3: $\mathbb{O} \leftarrow \emptyset$;
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: $S \leftarrow$ RELAXED-GRAPHMP$(\mathbb{G}, R)$;
8: **if** getPValue$(\mathbb{G}, S, R) \le \alpha$ **then**
9: $\mathbb{O} \leftarrow \mathbb{O} \cup (S,R)$;
10: **end if**
11: **end for**
12: **end for**

13: **return** $\mathbb{O}$;

Line 7 in Algorithm 4.1 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. [6] 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 in Theorem 4.1 and this approach is not applicable here.

THEOREM 4.1. *[] Given a specific window R, our objective function $F(S,R)$ defined in (4.2) 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(4.2) to a vector form. Let $\mathbf{x}$ be an n-dimensional vector $(x_1, x_2, \cdots, x_n)^\mathsf{T}$, 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:

$$\mathcal{P} = \left[\sum_{t\in R} a^t(1), \cdots, \sum_{t\in R} a^t(n)\right]^\mathsf{T}, \Lambda_a = [\lambda_a(1), \cdots, \lambda_a(n)]^\mathsf{T},$$

$$\mathcal{Q} = \left[\sum_{t\in R} b^t(1), \cdots, \sum_{t\in R} b^t(n)\right]^\mathsf{T}, \Lambda_b = [\lambda_b(1), \cdots, \lambda_b(n)]^\mathsf{T}.$$

Therefore, C_a, C_b, B_a, and B_b in Equation(4.2) can be reformulated as follows:

$$C_a = \mathcal{P}^\mathsf{T}\mathbf{x}, \quad C_b = \mathcal{Q}^\mathsf{T}\mathbf{x}, \quad B_a = |R|{\Lambda_a}^\mathsf{T}\mathbf{x}, \quad B_b = |R|{\Lambda_b}^\mathsf{T}\mathbf{x}$$

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

$$\hat{F}(\mathbf{x},R) = \mathcal{P}^\mathsf{T}\mathbf{x}\log\frac{\mathcal{P}^\mathsf{T}\mathbf{x}}{|R|{\Lambda_a}^\mathsf{T}\mathbf{x}} + |R|{\Lambda_a}^\mathsf{T}\mathbf{x} - \mathcal{P}^\mathsf{T}\mathbf{x} + \mathcal{Q}^\mathsf{T}\mathbf{x}\log\frac{\mathcal{Q}^\mathsf{T}\mathbf{x}}{|R|{\Lambda_b}^\mathsf{T}\mathbf{x}} + |R|{\Lambda_b}^\mathsf{T}\mathbf{x} - \mathcal{Q}^\mathsf{T}\mathbf{x} \tag{4.4}$$

We relax the discrete domain $\{0,1\}^n$ of S to the continuous domain $[0,1]^n$ of $\mathbf{x}$, 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 $\mathbb{O}$, the subset of keywords S_i is identified by solving the following problem with $S_i = supp(\hat{\mathbf{x}})$:*

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

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

$$\frac{\partial\hat{F}(\mathbf{x},R)}{\partial\mathbf{x}} = \log\frac{\mathcal{P}^\mathsf{T}\mathbf{x}}{|R|{\Lambda_a}^\mathsf{T}\mathbf{x}}\mathcal{P} + \Big(|R| - \frac{\mathcal{P}^\mathsf{T}\mathbf{x}}{{\Lambda_a}^\mathsf{T}\mathbf{x}}\Big)\Lambda_a + \log\frac{\mathcal{Q}^\mathsf{T}\mathbf{x}}{|R|{\Lambda_b}^\mathsf{T}\mathbf{x}}\mathcal{Q} + \Big(|R| - \frac{\mathcal{Q}^\mathsf{T}\mathbf{x}}{{\Lambda_b}^\mathsf{T}\mathbf{x}}\Big)\Lambda_b \tag{4.5}$$

ALGORITHM 4.2. (RELAXED-GRAPHMP) 1: **Input**: Graph instance $\mathbb{G}$, continous time window R;
2: **Output**: the co-occurrence subgraph $\mathbb{G}_S$;
3: $i \leftarrow 0$; $\mathbf{x}^i \leftarrow$ an initial vector;
4: **repeat**
5: $\quad\nabla\hat{F}(\mathbf{x}^i, R) \leftarrow \frac{\partial\hat{F}(\mathbf{x}^i,R)}{\partial\mathbf{x}^i}$ by Equation (4.5) ;
6: $\quad\mathbf{g} \leftarrow$ **Head**$(\nabla\hat{F}(\mathbf{x}^i, R), \mathbb{G})$; // *Head projection step*
7: $\quad\Omega \leftarrow \text{supp}(\mathbf{g}) \cup \text{supp}(\mathbf{x}^i)$;
8: $\quad\mathbf{b} \leftarrow \arg\max_{\mathbf{x}\in[0,1]^n} \hat{F}(\mathbf{x}, R)$ *s.t.* $\text{supp}(\mathbf{x}) \subseteq \Omega$;
9: $\quad\mathbf{x}^{i+1} \leftarrow$ **Tail**$(\mathbf{b}, \mathbb{G})$; // *Tail projection step*
10: $\quad i \leftarrow i+1, S \leftarrow \text{supp}(\mathbf{x}^i)$;
11: **until** halting condition holds;
12: **return** (S, R);

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 [5] subject to connectivity constraints. We call our method RELAXED-GRAPHMP which is analogous to GRAPHMP proposed by Chen et al. [2]. The high level of RELAXED-GRAPHMP is shown in Algorithm 4.2. It contains 4 main steps as described below.

- **Step 1:** Compute the gradient of relaxed GLRT problem (Line 5). The calculated gradient is $\nabla\hat{F}(\mathbf{x}^i, R)$. 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 $\mathbf{g}$ so that the corresponding subset $\text{supp}(\mathbf{g})$ can maximize the norm of the projection of gradient $\nabla\hat{F}(\mathbf{x}^i, R)$ (See details in [5]).
- **Step 3:** Solve the maximization problem without connectivity constraint. This step (Line 7,8) solves the maximization problem subject to the $\text{supp}(\mathbf{x}) \subseteq \Omega$, where Ω is the union of the support of the previous solution $\text{supp}(\mathbf{x}^i)$ with the result of head projection $\text{supp}(\mathbf{g})$ (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 $\mathbb{G}_S$ so that $\mathbf{b}_S$ is

close to **b** but with connectivity constraint. This tail projection guarantees to find a subgraph $\mathbb{G}_S$ with constant approximation guarantee (See details in [5]).

- **Halting:** The algorithm terminates when the condition holds. Our algorithm returns a connected subgraph $\mathbb{G}_S$ where the connectivity of $\mathbb{G}_S$ is guaranteed by Step 4.

Time Complexity Analysis: The GraphDPD algorithm is efficient as its time complexity is proportional to the total number of continous 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 4.2.

Theorem 4.2. *[]GraphDPD runs in $O(T^2 \cdot t(nT + nl + |\mathbb{E}|\log^3 n))$ time, where T is the maximal time window size, nT is the time complexity of Line 5 in Algorithm 4.1, nl is the run time of Line 8 using gradient ascent, $|\mathbb{E}|\log^3 n$ is the total run time of head projection and tail projection algorithms, and t is the total number of iterations needed in Algorithm 4.2.*

5 Experiments

Through experiments conducted on semi-synthetic data and real data, we evaluate 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.

5.1 Experimental Design Real world datasets:

Table 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 3.1.

Table 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.

Country	Daily Tweets	Daily News Articles	# of News Outlets
Mexico	84556	444	13
Venezuela	65916	91	6

Data Preprocessing: The preprocessing of the real world datasets has been discussed in detail in Section 3.1. 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 [15].

Our proposed Graph-DPD and baseline methods: We compare our proposed method with one baseline method LTSS [8], 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 [13] and NPHGS [1].

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:

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

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.

5.2 Semi-synthetic Data Evaluation

We evaluate the accuracy of our approach to detect the disrupted ground truth anomalies. We find that overall our approach consistently outperforms all other baseline methods. Here are the major findings: (1) Our approach

[2] http://www.nybooks.com/daily/2014/04/09/venezuela-protests-censorship/

[3] https://www.pri.org/stories/2014-03-27/venezuela-protests-shed-light-extent-media-censorship

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. (2) The recall of NPHGS becomes quite low when true subgraph is relatively large. (3) The recall of EventTree is among the best when perturbation intensity is low. (4) LTSS did well in average recall but poorly in average precision as the size of anomalous graph increases.

5.3 Real Data Evaluation As before, we apply three anomaly detection baseline methods, LTSS, NPHGS, and EventTree, 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 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.

5.4 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 5 summarizes a list of example instances of censorship identified by our approach in Mexico and Venezuela with significance level ≤ 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. 2a 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).

(a) Mexico 2014-05-01 **(b)** Venezuela 2014-02-18

Figure 2: Word cloud representing censored keywords in *News* identified by our method

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 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. 2b 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).

6 Conclusion

In this paper, 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.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract num-

Table 5: Example indicators of censorship identified by our approach in Mexico and Venezuela during Year 2014 (with significance level ≤ 0.05)

Mexico			
Date	Example censored keywords	Example local news media detected with censorship	Reasons for censorship in news media
2014-05-01	reforma(reform), gasolina(petrol), educación(education)	Noroeste	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.
2014-09-27	ayotzinapa, iguala, normalistas, desaparecidos(missing), detenidos(detained), protesta(protest)	El Mexicano Gran Diario Regional	43 students from the Ayotzinapa Rural Teachers' College went missing and kidnapped in Iguala on September 26, 2014.
2014-11-10	ayotzinapa, estudiantes(students), normalistas, desaparecidos(missing), protesta(protest), militares(military), iguala	El Mexicano Gran Diario Regional	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.
Venezuela			
2014-02-18	represión(repression), disparó(shooting), marchamos(march), heridos(wounded), nicolasmaduro, armados(armed), leopoldolopez, apresar(arrest)	Ultimas Notícias	Mass protests led by opposition leaders, including Leopoldo López, occurred in 38 cities across Venezuela asking for the release of the arrested students.
2014-05-01	muertes(deaths), cambio(change), caracas, presidente(president), labor	El Tiempo in Anzoategui	Thousands of Venezuelans demonstrated in Caracas to commemorate Labor Day and denounce shortages.
2014-08-12	gubernamental(government), anticontrabando, contrabando, ébola, muerte(death)	El Nacional	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 .

Table 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.

Country	LTSS	NPHGS	EventTree	GraphDPD
Mexico	0.722	0.667	0.556	0.278
Venezuela	0.714	0.786	0.643	0.357

ber D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proc. KDD*, pages 1166–1175. ACM, 2014.

[2] F. Chen and B. Zhou. A generalized matching pursuit approach for graph-structured sparsity. In *Proc. IJCAI*, pages 1389–1395, 2016.

[3] S. Das and A. Kramer. Self-censorship on facebook. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[4] Y. Gao, S. Zhao, Y. Yang, and T.-S. Chua. Multimedia social event detection in microblog. In *MMM*, pages 269–281. Springer, 2015.

[5] C. Hegde, P. Indyk, and L. Schmidt. A nearly-linear time framework for graph-structured sparsity. In *Proc. ICML*, pages 928–937, 2015.

[6] T.-W. Kuo, K. C.-J. Lin, and M.-J. Tsai. Maximizing submodular set function with connectivity constraint: Theory and application to networks. *IEEE/ACM Transactions on Networking (TON)*, 23(2):533–546, 2015.

[7] D. B. Neill. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, 8(1):20, 2009.

[8] D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

[9] A. Olteanu, C. Castillo, N. Diakopoulos, and K. Aberer. Comparing events coverage in online news and social media: The case of climate change. *ICWSM*, 15:288–297, 2015.

[10] S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *ICWSM*, 2013.

[11] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proc. KDD*, pages 1799–1808. ACM, 2014.

[12] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proc. KDD*, pages 1104–1112. ACM, 2012.

[13] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *Proc. KDD*, pages 1176–1185. ACM, 2014.

[14] M. Sleeper, R. Balebako, S. Das, A. L. McConahy, J. Wiese, and L. F. Cranor. The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 793–802, 2013.

[15] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *Proc. ICDM*, pages 613–622. IEEE Computer Society, 2006.

[16] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proc. CIKM*, pages 2541–2544. ACM, 2011.