Mass Movements and their Adoption in Social Networks

Fang Jin

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

Naren Ramakrishnan, Chair
Chang-Tien Lu
Chris North
Yang Cao
Feng Chen

2016 summer
Arlington, Virginia

Keywords: Information Propagation, Group Anomaly, Mass Movement, Event Detection, Social Networks

Mass Movements and their Adoption in Social Networks

Fang Jin

(ABSTRACT)

Online social networks have become a staging ground for modern movements, with the Arab Spring being the most prominent example. In an effort to understand and predict those movements, social media is regarded as valuable social sensor to disclose the underlying behavior and pattern. To fully understand the mass movement information propagation pattern in social networks, several problems need to be considered and addressed. Specifically, modeling mass movements that incorporate (i) multiple spaces (ii) dynamic network structure (iii) misinformation and (iv) swift outbreak/slowly evolving transmission would be highly propitious in understanding information propagation in social medias.

This dissertation explores four research problems underlying mass movement adoption in social media. First, how do mass movements get mobilized on Twitter, especially in a specific geographic area. Second, how do we detect protest activity in social networks by observing group anomaly in graph? Third, how do we distinguish real movements from rumors or misinformation campaigns? Fourth, how can we infer the indicators of a specific type of protest, say climate related protest?

A fundamental objective of this research has been to comprehensively study the mass movement adoption in social networks, it may cross multiple spaces, it may evolve with dynamic network structures, it can be swift outbreaks or long term slowly evolving transmissions, what is more, it may mixed with misinformation campaigns. Each of those issues requires the development of new mathematical models and algorithmic approaches which are explored here. It is my hope that this work will facilitate advancements in information propagation, group anomaly detection and misinformation distinction, and ultimately helps improve the understanding of mass movement and their adoptions in social networks.

Acknowledgments

Over the past four and half years I have received support and encouragement from a great number of individuals.

My deepest gratitude is to my advisor, Dr. Naren Ramakrishnan. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time pick me back when I got lost. His far-sighted research attitude and respectable personality lead Discovery Analytics Center a collaborative and productive place. His insight, wisdom and humour make my graduate study a rich and rewarding journey, which I will cherish forever.

I am grateful to Dr. Chang-Tien Lu, one of the best teachers that I have had in my life. He introduced data analytics and open a new window to me. He provided infinite support and encouragement while I was looking for an academic job.

I would like to thank Dr. Feng Chen, a mentor and a friend, for his support over the past several years as I moved from an idea to a completed study. I am deeply grateful to him for the countless discussions that helped me sort out the technical details of my paper.

I appreciate the efforts of Dr. Yang Cao, for his encouragement and practical advice in inspiring me to work on chapter **??**. I am also thankful to him sparing no effort in supporting my job hunting and I deeply appreciate his belief in me.

I would like to say thanks to Dr. Chris North, for his support, feedback, and valuable discussions that helped me understand my research area better. It has always been pleasant to have such a knowledgeable and amiable professor around.

I would like to thank Dr. Huzefa Rangwala, who is always generous to give advice and provide help. I hope one day I would become a good advisor as Dr. Rangwala has been to me.

I am also indebted to the members of Discovery Analytics Center, thanks for making our lab such a warm and joyful family.

I would like to thank my collaborator Edward Dougherty, one of the best collaborators ever. His enthusiasm and efficiency has always inspired me.

Most importantly, I would like to thank my parents for their constant source of strength. I am so grateful to my husband for his care, sacrifice, and love. I would like to say thanks to him for walking with me through hardship, challenge and setbacks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social microblogs such as Twitter and Weibo are experiencing explosive growth with billions of users around the globe sharing their daily status updates online. For example, Twitter has more than 255 million average monthly active users (78% from mobile) per month as of March 31, 2014, and an estimated growth of 25% per year. In the technology era, online social networks have become a staging ground for modern movements with the Arab Spring being the most prominent example. Interestingly, the role of social networks is not limited to helping organize the activities of disruptive elements. Many key government and news agencies have also begun to embrace Twitter and other social platforms to disseminate information. Without doubt, the analysis of social media networks has become a crucial and irreplaceable task in understanding the social movements.

Social network analysis is the process of gathering data from stakeholder conversations on digital media, and processing into structured insights. These lead to more information-driven decisions, which include but are not limited to understanding social sentiment, discovering topics, identifying ongoing events, and predicting future trends. Social media as a carrier of information, despite its various forms (Facebook, Twitter, Weibo, etc.) shares some common properties in information propagation, that can be approached using the methods of mathematical modeling and data mining.

As social media gains popularity, more and more mass movements are organized by social media, their slogans have become hashtags in Twitter. Here the mass movements more refer to social movements, which are supported by large group of individuals or organizations, focusing on specific political or social issues. In this dissertation, we intended to study the mass movements adoption patterns and other subsequent phenomenons from social media.

1.1 Motivation

This dissertation explores four research problems underlying mass movement adoption in social media. In contrast to popular memes, they constitute modelling protest mobilization, detect graph group anomaly pattern, infer protest causality, and distinguish real movement from rumors. (i) First, how do mass movements get mobilized on Twitter, especially in a specific geographic area, (ii) Second, how do we detect protest activity in social networks by observing group anomaly in graph? (iii) Third, how do we distinguish real movements from rumors or misinformation campaigns? and (iv) Fourth, how can we infer the pathways of climate related protests?

Modeling mobilizations: It is well known that network structure plays a key role in information propagation. Several interesting questions arise in this space. Which node is the key player who exerts influence over others? How do we realistically simulate information propagation process within a network? How do specific memes get adopted in the network? When do they translate into mass movements?

Group anomaly in graph: Group anomaly not only depends on each user's activity, but also closely associates with the graph structure. In recent year, a significant body of research on group anomaly has been focused on two aspects: (1) modeling users behaviors to define the group anomaly, but fail to pay attention to the underlying network structure; (2) define the group in local scale with distance-based restrictions such as distance, radius, or even nodes numbers, but fail to consider in the global perspective, as nodes with far distance could be highly associated. We pay attention to the global level group anomaly, without setting any restriction to the group definition, consider both the users' behavior and the underlying graph structure. Investigating this phenomenon of broad group anomaly behavior online holds enormous potential for understanding large-scale, disruptive societal events, such as mass movements.

Climate related protest pathways: The occurrence of either a shift in climate, extreme weather, or environmental catastrophe is not sufficient to guarantee that civil unrest is likely to follow. In general the causal mechanisms leading to civil unrest are very complex, and there is no easy way to determine a linear pathway to protest. What is climate related protest evolution pattern, thus how does the climate disasters lead to armed protests? What is the coherent correlations among the climate protests?

Misinformation campaigns propagation: As millions of users post various messages every second, every one of them is a potential information source, resulting in multiple propagation paths, mixed messages, innuendos, falsehoods, and rumors. How do we track the spread of rumors and misinformation campaigns and can we distinguish them from 'regular' or normal propagation patterns? Can we distinguish real movement from rumors or misinformation campaigns?

1.2 Methods

Here we present an overview of the methods used in this dissertation. They will serve as he foundation to the key new information diffusion models proposed in Chapters ?? - ??.

1.2.1 Geometric Brownian Motion

Brownian motion is the random motion of particles suspended in a fluid (a liquid or a gas) resulting from their collision with the quick atoms or molecules in the gas or liquid. This term can also refer to the mathematical model used to describe such random movements, which is often called a particle theory [19].

Geometric Brownian motion is a continuous-time stochastic process in which the logarithm of the randomly varying quantity follows a Brownian motion (also called a Wiener process) with drift. It is an important example of stochastic processes satisfying a stochastic differential equation (SDE); in particular, it is used in mathematical finance to model stock prices (such as the price of a stock over time), subject to random noise.

A stochastic process S_t is said to follow a geometric Brownian motion if it satisfies the following stochastic differential equation:

$$dS_t = \mu S_t dt + \delta S_t dW_t$$

we call W_t as a Wiener process (Brownian motion) and μ the drift, δ the volatility.

Consider a Brownian motion trajectory that satisfies the differential equation, $\mu S_t dt$ controls the 'trend' of this trajectory and the term $\delta S_t W_t$ controls the 'random noise' effect in the trajectory. The analytical solution of this geometric brownian motion is given by

$$S_t = S_0 exp((\mu - \frac{\delta^2}{2})t + \delta W_t)$$

According to the GBM properties, $ln(S_t^{ij})$ is a Gaussian variable given by:

$$ln(S_t^{ij}) \sim \mathcal{N}((\mu - \frac{\sigma^2}{2})t, \sigma^2 t)$$

1.2.2 Graph Wavelets

Graph wavelets are a form of graphical models bringing three kinds of benefits: (a) they can represent the social network (structure), (b) they perform inference between nodes/edges;

and (c) they can help capture the properties of the social network. We employ graphical models here for spatial information propagation and specifically graph wavelets to study absenteeism. The classic wavelet has been referred to as a mathematical microscope since it is capable of showing signal abnormality with different scales. Wavelets help analyze signals which contain features that vary in time, space, and frequency (scale). Graph wavelets are particularly suited to study complex networks, as they render the graph with good localization properties both in frequency and vertex (i.e. spatial) domains. Their scaling property allows us to zoom in/out of the underlying structure of the graph.

1.2.3 Epidemiological Models

Epidemiological models provide a foundational approach in social network analysis since they elucidate the embedded information diffusion process. These models typically divide the total population into several compartments which reflect the status of an individual. For instance, common compartments denote susceptible (S), exposed (E), infected (I), and recovered (R) individuals. Individuals transit from one compartment to another, with certain probabilities that have to be estimated from data. The simplest model, SI, has two states; susceptible (S) individuals get infected (I) by one of their neighbors and stay infected thereinafter. While conceptually easy to understand, it is unrealistic for practical situations. The SIS model is popular in infectious disease modeling wherein individuals can transition back and forth between susceptible (S) and infected (I) states (e.g., think of allergies and the common cold); this model is often used as the baseline model for more sophisticated approaches. The epidemic model SIR was firstly proposed to simulate the disease spreading on population groups in 1927 [15], which enables individuals to recover (R) but is not suited for modeling news cascades on Twitter since there is no intuitive mapping to what 'recovering' means. The SEIZ model (susceptible, exposed, infected, skeptic) proposed by Bettencourt et al. [3] takes the interesting approach of introducing an exposed state (E). Individuals in such a state take some time before they begin to believe (I) in a story (i.e., get infected).

1.3 Goals of the Dissertation

The overall aim of this dissertation is to identify modeling approaches and strategies that identify novel information propagate patterns as motivated earlier. We propose four mass movement topics here.

Topic 1: Mass Protest Adoption in Social Networks Modeling the movement of information within social media outlets, like Twitter, is key to understanding to how ideas spread but quantifying such movement runs into several difficulties. Two specific areas that elude a clear characterization are (i) the intrinsic random nature of individuals to

potentially adopt and subsequently broadcast a Twitter topic, and (ii) the dissemination of information via non-Twitter sources, such as news outlets and word of mouth, and its impact on Twitter propagation. These distinct yet inter-connected areas must be incorporated to generate a comprehensive model of information diffusion. We propose a bispace model to capture propagation in the union of (exclusively) Twitter and non-Twitter environments. To quantify the stochastic nature of Twitter topic propagation, we combine principles of geometric Brownian motion and traditional network graph theory. We apply Poisson process functions to model information diffusion outside of the Twitter mentions network. We discuss techniques to unify the two sub-models to accurately model information dissemination. We demonstrate the novel application of these techniques on real Twitter datasets related to mass protest adoption in social communities.

Topic 2: Protests Detection from Group Anomaly Event detection in online social media has primarily focused on identifying abnormal spikes, or bursts, in activity. However, disruptive events such as socio-economic disasters, civil unrest, and even power outages, often result in abnormal troughs involving group absenteeism of activity. We present the first study, to our knowledge, that models absenteeism and uses detected absenteeism as a basis for event detection in location based social networks (LBSN) such as Twitter. Our framework addresses the challenges of (i) early detection of absenteeism, (ii) identifying the point of origin, and (iii) identifying groups or communities underlying the absenteeism. Our approach uses the formalism of graph wavelets to represent the spatiotemporal structure and user activity in a LSBN. This formalism affords multiscale analysis, enabling us to detect anomalous behavior at different graph resolutions, which in turn allows identification of event location and anomalous groups underlying the network. We introduce a systematic two-pass detection method using graph wavelets to detect group absenteeism and then check if there is a subsequent activity spike.

Topic 3: Distinguish Real Movement in Social Networks Quantifying information diffusion on social network has been an interesting and unresolved problem for several years now. A better understanding of information diffusion, especially how news and rumors propagate through a network empower us to design strategies that can enhance spreading of news and curbing of rumors. Epidemic models have been used in the past to study information diffusion based on an assumption that rumor/news spreading is no different than the propagation of a contagious disease.

We use an enhanced epidemic model SEIZ that has been specifically designed for information diffusion. The model introduces one more compartment called exposed (E), which refers to the individuals who has been exposed to a story but have still not adopted/rejected it. We use five true news stories and three rumors from varied geographical locations and topics. We also introduce a one-step graph transfer model that can mimic step by step information propagation on Twitter. Our experimental results prove that SEIZ model is far

more accurate in describing information diffusion than the other baseline epidemic models. Further, our one-step graph transfer model imitates information cascades of the stories with a very reasonable error.

Topic 4: Pathways Inference to Climate Related Protest To infer climate protest pathways, we need to develop a classifier which is able to separate out climate related protests from others. By analyzing historical climate protest events, we identify that different climate diasters cause related protests with different time span, depends the climate diaster influence and frequency. From constructing knowledge graph to represent link relationships between entities, we discloses protest causalities in Latin American countries, illustrate the pathways from climate disasters to climate protests. We also identify the climate related protest patterns, discover the coherent relationship among different protests demands.

1.4 Organization of the Dissertation

The remainder of the dissertation proposal is organized as follows.

In Chapter **??** we review all of the related work.

In Chapter **??**, we address the problem of multiple spaces information dissemination, such as via social networks and outlets such as word of mouth. Specifically, we introduce a trust function to simulate how users are influenced by their friends through direct mention using the '@' symbol. We present how our bispace model can capture propagation in the union of (exclusively) Twitter and non-Twitter environments.

Chapter **??** defines social network movements by an undirected, weighted graph. We detect the group anomaly not only by observing the user activity, but also consider the whole network structure. We propose to use graph wavelet to detect the group anomaly from a global viewpoint. We pay attention to user activity vectors and model their behaviors on graphs and uses detected anomaly as a basis for event detection.

In Chapter **??**, we investigate the problem of distinguishing real movements from rumors in social networks. Here we design strategies that can enhance the spreading of news and the curbing of rumors. We present how to simulate the 'doubt' and 'believe' sentiment propagations. We also introduce a one-step graph transfer model that can mimic step by step information propagation on Twitter. Finally, we test the models using five true news stories and three rumors from varied geographical locations and topics. We also study the problem of misinformation propagation in the era of Ebola. All the experiments are conducted on Ebola-related rumors and all the evaluations are based on real-world data.

Chapter 2 describes how we deal with problem of identifying linkages between climate change related phenomena and climate protests. We build a climate protest classifier which is able to

separate out protests directly or indirectly resulting from a major climatic, severe weather, or environmental event. By analyzing large historical protest reports, we make use of knowledge graph to represent the link relationships between entities, and further locate and identify the causality of most climate protests.

Chapter 3 presents the concluding remarks and illustrates future research directions.

Chapter 2

Pathways Inference to Climate Related Protest

2.1 Introduction

Climate change, extreme weather, and the state of the environment directly impact the availability of food [2] [1], energy [18], and shelter [24]. As finite resources become scarce, the residual impacts on local economies can have disastrous and sometimes long-lasting effects on the fundamental livelihoods of inhabitants for decades [16]. In some cases, the resulting instability can severely detriment the ability of an established political system to maintain peace. The examples of this occurring are numerous. The extended drought in Syria in 2011 is cited as one of the principle causes of civil war [7, 14]. In a smaller scale example, the environmental impact of lead contamination in the drinking water in the United States led to protests in 2016 [1]. As we later show as an example of extreme weather, tropical storm Manuel devastated the western coasts of Mexico leading to subsequent protests over resources at times as long as 17 months after the initial event.

The path from climate, extreme weather and environmental effects to civil unrest is causally complex [10, 22] and involves various combinations of climate change [5], natural resources, human security, and social stability. In general, sensitivities to climate change, exposure to climate change, and the ability of a society to adapt are indicators of whether or not violence will erupt [11]. A commonly studied pathway is the effect of climate on food prices that then induces civil unrest. An example of this occurrence is the Arab Spring uprisings in 2011, and how weather effects food prices [13]. The pathway to civil unrest is also not limited to a local region, where one study shows the Chinese drought effecting the supply wheat causing prices to rise in the Egyptian break market leading to protest [23]. The pathways of food prices to protest have also been studied in the global south [6], Africa, and Asia [25, 9].

[1] http://www.cnn.com/2016/01/11/health/toxic-tap-water-flint-michigan/

However, even this path of climate effects on income level leading to conflict is not eminently clear [21].

Of course, the occurrence of either a shift in climate, extreme weather, or environmental catastrophe is not sufficient to guarantee that civil unrest is likely to follow. In general, the causal mechanisms leading to civil unrest are very complex, and there is no easy way to determine a linear pathway to protest. However, to date, little quantitative analysis has been performed on the residual effects of changes resulting from climate, extreme weather, and the environment using a large volume of data. In this analysis, we focus on the breadth of the climate events by looking at events generated from a large Gold Standard Report (GSR) [20] containing all of the protests that have occurred in Latin America from 2011-2015.

GSR is a gold standard report of protests organized by MITRE, using human analysts, to survey newspapers for reporting of civil unrest. The GSR includes many features, as shown in Figure 2.1, such as protest location, event date, protest type, status, crowd size, headline, date, population, protest description, first reported links, etc.. The description feature is brief description of the protest, generally, it tells us who, where, why and when protest. As Figure 2.1 shows, the protest description is 'small farmers want the bank to forgive their debts due to the drought, which has hampered production'.

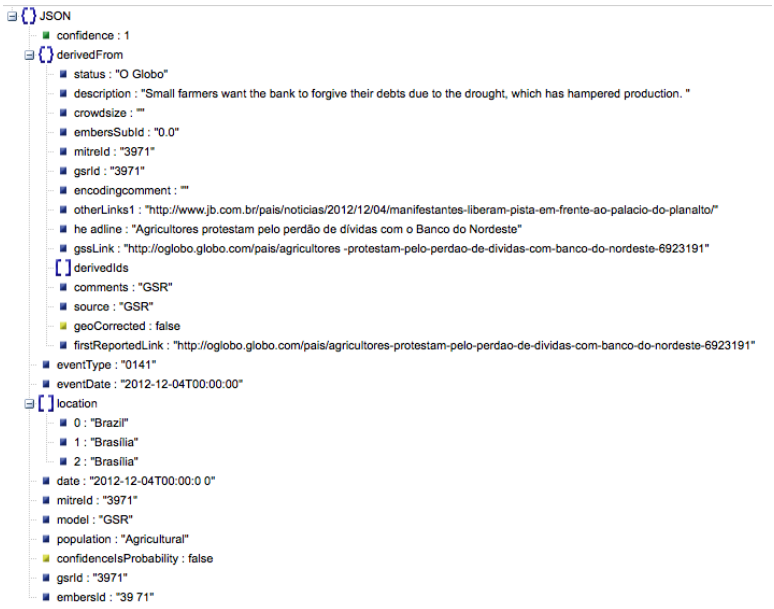


Figure 2.1: Gold standard report (GSR) format.

We address three foundational problems. First, we use machine learning to classify climate related protests. By developing a logistic regression classifier, 25352 GSR civil unrest events were classified as either being climate or non-climate related using terms in the description of the event. Second, we use the textual description of protests to extract the climate protest category for protests in each country. For each major climate category, we adopt

the knowledge graph approach to define linkage relationship between entities, and study the potentially related protest attributes. Third, we find that the massive climate protests show that certain protest types are shown to correlate with other protest types. Specifically, we find clues such as a lack of water is highly linked with power shortage. This is the first large-scale study of climate related protests to our knowledge. Generally, the main contribution of this chapter can be summarized as:

1. We develop a logistic regression classifier, which can classify climate protests from non-climate protests automatically based on protest event descriptions.
2. We analyze the climate protest spikes and disclose its relationship with climate disasters. For instance, the time span caused by storm and hurricane events in Mexico last much longer. However, for drought events in Brazil, the protests being initiated more swiftly, also last much shorter.
3. We figure out the climate protest pathways. By studying some major climate disasters, we also discover each protest category's evolution pattern, thus how does the climate disasters lead to armed climate protests.
4. We investigate the climate co-occurrence. For instance, the water related protests are often accompanies with electricity shortage, while land ownership protests are often associate with farmers.

2.2 Climate Protest Classifier

The classifier is designed to label text documents into two or more predefined categories. In this work, we only have two categories: climate or non-climate related protest. By sample analysis, more than 90% records belongs to non-climate related protest, thus the dataset can be ascribed as un-balanced dataset. So we consider majority assign classification as baseline, adopt other four classical classification methods: K-Nearest Neighbor, Naive Bayes, weighted SVM and Logistic regression.

2.2.1 Majority Assign

Majority assign method is taken as a baseline for the unbalanced classification dataset. It first calculates the climate related protest rate with the training dataset as p, and non-climate related protest as $1-p$, and then uses this distribution to randomly assign each testing event. Suppose there are N testing events, by this algorithms, the true-positive would be Np^2, false-positive and false-negative would both be $Np(1-p)$, on average. Hence the precision, recall and $F-measure$ would all be p, and the accuracy would be $p^2+(1-p)^2$.

For unbalanced dataset, since $p << (1-p)$, the accuracy approximately equals to $(1-p)^2$, while the $F-measure$ is p.

2.2.2 K Nearest Neighbor

K Nearest Neighbor (KNN) classifier is to label data based on a majority vote of its neighbors, where the vote is measured by a distance function. Distance function is used to measure similarity, which is not necessarily be Euclidean distance or the cosine value although they are most commonly used. Clearly, text similarity plays a fundamentally important role in labeling datastet. In our study, we calculate two GSR records' similarity using cosine similarity [17]. KNN is a supervised learning method which require to prepare training dataset. In the training dataset, each GSR record is transformed into vectors and are labelled as climate or non-climate protest. How to choose the optimal value for K is another factor influencing classification result. After a set of comparison, we found setting K to be 100 tends to get the best performance.

2.2.3 Naive Bayes

Essentially, Naive Bayes is to maximize a posteriori classifier, which can be represented as $c = \text{argmax}_c p(c|e)$. e is the protest description, and consists of multiple words w_i, and can be denoted as $e =< w_0, w_1, ... >$. c={climate protest, non-climate protest}. However, there is no trivial solution to measure the joint probability distribution for e, c considering the extremely complex underlying structures among w_i. Naive Bayes circumvents this problem by assuming the independency among w_i. Hence, the probability of each protest e being class c can be simplified as:

$$p(c|e) \propto p(c) \prod_i p(w_i|c)$$

, where $p(w_i|c) = \frac{f^c_{w_i}}{f^c_w}$ is the conditional probability of term w_i that appears in the description of e. $f^c_{w_i}$ is the occur frequencies of w_i in class c, and f^c_w is the total word number in class c. If a new term w_i does not occur in the training dataset, then $p(w_i|c) = 0$. $p(w_i|c)$ measures how much likeness of being c for the existence of term w_i. To mitigate the zeroing affects, $Laplace - smoothing$ modifies $p(w_i|c)$ as

$$p(w_i|c) = \frac{f^c_{w_i} + 1}{f^c_w + W}$$

, where W is the total word number for climate and non-climate protests together. Usually, the conditional probability is small which might results in float point underflow. In reality, it is converted as:

$$c = \text{argmax} p(c|e) = \text{argmax}_c \{log(p(c)) + \sum_i log(p(w_i)|c)\}.$$

2.2.4 Weighted Support Vector Machine

Traditionally, the Support Vector Machine works in this way. The training data consists of N pairs $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, with $x_i \in \mathbb{R}^p$, and $y_i \in \{-1, +1\}$. By introducing a hyperplane of $P := \{x | x^T\beta + \beta_0 = 0\}$, the classification rule is defined as $G(x) = \text{sign}[x^T\beta + \beta_0]$. To find the hyperplane P for inseparable sets, it is often converted into the following quadratic convex optimization problem by defining the slack variables $\zeta = (\zeta_1, \zeta_2, .., \zeta_N)$ [8].

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\zeta_i \\ \text{subject to} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \zeta_i, \quad \forall i \\ & \zeta_i \geq 0, \qquad\qquad\qquad\quad \forall i \end{aligned} \tag{2.1}$$

where C is the penalty parameter. For separable sets, C corresponds to ∞.

The problem with above classifier is that the penalty for misclassification are the same. However, in cases with unbalanced dataset, the miss alarm should have a much higher cost than the false alarms. As in our study, the climate related protest is more important. To consider these scenarios, we introduce two different penalties for miss alarm and false alarm. For simplicity, I and J denotes the subscript of positive and negative set. Thus, the problem of 2.1 can be re-formulated as:

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C_1\sum_{i\in I}\zeta_i + C_2\sum_{j\in J}\eta_j \\ \text{subject to} \quad & x_i^T\beta + \beta_0 \geq 1 - \zeta_i, \forall i \in I \\ & x_j^T\beta + \beta_0 \leq -1 + \eta_j, \forall j \in J \\ & \zeta_i \geq 0, \forall i \in I \\ & \eta_j \geq 0, \forall j \in J \end{aligned} \tag{2.2}$$

The Lagrange function of 2.2 is

$$\begin{aligned} L_p = & \frac{1}{2}||\beta||^2 + C_1\sum_{i\in I}\zeta_i + C_2\sum_{j\in J}\eta_j \\ & - \sum_{i\in I}\alpha_i[x_i^T\beta + \beta_0 - (1 - \zeta_i)] - \sum_{i\in I}\mu_i\zeta_i \\ & + \sum_{j\in J}\theta_j[x_j^T\beta + \beta_0 - (1 - \eta_j)] - \sum_{j\in J}\tau_j\eta_j \end{aligned} \tag{2.3}$$

2.2.5 Logistic Regression

The classifier is built based on logistic regression method. To reduce the computation complexity, we only apply the GSR description text as input to the logistic regression classifier. First of all, we construct a bag of words from the training dataset descriptions by deleting meaningless stop-words, like "the", "a/an", "at", and etc.. The bag of words is composed of M words denoted as $[w_1, w_2, ..., w_M]$. Each GSR description X is considered as a vector of length M. If word w_i occurs in its description, then $X(i)$ will be assign with 1; otherwise 0. Further each protest in the training dataset is assigned $Y = 1$ as climate protest, or $Y = 0$ as non-climate protest by manually checking its description meaning. In this way, each GSR record is converted to a corresponding vector based on the bag of words. Second of all, we estimate k_i and b based on maximum likelihood criterion. This process is usually converted to convex optimization problem with efficient solutions [12]. Once the coefficients of k_i and b are estimated, the probability for each class is calculated by:

$$P(Y = 0|X = x) = \frac{1}{1 + exp(\sum_{i=1}^{N} k_i x_i + b)}$$

$$P(Y = 1|X = x) = \frac{exp(\sum_{i=1}^{N} k_i x_i + b)}{1 + exp(\sum_{i=1}^{N} k_i x_i + b)}$$

If the probability of $P(Y = 1|X = x)$ is larger than 0.5, then the protest event will be classified as climate protest, otherwise, non-climate protest.

2.2.6 Evaluation

We manually labelled 1700 GSR protest records as climate or non-climate protests. Using 70% dataset as training, and the rest 30% as test. To ensure we have a trustworthy classification results, we evaluate the performance carefully by cross evaluation. The evaluation criteria are precision (positive predictive value), recall (true positive rate), F-measure (a measure that combines precision and recall) and accuracy (the proportion of true results both true positives and true negatives among the total number of cases examined). We compare with four well-known classification methods: majority assign, K-nearest neighbor, Naive Bayes, and weighted support vector machine (SVM). Since the climate events only account for a small portion of all the events, which make it an unbalanced classification problem, so we change the traditional support vector machine into weighted SVM, by adding more importance to the climate portest events (we set the class weight to be 100). The experiment results are shown in Table 2.1, where we demonstrate logistic regression method outperforms other methods uniformly.

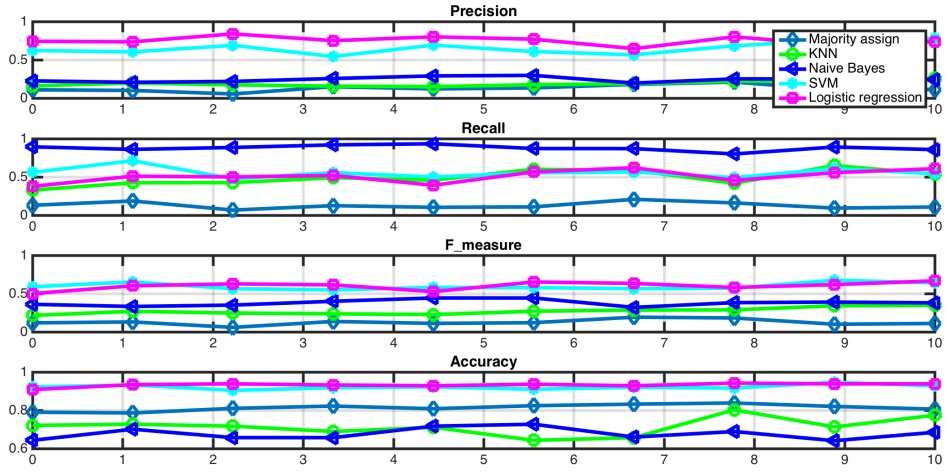


Figure 2.2: Classification methods comparison.

Table 2.1: Classification methods comparison.

	Precision	**Recall**	**F_ measure**	**Accuracy**
Majority assign	0.1274	0.1289	0.1258	0.8136
KNN	0.1906	0.4913	0.2723	0.7154
Naive Bayes	0.2432	0.8779	0.3798	0.6777
Weighted SVM	0.6543	0.5565	0.5966	0.9218
Logisitic Regression	0.7513	0.5102	**0.6018**	**0.9322**

2.3 Experimental Results

There were a total of 25352 recorded civil unrest events in Latin American countries from July 2011 to March 2015 that were included in our dataset. Using our climate protest classifier, we were able to separate out protests directly or indirectly resulting from a major climatic, severe weather, or environmental event. In the subsequent analysis, these three categories of event types are labeled with a common definition of "climate event". Of the candidate civil unrest events, 991 (3.9%) events are classified as climate-motivated across all Latin American countries for that time period. In the subsequent sections, we conduct a multi-dimensional analysis of these protests to understand potential implications of the breadth of impact resulting from climate motivated protests.

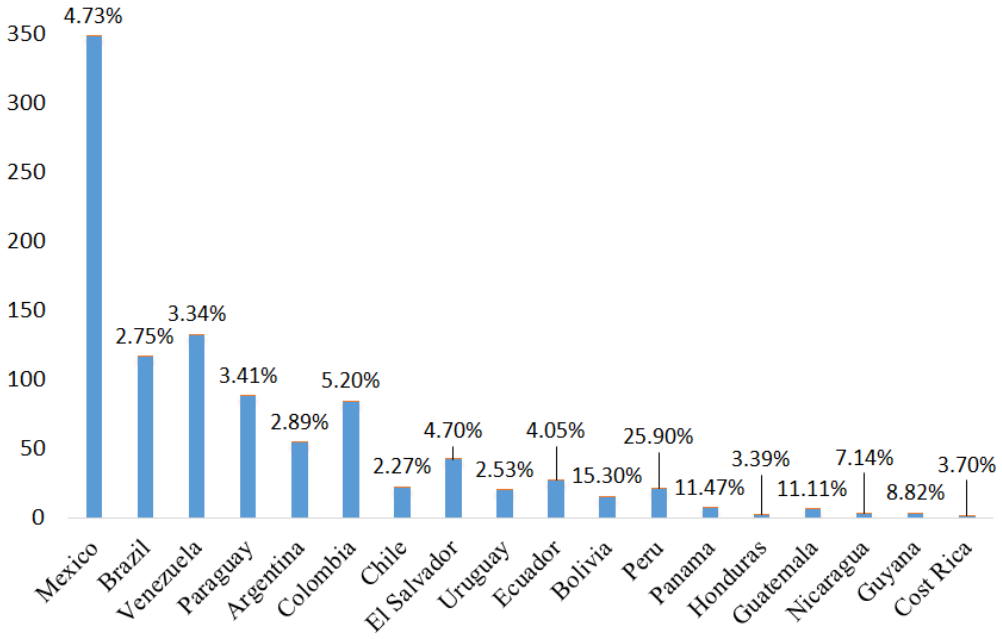
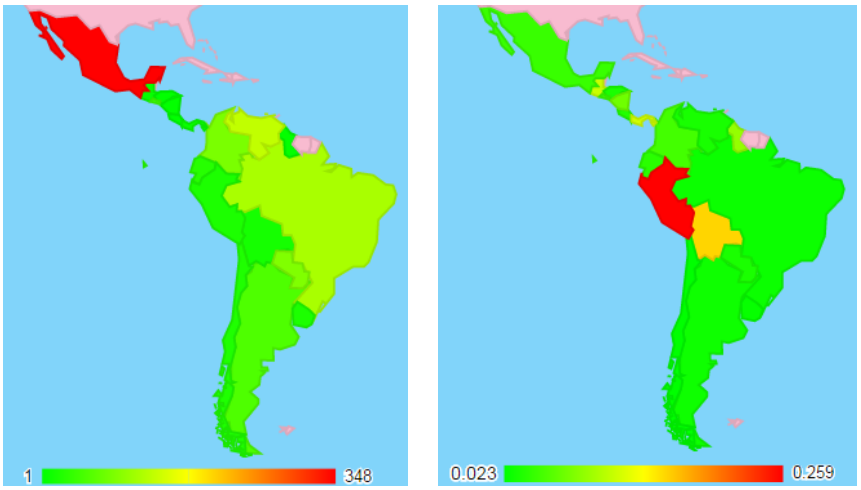


Figure 2.3: Blue bar shows all the climate related protests for each country, yellow bar shows the climate protest percentage over its total protest, from Jan 2011 to March 2015.

2.3.1 Frequency Analysis by Country

The first analysis we conduct is a comparison of the representative number of protests within and across each country. The results of our classifier selection show the number of climate motivated protests and the percentage over all the protests in that country, as can be seen in Figure 2.3. The country with the most climate protests overall is Mexico, and Costa Rica has the least. As evidenced by the climate to non-climate protest ratio, the portion of protests related to climate remains fairly constant across countries with the exception of Peru. In this particular case, there were numerous protests centered on mining and its effect on the environment that dominate the overall protest landscape. As the number of total protests decrease, we see more variability in the ratio as expected. For these countries, which typically have smaller populations, the significance of a single type of protest has more of an impact on the measure than larger countries.

To show the effect of the population on the number of climate protests, we plot the result of a linear regression in Figure 2.5. The result of this shows an $R^2 = 0.64$, showing a slight linear relationship. However, the interesting part of this analysis lies in the residual errors. The set of countries including Mexico, Venezuela, Paraguay, and Colombia, all demonstrate the occurrence of more climate protests than would be expected given the entire dataset. On the contrary, Brazil has fewer climate protests given the size of their population. There could be a number of reasons for these findings such as socio-political stability, environmental sensitivity, and the type of climate events. All of these are potential avenues for further causal



(a) (b)

Figure 2.4: (a) Climate related protests events numbers; (b) Climate related protests percentage in Latin American countries, from Jan 2011 to March 2015.

or anecdotal studies. In the following, however, we choose Brazil, Mexico, and Venezuela for further analysis into overall trends of climate protests, and how these are shaped in the data recovered by the classifier.

2.3.2 Spatial Distribution of Climate Protests

In this manuscript we are defining the climate protest as being different from a regular civil-unrest event by a relation to an climate event. Next, we investigate if there is any fundamental difference in terms of where these protests occur in relation to protests in general. For this analysis we use Mexico, Brazil, and Venezuela which all have many protest events, and the percentage of those that are related to the climate are all at about 4%. The spatial distribution of events is shown in Figure 2.6. Both the total number of protests and those that are climate motivated are shown and represented by the size of the blue and red shaded circles, respectively.

In both Brazil and Venezuela, many of the protests appear at or near their coastal boundaries, and Mexico has more inland activity. However, we have already established a connection between population and protests. This is no different for the spatial distribution, where much of the population of Brazil and Venezuela is located in coastal regions. The protests in Brazil mainly center at two major cities Sao Paulo and Rio de Janeiro. In Mexico and Venezuela climate protests have a more uniform distribution across the cities. Therefore, there is no particularly strong evidence to suggest that certain regions of these countries are more prone to protest with respect to the climate than they would normally be willing to protest in general. In terms of the climate events defined in this study, effects of climate,

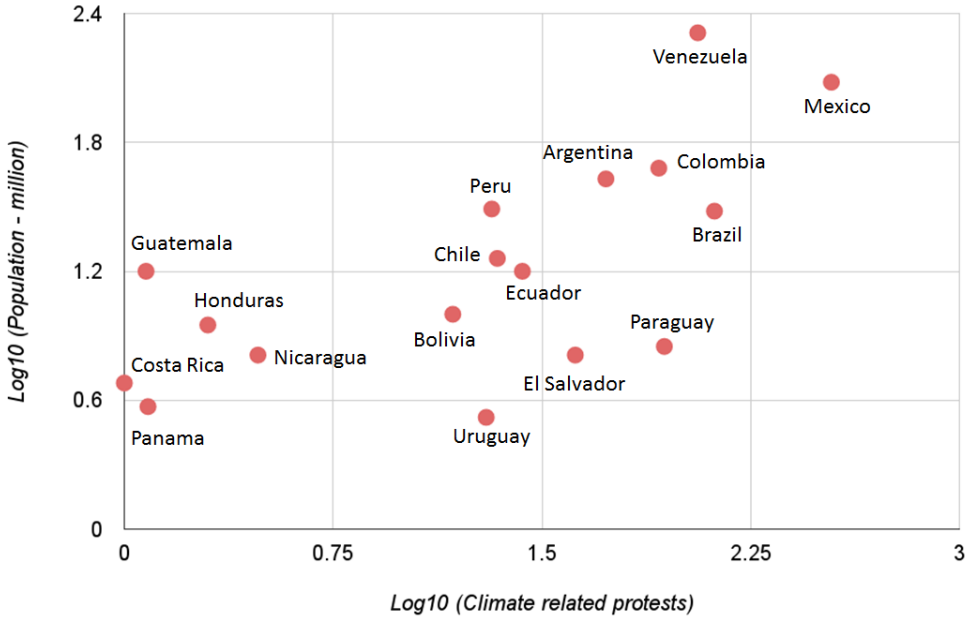


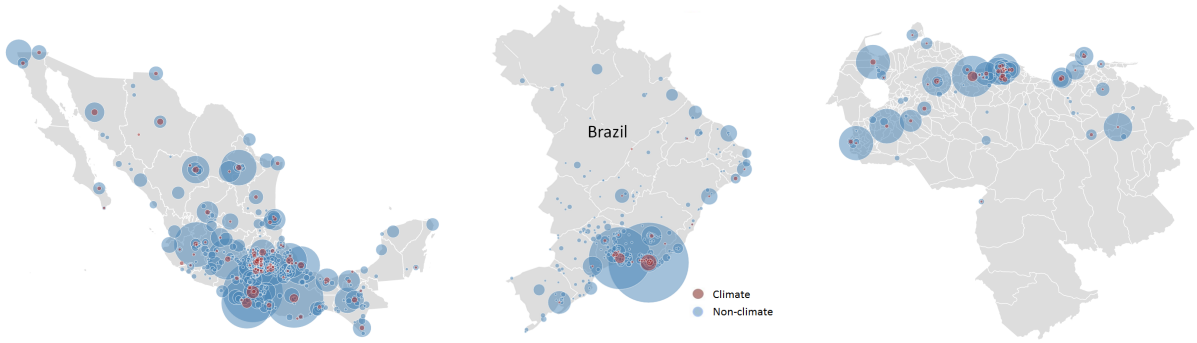
Figure 2.5: Log10 (Climate protest events) and log10 (population - million) of each country. The two series have a Pearson correlation coefficient 0.70.

the environment, and extreme weather are not regionally exclusive to certain populations. Through complex channels such as food supply, the effects of climate impact can ripple across spatial networks.

2.3.3 Temporal Dependency on Climate Events

The temporal dependency of climate protest occurrences is analyzed for each country. As with the spatial domain, the effects of climate events are non-local in time in some cases. The ground truth for the events was established for extreme weather only, as the event itself is more local in time than climate and environmental changes. This data is available by combining the following sources: International Disaster Database EMDAT[2], World Disasters Timeline[3] and European Commission's Humanitarian Aid and Civil Protection department (ECHO)[4]. The official climate disaster report for each country is shown with climate related protests in Figure 2.7.

[2]http://www.emdat.be/database
[3]http://www.mapreport.com/
[4]http://ec.europa.eu/echo/



(a) Mexico (b) Brazil (c) Venezuela

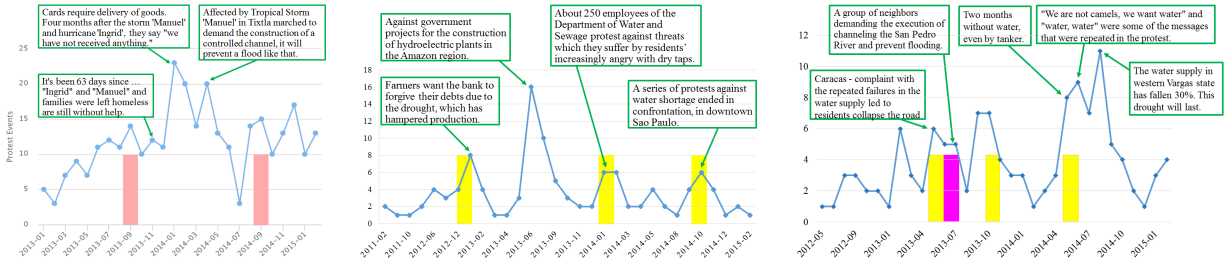
Figure 2.6: Climate and non-climate protests from July 2012 to March, 2015. Red circle represents climate related protest events, and blue circle represents non-climate related protests.

Mexico climate disasters Figure 2.7(a) shows the Mexican extreme weather events and protests where the blue time series represents the climate protests events, and the two red bars shows the occurrence of two storms. The first storm is the combined tropical storm Manuel (category 1) and hurricane Ingrid in September 17, 2013. The track maps can be seen in Figure 2.8(a). Tropical storm Manuel crossed the west coast of Mexico and resulted in more than 23,000 people fleeing their homes due to heavy rains spawned by what had been Hurricane Ingrid. Of those displaced 9,000 went to emergency shelters. In terms of infrastructure, at least 20 highways and 12 bridges had been damaged[5]. After the storm, related protests and other civil unrest events broke out and lasted for more than 17 months because the government's response had been inadequate. The storm related protests reached a climax in January 2014, and second climax in April 2014. On November 19, 2013, there was report saying "it's been 63 days since the onslaught of 'Ingrid' and 'Manuel' and families were left homeless are still without help"[6] Four months after the storm Manuel and the effects of Hurricane Ingrid, they say "we have not received anything". On April 7, protest descriptions said "Affected by Tropical Storm 'Manuel' in the municipal head of Tixtla marched to demand the construction of a controlled channel, it will prevent a flood like that caused the overflow from the Black Lagoon in September 2013". The last protest event we have on record from the climate protest classifier occurred 17 months after the original event. This demonstrates that the residual capacity of these events to impact the livelihoods of people is not guaranteed to be local in time. As we show, the range of impact can extend even beyond the occurrence of other storms.

In Figure 2.7(a), the second red bar shows hurricane Odile. It is a category 3 storm that occurred in 2014, and the track of the storm's path is shown in Figure 2.8(b). Despite

[5]https://weather.com/storms/hurricane/news/tropical-storm-manuel-hurricane-ingrid-hit-mexico-opposite-coasts-20130916

[6]Quotes are translated from the native language of the country.



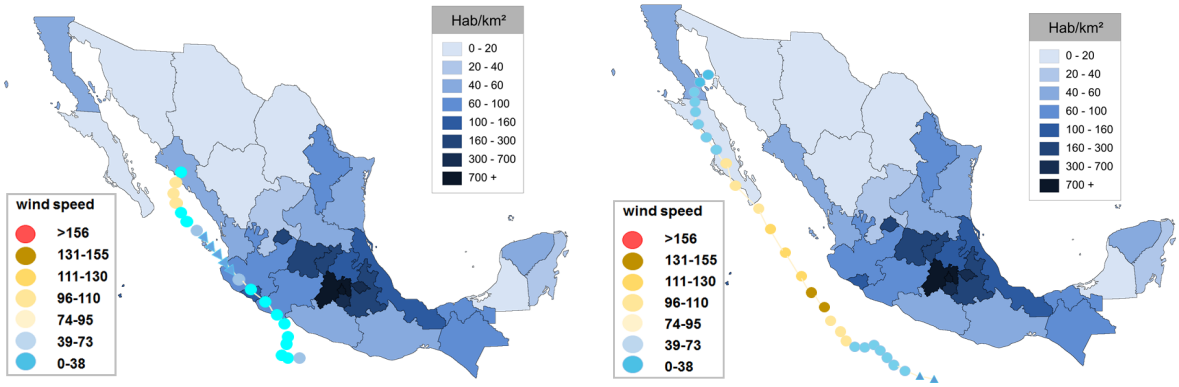
(a) Mexico (b) Brazil (c) Venezuela

Figure 2.7: (a) Mexico climate disasters and climate protests. The blue time series shows the climate related protest events, and light red vertical lines show two storm diasters in Mexico, storm Manuel in September 17, 2013 and hurricane Odile in September 15, 2014 respectively. (b)Brazil climate disasters and climate protests. The blue time series shows the climate related protest events, and yellow vertical lines show three drought diasters in Brazil, drought in Feb 2012, Heat wave in Feb 2014, and drought in Oct 2014, respectively. (c) Venezuela climate disasters and climate protests. The blue time series shows the climate related protest events, and rose vertical line show local area flood diaster, and yellow vertical lines drought disasters.

hurricane Odile being a more intense storm, there were not many protests related to the event. Comparing the storm's paths in Figure 2.8, Tropical Storm Manuel hit Mexico's mainland, which caused more destruction. Hurricane Odile 2014 had less of an impact on the Mexican mainland, even though it crossed the state of Baja California. However, this is the second smallest Mexican state by population. This can explain why storm 2013 lead to tremendous protests, while hurricane 2014 does not.

Brazil climate disasters Figure 2.7(b) shows the relationship between protests classified by our algorithm and actual extreme weather events in Brazil. The three yellow bars show three separate drought events in Brazil, which resulted in drought related protests almost immediately. The drought in February 2012 hampered production, which caused farmers to protest. The heat wave in February in 2014, and drought in October 2014 resulted in water shortages, causing civil unrest. The biggest spike in June 2013 described protests against government's projects for the construction of hydroelectric plants in the Amazon region[7] and is more of an environmental impact type of event. In general, for these events we see predominantly local relationships in time between the protest and the preceding event. For Brazil in particular, the extreme weather event matches fairly well with the onset of drought.

[7]http://www.bloomberg.com/news/articles/2013-06-05/protests-over-brazil-hydropower-leads-to-delays-and-boosts-costs



(a) Storm Manuel, Sept 2013 (b) Hurricane Odile, Sept 2014

Figure 2.8: The map shows population density of all Mexico's 32 states. The track shows Tropical Storm Manuel of 2013 and Hurricane Odile of 2014, points in different color represent the wind speeds.

Figure 2.9: Word cloud of Mexico storm Manuel, Sept 13, 2013.

Venezuela climate disasters In Figure 2.7(c), the climate motivated events are shown in relation to relevant extreme weather events for Venezuela. The pink bar represents sudden onslaught of rain in June 2013 that caused a heightened risk of flooding and landslides in the densely populated communities on the outskirts of Caracas. It triggered a small portion of protests to prevent flooding. The yellow bars denote drought disasters. The drought in May 2014 triggered rationing of tap water in the capital, Caracas, where residents formed lines lasting hours to fill jugs of water[8]. This drought disaster lasted so long that related protests reached a climax in September 2014. Unlike Brazil, the data in Venezuela on droughts proved tough to ascribe to a particular drought event. They occur rather frequently and there is a substantial amount of overlap in the residual protest events that it was difficult to distinguish to which it was referring.

[8]http://www.breitbart.com/national-security/2014/05/31/severe-scarcity-prompts-venezuelan-government-to-ration-water/

2.4 Results Discussion

Figure 2.10: Word cloud of all the climate related protests, from GSR descriptions.

Of the climate related protests, we are interested in what are the protesters demanding. To have a birds view of climate protests, we extract all the climate protest descriptions and plot the word cloud, as shows in Figure 2.10. We can see words like 'water', 'storm', 'mining', 'rain', 'construction', 'power', 'heat', 'gas', 'environment', 'electricity', and other weather, environment related keywords are dominant, which gives us a general idea of what protesters are demanding.

As stated previously, we are not blind to the realization that the causes of climate motivated protests are in general complex. In the following, we analyze the descriptions of the protest events in order to gain insight into the general pathways by which protests within our corpus have occurred. Shown in Figure 2.11 is a weighted Sankey diagram showing the bipartite graph of the most common keywords in the descriptions of protests from each country. Apparently, many of the protests identified by the classifier in one way or another have something to do with lack of water followed by climateal effects in general. Other prominent keywords include mentions of power and energy issues. Each country also exhibits its own protest keyword categories. In Mexico, the most notable protest keywords involve are lack of water, environmental concern, storm and hurricane. In Venezuela, apart from lack of water and environment problems, the dominant keywords are blackout and energy issues. In Peru, more than half of climate protests are about a mining project, which is an environment concern. While in Argentina, 35% events protest against blackout issues. We expand on these observations in the following where we analyze several dimensions of the keywords to extract details about pathways to protest.

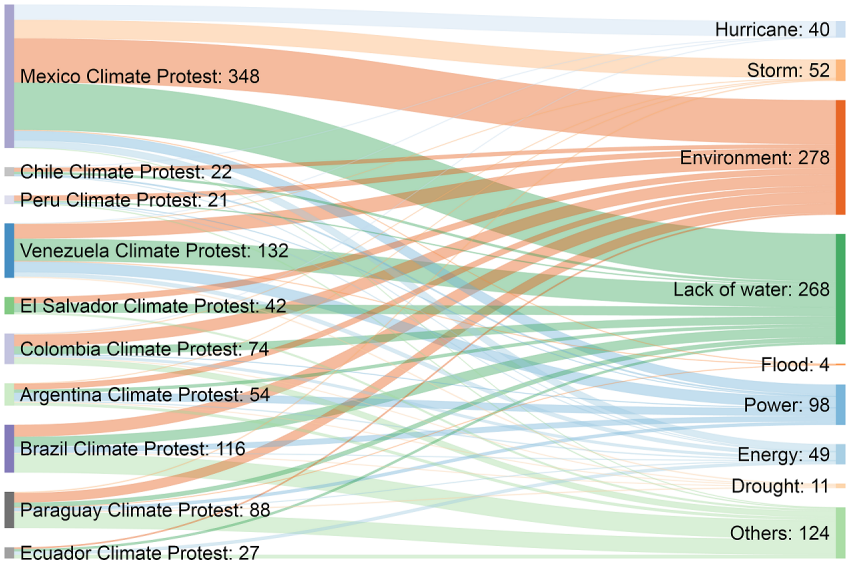
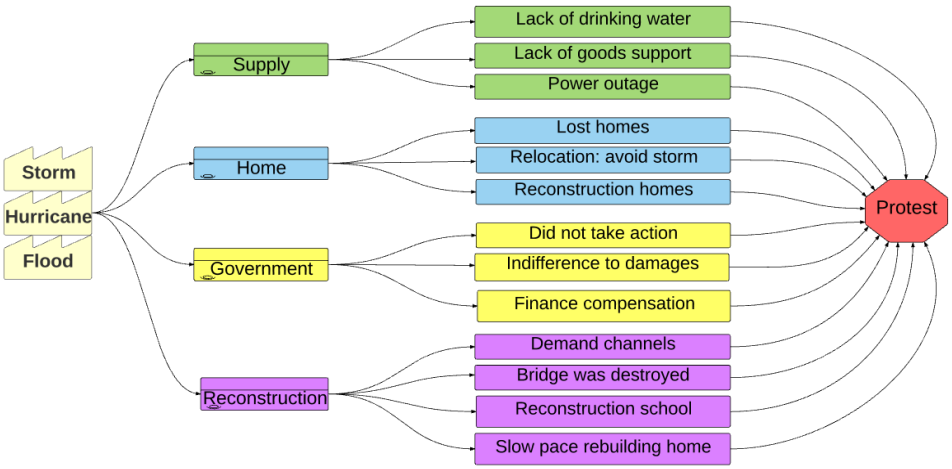


Figure 2.11: Climate motivated protest keywords diagram. Countries on the left are matched with keywords appearing in the description of the protest event on the right.

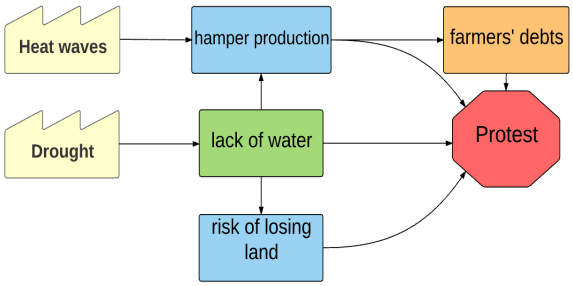
2.4.1 Climate Protest Precursors

For some severe and dominant climate events, such as storm, hurricane, flood, and drought events, we employ the knowledge graph to represent the link relationships between entities. By matching the object or subject with climate related keywords, and predicate to be causality relationship like result of, cause by, lead to, blamed, accused of, demanding, against, request, we can locate and further identify the pathways of most protest descriptors.

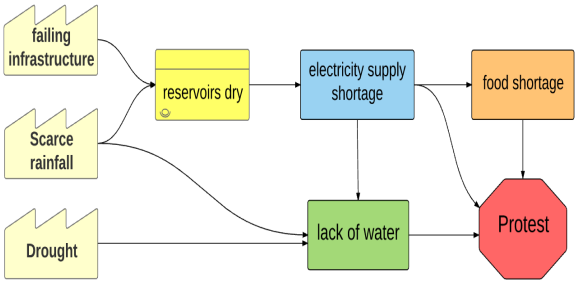
Figure 2.12(a) shows the storm caused protests demands in Mexico, which generally falls into four categories: supply, home, government and reconstruction. In the supply related protest, the causality includes but not limited to: lack of drinking water, lack of good support, and power outage. The second category is about home, they protest either because of lost homes, or request to relocate to avoid storm, or request to reconstruct homes. Another protest type targets at government, they either fight because government did not take action, or blame government's indifference to damages, or request finance compensation to the damages. In the reconstruction category, residents demand reconstruct channels to avoid more storms, request to reconstruct bridges, roads, schools, or unsatisfied with the slow pace of rebuilding homes. Figure 2.12(b) describes the pathways of Brazil climate related protests. One line is heat wave hampered production, which cause farmers' protests, the other line is drought causing residents lack of drinking water thus lead to protest, and the third line is lack of water causing farmers facing the risk of losing land, which result in protest. In Venezuela, the protests are more water-electricity centralized, as shown in Figure 2.13(c). Scarce rainfall, drought plus failing infrastructure, which makes water shortage and blackout is an everyday



(a) Mexico



(b) Brazil



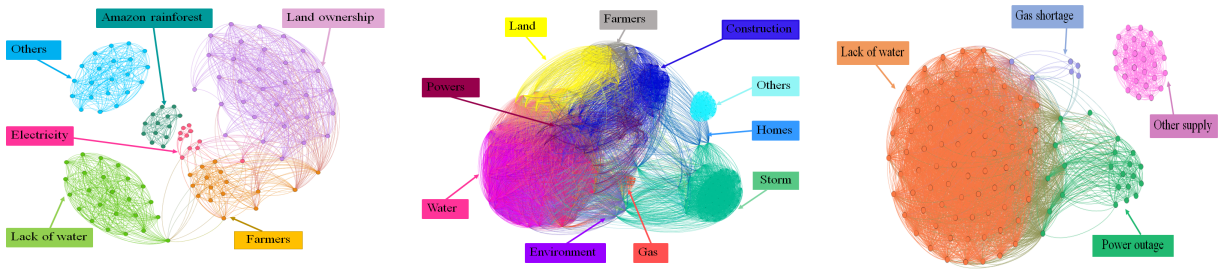
(c) Venezuela

Figure 2.12: Climate protest pathways diagram

fact of life in Venezuela. The electricity shortage deteriorates water shortage, leads to food shortage, and worsens food quality, and so forth. All those situation touches off climate related protests.

2.4.2 Climate Protest Pattern

The above analysis shows events that commonly co-occur with protests; however, we intend to further illuminate the correlations surrounding protest activity. In the following we take a graph theoretic approach, where we treat each protest event as a node and connect two nodes with a weight based on their protest descriptions and text similarity. Specially, we pay attention to the protest themes or protest demands. If two descriptions have the same protest demand, their weight will be very high. Otherwise, their connection weight tends to be 0. In this way, we build a weighted undirected network $G(V, E, W)$, with each protest as node V, and their connection as edge E, their weight as W. If the weight between two nodes is 0, there will be no edge. We employ Louvain method [4] to split the network into several clusters.



(a) Brazil (b) Mexico (c) Venezuela

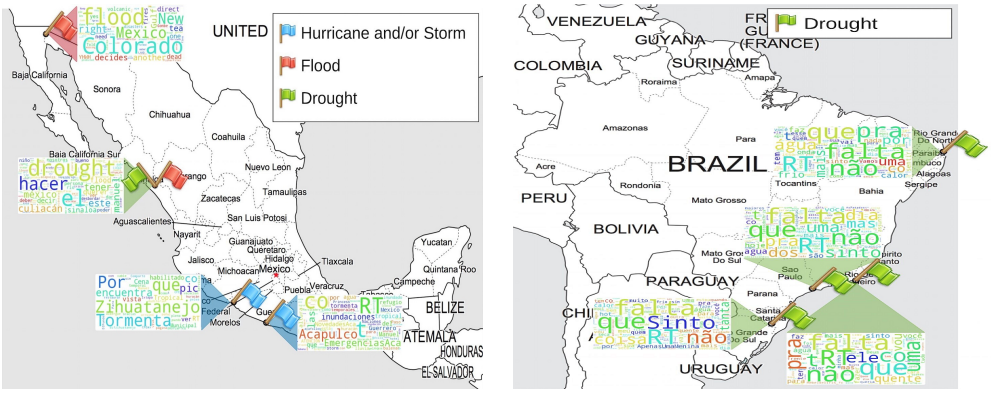
Figure 2.13: Climate protest clustering results

We show in Figure 2.13 the climate protest clustering results that provides the protest proportion and coherent correlations among different protest types. Figure 2.13(a) illustrates Brazil's climate protest pattern. We see that the largest protest cluster is about land ownership which accounts for 26.7% of the protests. The second cluster is lack of water and takes up 20.7%. The farmers cluster occupies 13.8%. We note that land and farmer clusters are closely correlated, and lack of water is also closely binded to farmers as well. Amazon rainforest is another striking protest which is responsible for 11.3%. Figure 2.13(b) shows the protest patterns of Mexico, which has the most climate protest events and most complex patterns. We can see the red cluster which denotes lack of water is the most dominant protest, accounting for 20.5%. The green cluster represents the tropical storm, and is the second largest protest type, taking up 19.0%. The dark blue cluster construction accounts for 17%, and the yellow cluster land is responsible for 11.8%. We find that water related protests are intertwined with environment protests and power protests. Land protests are closely related with farmers, while construction clusters are coherent with homes (2.6%). Figure 2.13(c) gives the overview clustering results of Venezuela climate protests. We see the yellow cluster representing lack of water protests takes up the largest portion, as high as 55.8%, and the green cluster denoting power outages accounts for the second part at 22.1%. The blue cluster, which stands for gas shortage accounts for 5%. The purple cluster shows the rest climate protest portion, which includes food shortage, medicine shortage, water tank robbery behavior, etc. The lack of water protest is intertwined with power outage protest, which corresponds to the fact that lack of water and power shortage is everyday life in Venezuela.

2.4.3 Climate Protest Influence

We are also interested in climate events influence on social media, such as Twitter. Using keywords list we are able to filter tweets, then cluster tweets into different partitions based on similarity among tweets using distance function, taking tweets content, geolocation and other features into consideration.

Events Clustering is used to separate events happened at same place or at same time, or the separate different events happened simultaneously on local and entire country. By measure the distance among tweets based on similarity, tweets collection can be clustered into subsets in which tweets are exactly related and similar. Each partition includes similar tweets stand for a specific event. Without events clustering, different events will be mixed. As shown in Figure 2.14(a), Mexico Hurricane were mixed with severe drought happen in Culiacan, with the aid of event clustering, we are able to distinguish those distinct extreme weather events, even though they may happen at the same time. For each event, we plot the related Tweets word cloud besides the flag. Figure 2.14(b) illustrates four drought events in Brazil, on May 2012.



(a) Mexico (b) Brazil

Figure 2.14: Climate protest events in Mexico, Sept 2013 and Brazil, May 2012. Different flag represents different climate disasters. The adjacent world cloud shows Twitter discussion as per that event.

2.5 Conclusion

Climate changes, extreme weather and environmental catastrophes can all exert a devastating amount of harm to people around the world. To better understand how these events lead to protest behaviors, we show different pathways to protest following severe events in Latin America from 2011 to 2015. Our analysis differs from those previously published in that we consider the breadth of climate protests over a wide spatial and temporal domain. This is accomplished by identifying climate related protests using a logistic regression classifier acting over keyword vectors of protests descriptions in our protest GSR dataset. We found this approach achieved an F-score of 0.60 and accuracy of 0.93, which was the best preforming of other common binary classifiers. The results of the classifier indicate a number of broad properties about climate related protests.

From our analysis, we found different climate disasters may cause related protests with

different time span. For example, the tropical storm Manuel in Mexico was followed by a wave of climate related protests lasting as long as 17 months. In Venezuela, the protests caused by one drought always overlap with the other drought. This chapter discloses protest causalities in Latin American countries, and illustrates the pathways from climate disasters to climate protests. This chapter also identifies the climate related protest patterns, and discovers the coherent relationships among different protests.

Chapter 3

Conclusions and Future Directions

3.1 Contributions

As social media (e.g., Twitter) continues to increase in popularity, it is becoming employed as a social sensor into real-world mass movement event detection. Modeling and studying their adoption patterns gives us insight into investigating social and physical aspects of those events and precursors. This dissertation has presented several approaches and strategies with the goal of detecting and predicting mass movements and further inferring its causality, with given information mixed with real news and rumors. Those include techniques to capture information propagation across multiple spaces, as well as a graph wavelet approach that broadens predictive capabilities to capture group anomaly within dynamic changing networks. Numerous forms of mass movements have been investigated and diverse aspects of modeling and detecting have been addressed.

Using social media as indicators for real-word event detection is indeed helpful tool, however, they do possess limitations, perhaps most notably when applied to a specific event type, such as mass movement studied here. First, modeling protest-related topic propagation on networks is never trivial. One challenge is social protest propagation through online media can spread over large areas more quickly than traditional methods since users are geographically distributed, the other challenges include mass protest information can be spread by multiple social medias and lot of paths, like word of mouth, TV and news broadcast. Second, detect the group abnormality on social media is challenging. One challenge is Twitter's user network embodies many subgraphs based on social ties which is dynamically changing the graph structures since users are active. The other challenge includes real world events are not only correlated with burst signals, but can also exhibit unusually low levels of activity in social networks. Despite these restrictions, graph wavelet have in fact provided powerful capacity in capturing graph abnormality (considering burst behavior and absenteeism behavior), even on dynamic changing networks.

A fundamental objective of this dissertation has been to model mass movement adoption behavior, and in doing so, several significant advantages are gained beyond the target. One contribution is the ability to model information propagation across multiple networks/spaces, and capture the propagation speed and possible propagation paths, which is demonstrated in Chapter **??**. Another benefit that enhanced the mass movement detection is though group anomaly detection, as introduced in chapter **??**. Graph wavelet provides appropriate definition of group anomaly which can cover both burst and absenteeism with different scales, thereby increasing the probability of capturing protest behaviors. Another benefit is the capability of quantify compartment transition dynamics using epidemic model SEIZ, and facilitate the development of screening criteria for distinguishing real movements from rumors happenings on Twitter, as demonstrated in Chapter **??**.

Understanding information propagation over dynamic social network is highly-popular for addressing real-world problems in social network analysis. This dissertation analyzes several fundamental questions underlie the propagation-like processes, such as mass movement adoption, rumors transmission. These methodologies can be extended to other applications such as infectious diseases, public health, marketing, and so on.

3.2 Future Directions

One of the major attractive areas would be continuing focusing on social network analysis, specifically the information propagation research over dynamic changing social network. Thereby the future research directions will fall into two categories, one is to deepen the existing theory and algorithm, the other is to broaden the current research.

Extend GBM model What would happen to the geometric Brownian motion model if the underlying mention network changes over time? How to adopt or modify this model when apply into multiple networks? As well as those theoretical questions, there are also some applications worth further investigation, such as, can we introduce the GBM model into infectious disease domain, for example, zika virus spreading? Assume the Bispace is composed with connection network and the other is physical space, can we train the GBM model to estimate each use's infection probability based on their environment?

Further study graph wavelet We hope to extend the graph wavelet applications into other areas, based on the two distinguish properties of graph wavelet. One is the ability to detect graph abnormality, such attribute can be adopted into detecting the wealth gap between rich and poor in one region, identifying the brain neural network abnormality, or detecting traffic congestion through road network analysis; the other property is the ability to identify the central point of a subgraph, which can be employed to rank key players over

networks, detect the rumor spreaders in some cascade, or find the source of infection as per certain disease.

Broaden rumor detection scenarios In stead of predict a story is true or false, it is more practicable to label how much people tend to believe it. Newspapers would find it very useful, especially when it comes to some breathtaking news yet not being confirmed. Before reporting to the public, they would like to grasp how much the story is believable. Also it is valuable as to decide whether vendors are cheating during online shopping.

Deep understand personalized information propagation. We would like to understand how users' behavior lead to the information propagation delay or boost, especially when accompanied with strong sentiment. This may help formulate advisement strategy, if we can find a way to manipulate the information flow. Further, we would like to explore opportunities to extract personalized information spreading pattern. What kind of news may arouse his/her interest, if so, what kind of role he/she may play, what kind of push strategy may stimulate his/her activity? This study is propitious for precision marketing or personalized recommendations, provided refined content filtering.

Build an intelligent disaster detection system We would like to build an intelligent system which is efficient at event detection, especially for some diasters, protests, extraordinary events. It cannot only do immediate reporting, but also able to track events and do causality analysis and even do future prediction. The system has some critical building blocks: natural disaster detection using graph wavelet by detecting group anomaly, rumor or news detection by employing epidemic SEIZ model, story causality analysis, and event coding. Take the flint water crisis for example, firstly it will identify the water crisis events from social media analysis, then confirm it is true story, next, it will trace all the historical news to identify the causality, finally, it will generate a complete report using event coding.

Combine social network analysis with physical data. The advent of social media provides unprecedent opportunities to access vast information which can benefit our research. Given such a convenience, we would like to explore the possibilities to renovate some traditional research, hoping to have some extraordinary discoveries and bring more vitality. Take the vaccine and its adverse effects study for example. Traditionally, vaccine research heavily depends on the raw data collected by CDC, hospitals, patients report, and vaccine adverse event reports. However, this data usually suffers some problems, such as time delay, some information is incomplete. Worse more, most of data is isolated. If we can find a way to combine those statistic data with social media data(Tweets, Facebook, etc.), hopefully, we can pull out more information and form a complete picture, including the information of what kind of people are vulnerable to a specify vaccine. In this way, we will be able to better predict the adverse events, or even help to design new vaccine approaches that minimize or

eliminate serious vaccine-related reactions. Given current advancements in data mining, this is a revolutionary time for research in real-world applications using social network analysis.

Bibliography

[1] L. Akil, H. A. Ahmad, and R. S. Reddy. Effects of climate change on salmonella infections. *Foodborne pathogens and disease*, 11(12):974–980, 2014.

[2] P. Antwi-Agyei, E. D. Fraser, A. J. Dougill, L. C. Stringer, and E. Simelton. Mapping the vulnerability of crop production to drought in ghana using rainfall, yield and socioeconomic data. *Applied Geography*, 32(2):324 – 334, 2012.

[3] L. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *PHYSICA A*, 364:513–536, 2006.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] M. Burke, S. M. Hsiang, and E. Miguel. Climate and conflict. Technical report, National Bureau of Economic Research, 2014.

[6] R. BUSH. Food riots: Poverty, power and protest1. *Journal of Agrarian Change*, 10(1):119–129, 2010.

[7] P. H. Gleick. Water, drought, climate change, and conflict in syria. *Weather, Climate, and Society*, 6(3):331–340, 2014.

[8] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[9] C. Hendrix, S. Haggard, and B. Magaloni. 1 grievance and opportunity: Food prices, political regime, and protest, 2009.

[10] S. M. Hsiang, K. C. Meng, and M. A. Cane. Civil conflicts are associated with the global climate. *Nature*, 476(7361):438–441, 2011.

[11] IPCC. *Climate Change 2007: Climate Change Impacts, Adaptation and Vulnerability*. Cambridge University Press, 2007.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[13] S. Johnstone and J. Mazo. Global warming and the arab spring. *Survival*, 53(2):11–17, 2011.

[14] C. P. Kelley, S. Mohtadi, M. A. Cane, R. Seager, and Y. Kushnir. Climate change in the fertile crescent and implications of the recent syrian drought. *Proceedings of the National Academy of Sciences*, 112(11):3241–3246, 2015.

[15] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[16] P. Le Billon. The political ecology of war: natural resources and armed conflicts. *Political geography*, 20(5):561–584, 2001.

[17] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[18] C. C. Mitigation. Ipcc special report on renewable energy sources and climate change mitigation. 2011.

[19] P. Mörters and Y. Peres. *Brownian motion*, volume 30. Cambridge University Press, 2010.

[20] 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'14*, pages 1799–1808, 2014.

[21] H. Sarsons. Rainfall and conflict, 2011.

[22] J. Scheffran, M. Brzoska, J. Kominek, P. M. Link, and J. Schilling. Climate change and violent conflict. *Science*, 336(6083):869–871, 2012.

[23] T. Sternberg. Chinese drought, bread and the arab spring. *Applied Geography*, 34:519 – 524, 2012.

[24] K. Warner, C. Ehrhart, A. d. Sherbinin, S. Adamo, T. Chai-Onn, et al. In search of shelter: Mapping the effects of climate change on human migration and displacement. *In search of shelter: mapping the effects of climate change on human migration and displacement*, 2009.

[25] G. Wischnath and H. Buhaug. On climate variability and civil war in asia. *Climatic Change*, 122:709–721, 2014.