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

May 30, 2016
Arlington, Virginia

Keywords: Information Propagation, Absenteeism, 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) swift outbreak/slowly evolving transmission (iv) misinformation 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 abnormality in graph? Third, how can we infer the causality of a specific type of protest, say climate related protest? Fourth, how do we distinguish real movements from rumors or misinformation campaigns?

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 abnormality detection and misinformation distinction, and ultimately helps improve the understanding of mass movement and their adoptions in social networks.

Acknowledgments

I would like to acknowledge and express my gratitude to several individuals for their contributions and support through my years as a doctoral student.

My advisor, Naren Ramakrishnan, for giving me the independence to pursue my research interests. I am especially grateful for his advice and wisdom over the past four years.

My doctoral committee - Chang-Tien Lu, Yang Cao, Chris North, Feng Chen - for their valuable comments, guidance and encouragement.

Contents

1 Causality Inference to Climate Related Protest **1**

1.1 Introduction 1

1.2 Related research 3

1.3 Climate protest classifier 3

1.3.1 Majority assign 3

1.3.2 K Nearest Neighbor 4

1.3.3 Naive Bayes 4

1.3.4 Weighted Support Vector Machine 5

1.3.5 Logistic regression classifier 7

1.3.6 Evaluation 7

1.4 Climate Motivated Protests 8

1.4.1 Frequency Analysis by Country 9

1.4.2 Spatial Distribution of Climate Motivated Protests 10

1.4.3 Temporal Dependency on Climate Events 12

1.5 Climate protests causality 15

1.5.1 Word cloud 15

1.5.2 Analysis of Protest Descriptions 15

1.5.3 Pathways to Climate Motivated Protest 16

1.6 Climate protest pattern 18

1.7 Climate protests in Twitter 19

1.8 Discussion 20

List of Figures

1.1 Gold standard report (GSR) format. 2

1.2 Classification methods comparison. 8

1.3 Blue bar shows all the GSR protest events, yellow bar shows climate related protest events, green area shows the climate protest percentage over all the Latin American countries, from July 2012 to March 2015. 9

1.4 (a) Climate related protests events numbers; (b) Climate related protests percentage in Latin American countries, from July 2012 to March 2015. . . 10

1.5 Climate protest events and population (million) of each country. The two series have a Pearson correlation coefficient 0.64. 10

1.6 climate protest and non-climate protest time series from 2011 to March 2015. 11

1.7 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. 11

1.8 (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. 12

1.9 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. 13

1.10 Word cloud of Mexico storm Manuel, Sept 13, 2013. 14

1.11 Word cloud of all the climate related protests, from GSR descriptions. 15

1.12 Climate protest causality diagram. Left bar shows ten countries' climate protest numbers, and right bar shows nine climate event categories which cause climate protests. 16

1.13 Climate protest causality diagram 17

1.14 Climate protest clustering results 18

1.15 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. 20

List of Tables

1.1 Classification methods comparison. 8

Chapter 1

Causality Inference to Climate Related Protest

1.1 Introduction

Climate change, extreme weather, and the state of the environment directly impact the availability of food [1], [2], energy [14], and shelter [18]. As finite resources become scarce, the residual impacts on local economies can have disastrous and long-lasting effects on the fundamental livelihoods of inhabitants for decades [12]. 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 [6, 11]. In a smaller scale example, the environmental impact of lead contamination in the drinking water in the United States led to protests in 2016. The extreme weather event, Hurricane Manuel, that devastated the western coasts of Mexico led to subsequent protests over resources at points as long as 1 year after the initial event.

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) 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 reportings of civil unrest. The GSR includes many features, as shown in Figure 1.1, 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 1.1 shows, the protest description is 'small farmers want the bank to forgive their debts due to the drought, which has hampered production'.

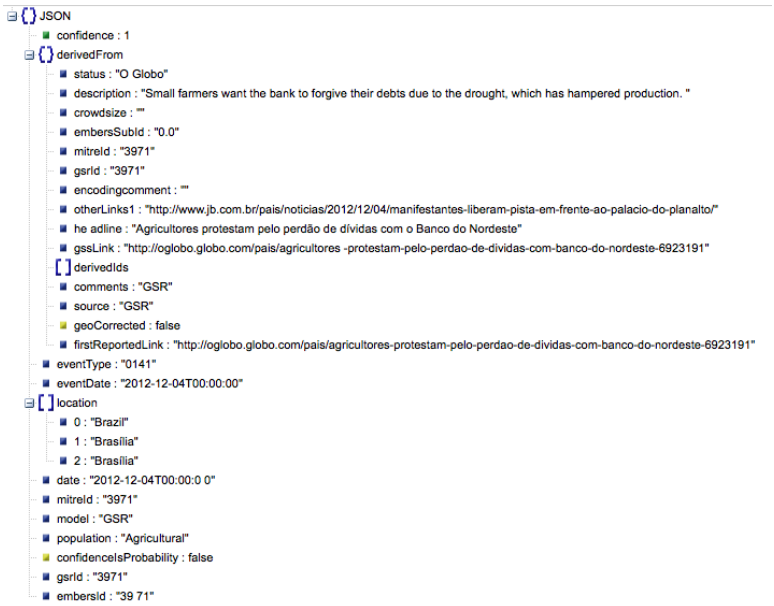


Figure 1.1: Gold standard report (GSR) format.

We address three foundational problems: first, the identification of climate related protest. 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, from analyzing large historical climate protests records, we look into the text description of protests and extract the climate protest category for each country. For each major climate category, we adopt the knowledge graph approach to define linkage relationship between entities, and study the possible protest causalities. Third, we find that the massive climate protests have coherent correlations within them, we also show that certain protest types are more prone to associate with certain other protest types, such as lack of water is highly linked with power shortage, and so on. Generally, the main contribution of this paper 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 proportion of protest causality. 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.

1.2 Related research

The path from climate, extreme weather and environmental effects to civil unrest is causally complex [8, 16] and involves various combinations of climate change [4], 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 [9]. A commonly studied pathway is the effect of climate on food prices which then induces civil unrest. An examples of this occurrence is the Arab Spring uprisings in 2011, and how weather effects food prices [10]. 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 [17]. The pathways of food prices to protest have also been studied in the global south [5], Africa, and Asia [19, 7]. However, even this path of climate effects on income level leading to conflict is not eminently clear [15].

1.3 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.

1.3.1 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 data-set 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 data-set, since $p << (1-p)$, the accuracy approximately equals to $(1-p)^2$, while the $F-measure$ is p.

1.3.2 K Nearest Neighbor

To classify a class-unknown document X, the K-Nearest Neighbor (KNN) classifier algorithm ranks the document's neighbors among the training document vectors, and uses the class labels of the k most similar neighbors to predict the class of the new document. The classes of these neighbors are weighted using the similarity of each neighbor to X, where similarity is measured by Euclidean distance or the cosine value between two document vectors [13].

KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970s as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. Choosing the optimal value for K is best done by first inspecting the data. We first manually identify 100 climate-related protest events as the training sets. In our protest filter design, text similarity measures play a fundamentally important, where apply Corpus-Based similarity for distance computation between different event descriptions. In our experiment, we set K to be 100.

1.3.3 Naive Bayes

Essentially, Naive Bayes is maximum 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 $e =< w_0, w_1, ... >$. $c = \{climateprotest, non-climateprotest\}$. 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:

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

1.3.4 Weighted Support Vector Machine

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 the 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 hperplane P for unseperable sets, it is often converted into the following quadratic convex optimization problem by defining the slack variables $\zeta = (\zeta_1, \zeta_2, .., \zeta_N)$.

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\zeta_i \\ \text{subject to} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \zeta_i, \quad \forall i \\ & \zeta_i \geq 0, \qquad\qquad\qquad\quad \forall i \end{aligned} \tag{1.1}$$

C is the penalty parameter. For separable sets, C corresponds to ∞.
The problem with above classifer is that the penalty for misclassification are the same. However, there are a lot of cases the miss alram should have a much higher cost than the false alarms. To considet those scenarios, we introduce two different penalty for miss alarm and false alarm. For simplicity, I and J denotes the subscript of positive and negative set. Thus, the problem of 1.1 can be re-formulated as:

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}||\beta||^2 + C_1\sum_{i\in I}\zeta_i + C_2\sum_{j\in J}\eta_j \\ \text{subject to} \quad & x_i^T\beta + \beta_0 \geq 1 - \zeta_i, \forall i \in I \\ & x_j^T\beta + \beta_0 \leq -1 + \eta_j, \forall j \in J \\ & \zeta_i \geq 0, \forall i \in I \\ & \eta_j \geq 0, \forall j \in J \end{aligned} \tag{1.2}$$

The Lagrange function of 1.3 is

$$
\begin{aligned}
L_p =& \frac{1}{2}||\beta||^2 + C_1 \sum_{i \in I} \zeta_i + C_2 \sum_{j \in J} \eta_j \\
& - \sum_{i \in I} \alpha_i [x_i^T \beta + \beta_0 - (1 - \zeta_i)] - \sum_{i \in I} \mu_i \zeta_i \\
& + \sum_{j \in J} \theta_j [x_j^T \beta + \beta_0 - (1 - \eta_j)] - \sum_{j \in J} \tau_j \eta_j
\end{aligned} \tag{1.3}
$$

The input of the SVM experiment is GSR json file, and the input file includes lots of features. Among these features, several of them are more important than others, including status, description, crowdsize, headline, eventType, eventDate, location, date, population etc. The description feature is brief description of the events, which plays a dominant role in the entire dataset.

In order to apply SVM algorithm on GSR dataset, we need to vectorize text data in the dataset. First of all, we construct a word corpus which includes every word shown in the dataset (including non-words). We accept non-words because most coinages come from Internet and some of them might be important for the events. As we accept non-words, the corpus might be large than our corpus vocabulary. Then, for every fields, the content is converted to a corresponding vector based on the corpus. If one word exists in specific field, a calculated value will be assign to the corresponding element in the vector. Other numerical values will also be added in the vector. By this way, we could convert the dataset into a huge matrix.

Minimizing (2) can be rewritten as a constrained optimization problem with a differentiable objective function in the following way.

For each $i \in \{1, \ldots, n\}$ we introduce the variable ζ_i, and note that $\zeta_i = \max(0, 1 - y_i(w \cdot x_i + b))$ if and only if ζ_i is the smallest nonnegative number satisfying $y_i(w \cdot x_i + b) \geq 1 - \zeta_i$.

Thus we can rewrite the optimization problem as follows

$$\text{minimize } \frac{1}{n} \sum_{i=1}^{n} \zeta_i + \lambda \|w\|^2$$

,

$$\text{subject to } y_i(x_i \cdot w + b) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0, \text{ for all } i.$$

This is called the "primal" problem.After the preprocessing, we could apply SVM algorithm on GSR matrix. We randomly label 1000 events from GSR dataset and use ten-fold cross validation to train out SVM. Figure 1 shows the result of SVM experiment.

1.3.5 Logistic regression classifier

From GSR events, we devised a climate protest classifier to identify the climate related protest events automatically. The classifier is built based on logistic regression model. With input of GSR descriptions, we aim to train a classifier which can label a protest description as climate related or not. The GSR includes many potential important features, such as status, description, crowd size, headline, event Type, event Date, location, date, population etc. The description feature is brief description of the events, which plays a dominant role in the entire dataset. In order to adopt logistic regression on GSR dataset, we need to vectorize text data in the dataset. First of all, we construct a word corpus which includes every word x_i shown in the training dataset (including non-words). We accept non-words because most coinages come from Internet and some of them might be important for the events. As we accept non-words, the corpus might be large than our corpus vocabulary. The word corpus is composed with $[x_1, x_2, ..., x_i, ..., x_N]$. Second, take each GSR description as a vector, we assign values to each vector, if x_i appeared in GSR record, the corresponding value will be assigned as 1, otherwise 0. In this way, every GSR record being converted to a corresponding vector based on the corpus. Third, set climate protest as $Y = 1$, non-climate protest as $Y = 0$, the weight for each term x_i as k_i, then $Y_j = \sum_{i=1}^{N} k_i x_i$. By training process, we calculate the weight k_i for each term x_i. The last step is test. Given a new GSR description, the probability of classification is:

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

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

1.3.6 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 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). From Table 1.1, we prove logistic regression method outperforms other methods uniformly.

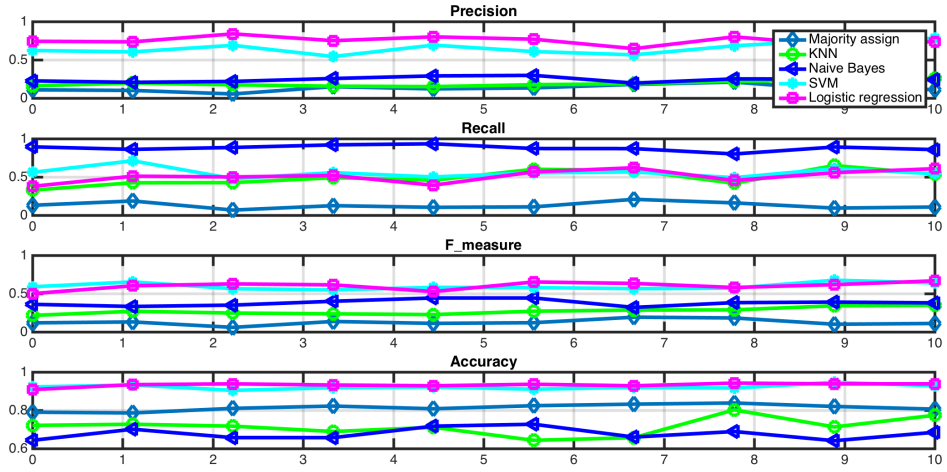


Figure 1.2: Classification methods comparison.

Table 1.1: 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**

1.4 Climate Motivated Protests

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.

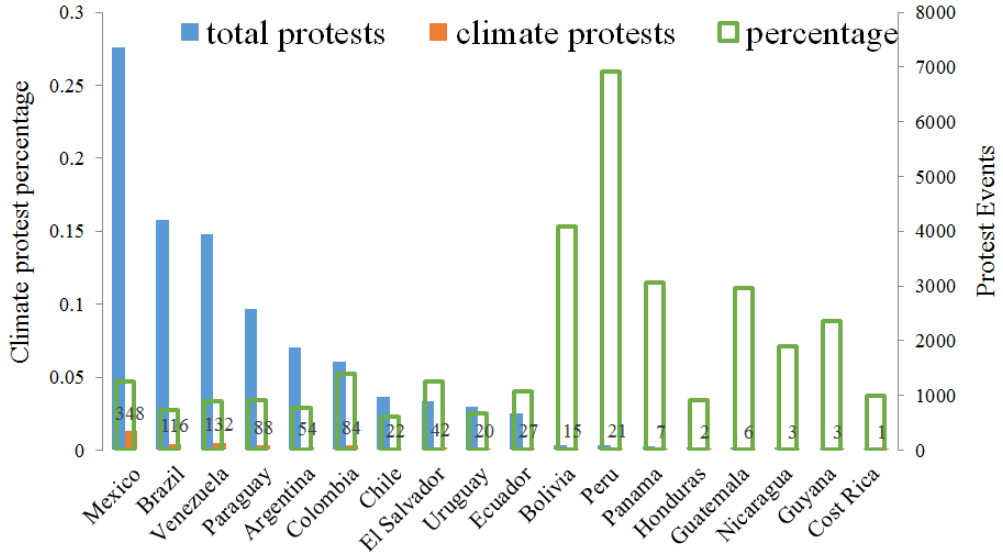
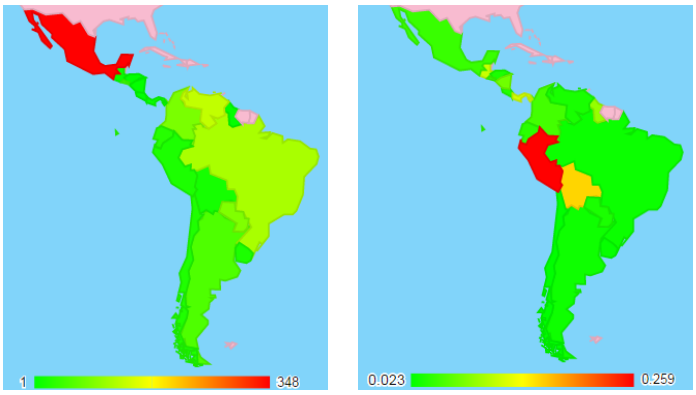


Figure 1.3: Blue bar shows all the GSR protest events, yellow bar shows climate related protest events, green area shows the climate protest percentage over all the Latin American countries, from July 2012 to March 2015.

1.4.1 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 total number of protests and the percentage of those that are climate motivated in Figure 1.3. The country with the most protests overall is Mexico, and Costa Rica has the least. A similar trend is also seen in terms of climate motivated protests. 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 1.5. 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 or anecdotal studies. In the following, however, we choose Brazil, Mexico, and Venezuela for



(a) (b)

Figure 1.4: (a) Climate related protests events numbers; (b) Climate related protests percentage in Latin American countries, from July 2012 to March 2015.

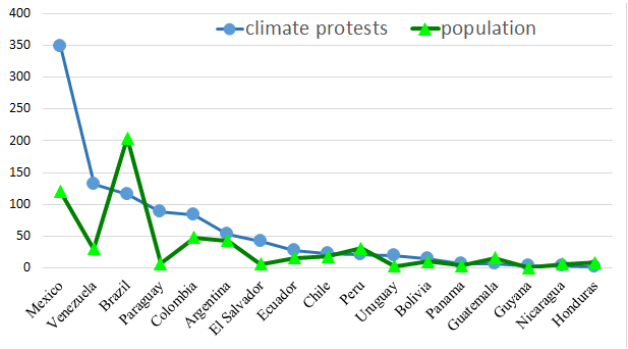


Figure 1.5: Climate protest events and population (million) of each country. The two series have a Pearson correlation coefficient 0.64.

further analysis into overall trends of climate protests, and how these are shaped in the data recovered by the classifier.

We investigate the protest event time series in South American. As shown in Figure 1.6, on average, February, June, July, and August see the most climate-related protest events. We can also see the for the none-climate protest, the temporal distribution is different since it see most protests in March.

1.4.2 Spatial Distribution of Climate Motivated 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

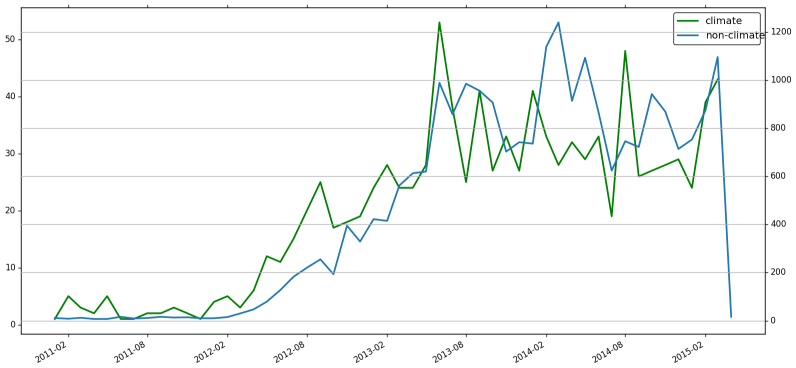
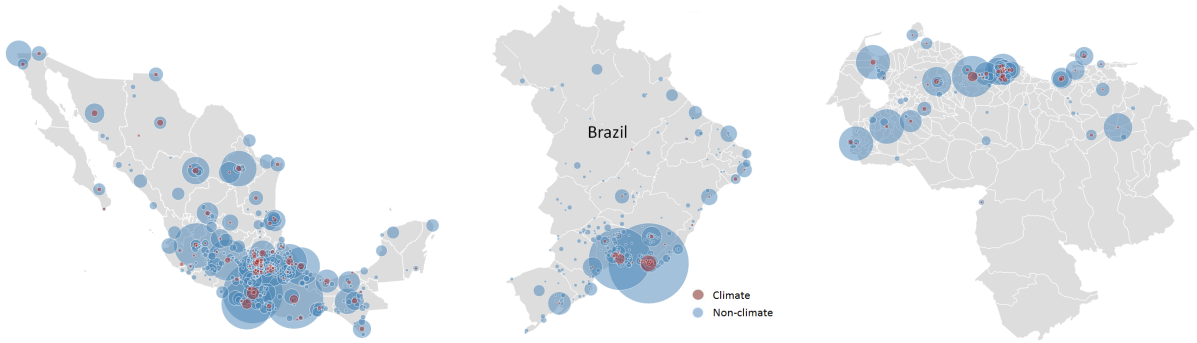


Figure 1.6: climate protest and non-climate protest time series from 2011 to March 2015.

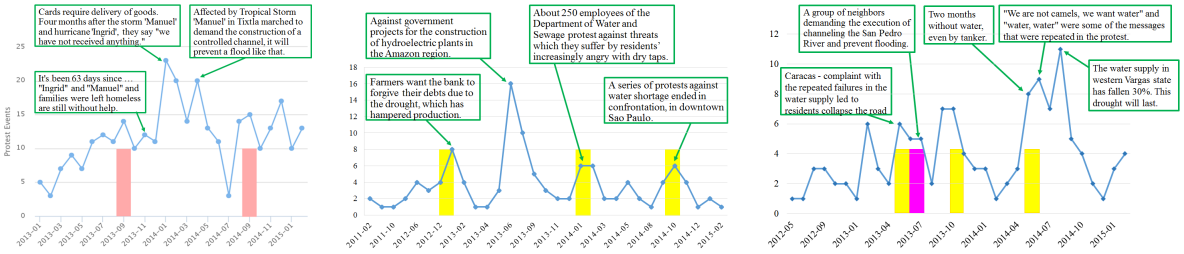


(a) Mexico (b) Brazil (c) Venezuela

Figure 1.7: 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.

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 1.7. 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



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

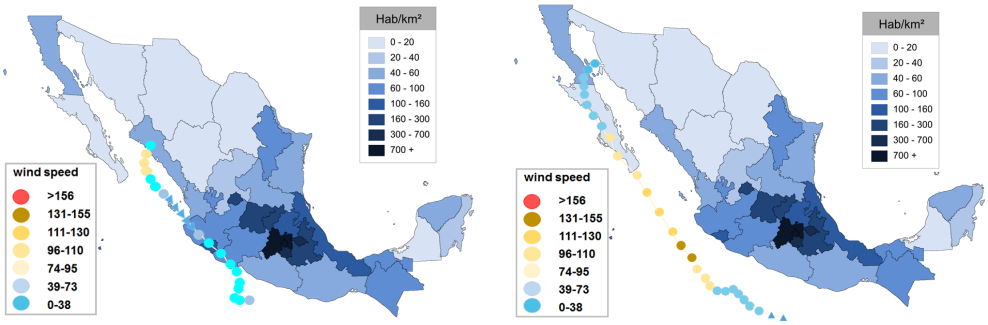
Figure 1.8: (a) Mexico climate disasters and climate protests. The blue time series shows the climate related protest events, and light red vertical lines show two storm diasters in Mexico, storm Manuel in September 17, 2013 and hurricane Odile in September 15, 2014 respectively. (b)Brazil climate disasters 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.

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, 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.

1.4.3 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[1], World Disasters Timeline[2] and European Commission's Humanitarian Aid and Civil Protection department (ECHO)[3]. The official climate disaster report for each country is shown with climate related protests in Figure 1.8.

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



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

Figure 1.9: 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.

Mexico climate disasters Figure 1.8(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 1.9(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[4]. 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"[5] 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 1.8(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 1.9(b). Despite hurricane Odile being a more intense storm, there were not many protests related to the

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

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

Figure 1.10: Word cloud of Mexico storm Manuel, Sept 13, 2013.

event. Comparing the storm's paths in Figure 1.9, 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 1.8(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[6] 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.

Venezuela climate disasters In Figure 1.8(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[7]. 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

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

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

there is a substantial amount of overlap in the residual protest events that it was difficult to distinguish to which it was referring.

1.5 Climate protests causality

Figure 1.11: Word cloud of all the climate related protests, from GSR descriptions.

1.5.1 Word cloud

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 1.11. 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.

1.5.2 Analysis of Protest Descriptions

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 1.12 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

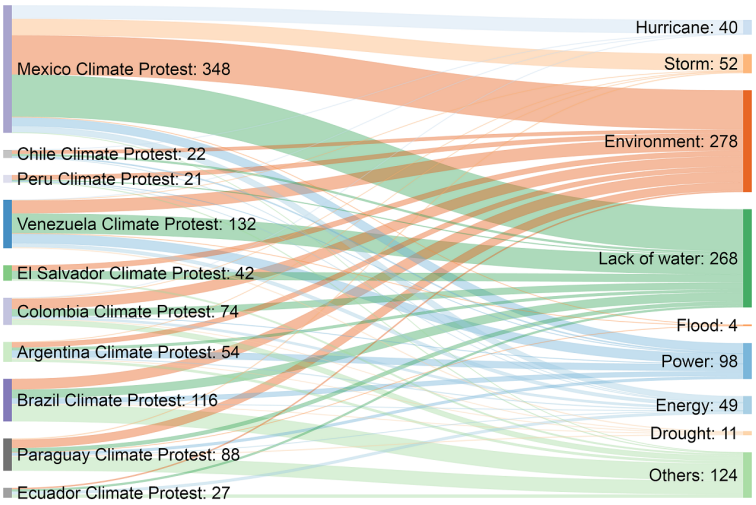


Figure 1.12: Climate protest causality diagram. Left bar shows ten countries' climate protest numbers, and right bar shows nine climate event categories which cause climate protests.

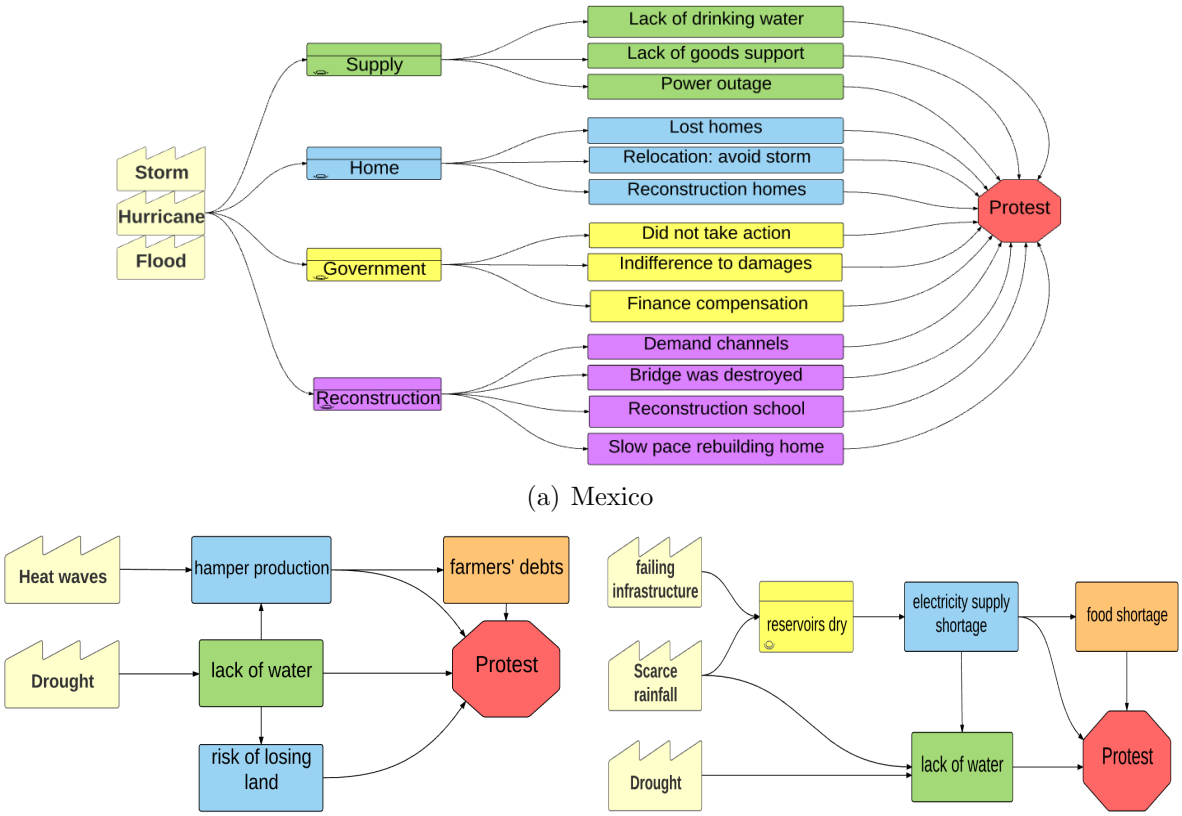
keywords include mentions of power and energy issues. Each country also exhibits its own protest keyword categories. In Mexico, the most notable protest keywords involve are lack of water, environmental concern, storm and hurricane. In Venezuela, apart from lack of water and environment problems, the 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.

1.5.3 Pathways to Climate Motivated Protest

To determine the different pathways to climate motivated protest events, a three step process was implemented to extract the relevant information.

a. Text enrichment. Messages with textual content (Tweets, Newsfeeds, Blog postings, etc.) are subjected to shallow linguistic processing prior to analysis. Applying BASIS technologies' Rosette Language Processing (RLP) tools, the language of the text is identified, the natural language content is tokenized and lemmatized and the named entities identified and classified. Finally, messages are geocoded with a specification of the location (city, state, country), being talked about in the message.

b. Knowledge graph construction. In the scope of this study, an entity network is a graph $G(E, R)$ where entities $E = e_1, ..., e_n$ can be linked to one another through relationships $R = r_1, ..., r_n$ defined by conceptual interactions, and thus called a knowledge graph. knowledge graphs are used to represent its logic based on semantic networks. How to repre-



(a) Mexico

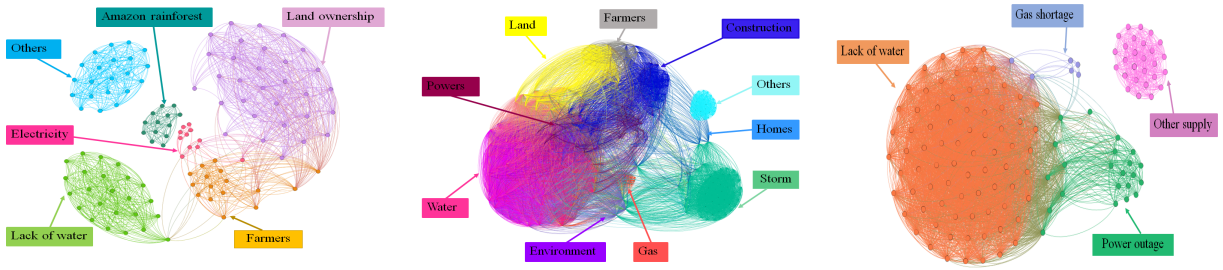
(b) Brazil

(c) Venezuela

Figure 1.13: Climate protest causality diagram

sent knowledge and extract entities and their relationship from text? The general extraction patterns rule is Noun Phrase - Verb Phrase - Noun Phrase. We can also employ multilingual rules for phrase patterns learnt using association mining based on collocations. Using an open source, GraphDb (Neo4j) to store and query knowledge triples. The general scheme is Subject - Predicate - Object. Subject is a concept, which can be person, organization, location, event, noun phrase; Predicate can be either a verb action or (predefined) relation type; Object can be either a concept or (some) value. With all the enriched articles as input, since their named entities and verb phrases being identified and classified, the system can automatically extract the structural knowledge and load them into Graph database. It worths to note the entities may evolve along space and time.

c. Causality query. With a complete knowledge database, we are going to poses domain specific questions, for example, who are the main players in the article? what are the reasons for the protest? To such questions like that, we are seeking causal (base) relations between concepts. Specifically, we are interested in entities to be storm, hurricane, heat, draught, etc,, climate related keywords. 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,



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

Figure 1.14: Climate protest clustering results

demanding, against, request, we can locate and further identify most of protest causalities.

Figure 1.13(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 1.13(b) describes the causality 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 1.14(c). Scarce rainfall, drought plus failing infrastructure, which makes water shortage and blackout is an everyday 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.

1.6 Climate protest pattern

The above analysis shows the general protest causality, however, we intend to further discover the coherent correlations of protest reasons, hoping to answer questions like is there any protest pattern, or are there some protests associate with others. We treat each protest event as a node and connect two nodes with weight based on their protest description text similarity. Specially, we pay attention to the protest themes or protest demands, if two descriptions have the same protest demanding, 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 [3] to split the network into several clusters.

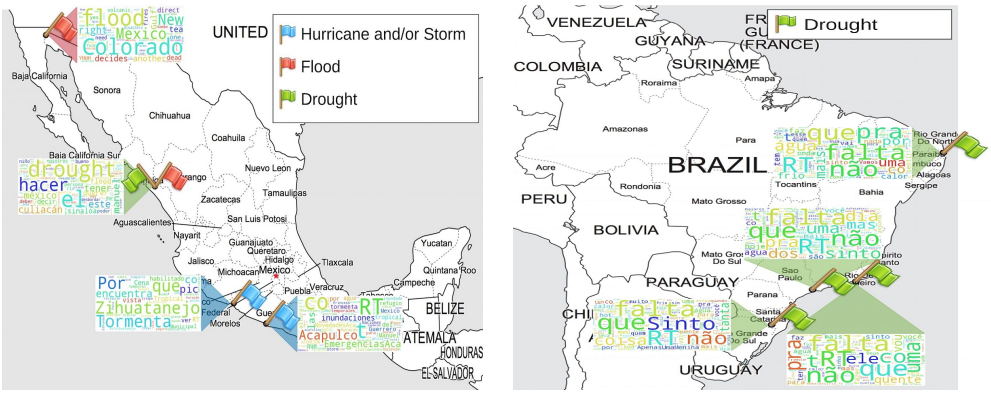
We show in Figure 1.14 the climate protest clustering results that provides the protest proportion and coherent correlations among different protest types. Figure 1.14(a) illustrates Brazil's climate protest pattern, the results shows that in Brazil, the largest protest cluster is about land ownership which accounts for 26.7%, the second cluster lack of water takes up 20.7%, farmers cluster occupies 13.8%, of which, one interesting discover is land and farmers clusters are closely coherent, and lack of water is also closely bind with farmers. Amazon rainforest is another striking protest which is responsible for 11.3%. Figure 1.14(b) shows the protest pattern of Mexico, which has the most climate protest events and complex patterns. We can see the rose red cluster which denotes lack of water is the most dominant protest, accounts for 20.5%, the green cluster represents tropical storm is the second largest protest type, takes up 19.0%, the dark blue cluster construction accounts for 17%, and the yellow cluster land is responsible for 11.8%. We find, water protest is intertwined with environment protest and power protest, land protest is closely related with farmers, while construction cluster is coherent with baby blue cluster which denotes homes (2.6%). Figure 1.14(c) gives the overview clustering results of Venezuela climate protests. We can see, the yellow cluster which represents lack of water protests takes up the largest portion, as high as 55.8%, the green cluster which denotes power outage accounts for second part, 22.1%, and the blue cluster which stands for gas shortage accounts for 5%, the purple cluster shows the rest climate protest portion, which include food shortage, medicine shortage, water tank robbery behavior, etc.. Clearly, as expected, 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.

1.7 Climate protests in Twitter

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 1.15(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 1.15(b) illustrates four drought events in Brazil, on May 2012.



(a) Mexico (b) Brazil

Figure 1.15: 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.

1.8 Discussion

Climate changes, extreme weather and environmental catastrophes can all exert a devastating amount of harm to people around the world. To better understand this process, 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 diasters may cause related protests with different time span, for instance, the Mexico storm Manuel aroused climate related protest as long as 17 months, while in Venezuela, the protests caused by one drought always overlap with the other drought. This paper discloses protest causalities in Latin American countries, illustrate the pathways from climate disasters to climate protests. This paper also identifies the climate related protest patterns, discover the coherent relationship among different protests demanding, such as in Venezuela, the majority protests are against lacking of water, which has high co-occurrence with protests against power outage.

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] 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.

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

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

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

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

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

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

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

[11] 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.

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

[13] Y. Liao and V. R. Vemuri. Using text categorization techniques for intrusion detection. In *USENIX Security Symposium*, volume 12, pages 51–59, 2002.

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

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

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

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

[18] 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.

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