Algorithms for Modeling Mass Movements and their Adoption in Social Media

1. **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 media.

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.

1. **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 5. 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.

1. **Introduction**

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 phenomenon from social media.

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 activities 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?

**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?

1. **Conclusions and future directions**

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 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 movements 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 anomaly 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 anomaly (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 2. Another benefit that enhanced the mass movement detection is though group anomaly detection approach, as introduced in chapter 3. 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 5.

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.

**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 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 anomaly, such attribute can be adopted into detecting the wealth gap between rich and poor in one region, identifying the brain neural network anomaly, 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**

Instead 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 advertisement 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 disasters, 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 from 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, such as 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.