Algorithms for Modeling Mass Movements and their Adoption in Social Media

1. **Abstract**

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

This dissertation explores four research problems underlying efforts to identify and track the adoption of mass movements in social media. First, how do mass movements become mobilized on Twitter, especially in a specific geographic area? Second, can we detect protest activity in social networks by observing group anomalies in graph? Third, how can we distinguish real movements from rumors or misinformation campaigns? and 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 conduct a comprehensive study of how mass movement adoption functions in social networks. For example, it may cross multiple spaces, evolve with dynamic network structures, or consist of swift outbreaks or long term slowly evolving transmissions. In many cases, it may also be mixed with misinformation campaigns, either deliberate or in the form of rumors. Each of those issues requires the development of new mathematical models and algorithmic approaches such as those explored here. It is my hope that this work will facilitate advances in information propagation, group anomaly detection and misinformation distinction and, ultimately, help to improve our understanding of mass movements and their adoption in social networks.

1. **Acknowledgments**

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

My deepest debt is to my advisor, Dr. Naren Ramakrishnan. I have been amazingly fortunate to have an advisor who has given me the freedom to explore on my own, while at the same time pulling me back on track when I strayed too far afield. His far-sighted research attitude and the respect he has earned from his peers over the years have made the Discovery Analytics Center he leads a collaborative and productive place. His insight, wisdom and humor have made my graduate study a rich and rewarding journey that I will cherish forever.

I am especially grateful to Dr. Chang-Tien Lu, one of the best teachers that I have ever had. He introduced me to data analytics, opening up a whole new world of research and ideas to me. He has also provided unstinting support and encouragement throughout my search 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 research.

I appreciate the efforts of Dr. Yang Cao, for his encouragement and practical advice during the work reported in Chapter 5. He has also spared no effort in supporting my job hunt and his belief in me and what I can achieve has lifted my spirits when I became discouraged.

I would like to say thank you to Dr. Chris North for his support, feedback, and many valuable discussions that have helped me understand my research area better. It has always been a pleasure to get to know such a knowledgeable and amiable professor.

I would like to thank Dr. Huzefa Rangwala, who has always been generous with advice and ready to provide help. I hope one day I will become as good an advisor to others as Dr. Rangwala has been to me.

I am especially indebted to all the members of Discovery Analytics Center. Thank you for making our lab group 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 hope he knows how much I appreciate the way he has walked alongside me through hardship, challenge and setbacks. His faith that I will succeed has kept me going. Thank you.

1. **Introduction**

As social media gains popularity in societies around the world, more and more mass movements are beginning to be organized via social media, with their slogans becoming hashtags in Twitter. In this context, “mass movements” are more akin to social movements supported by large groups of individuals or organizations that focus on specific political or social issues. The research conducted for this dissertation studied mass movement adoption patterns and their related subsequent phenomena from social media.

This dissertation consists of an exploration of four research problems underlying mass movement adoption in social media. In contrast to popular memes, these constitute modeling the process of protest mobilization, detecting graph group anomaly patterns, inferring protest causality, and distinguishing real movements from rumors. (i) How are mass movements mobilized on Twitter, especially those within a specific geographic area? (ii) How can we detect protest activities in social networks by observing group anomalies in graphs? (iii) How can we distinguish real movements from rumors or misinformation campaigns? and (iv) How can we infer the pathways of climate related protests?

**Group anomalies in graphs**

Group anomalies not only depend on each user's activity, but are also closely associated with a graph’s structure. In recent year, a significant body of research on group anomaly has focused primarily on two aspects of this problem: (1) modeling users’ behaviors to define group anomalies, paying little or no attention to the underlying network structure; and (2) defining groups at a local scale by imposing distance-based restrictions such as distance, radius, or even node numbers while failing to consider a more global perspective, as nodes located some distance away may still be highly associated. The work reported here looks at global level group anomalies but sets no restrictions on the group definition, considering both the users' behavior and the underlying graph structure. Investigating this phenomenon of broad group anomalous online behavior 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 will inevitably 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. Identifying the pattern involved in the evolution of a climate related protest could provide valuable information regarding the factors involved in a situation escalating from a climate disaster to become an armed protest.

1. **Conclusions and future directions**

As social media such as Twitter continue to increase in popularity, they are effectively becoming social sensors that can be utilized for real-world mass movement event detection. Modeling and studying the related adoption patterns provide new insights for investigating both the social and physical aspects of these events and their precursors. This dissertation has presented several approaches and strategies aimed at detecting and predicting mass movements and further inferring their causality, given information composed of a volatile mixture of real news and rumors. Techniques have been presented that are designed to capture information propagation across multiple spaces, as well as a new graph wavelet approach that broadens predictive capabilities to capture group anomalies within networks. A number of different types of mass movements have been investigated and diverse aspects of modeling addressed.

Using social media as indicators for real-word event detection can indeed be a helpful tool, but several limitations apply, perhaps most notably when applied to a specific event type such as the mass movements studied here. First, modeling protest-related topic propagation across networks is never trivial. One challenge is that social protest propagation through online media can spread over large areas far more quickly than traditional methods since users are geographically distributed, while other challenges include the way mass protest information can be spread by multiple social media types and paths, such as word of mouth, and radio and TV news broadcasts. Second, even detecting group anomalies on social media is challenging. For example, Twitter's user network embodies many subgraphs based on social ties with dynamic graph structures that are constantly changing since users are actively contributing content. The other challenge includes real world events that are not only correlated with burst signals, but can also exhibit unusually low levels of activity in social networks. Despite these restrictions, graph wavelets have in fact proven to offer a powerful capacity for capturing graph anomalies (in terms of both burst behavior and absenteeism behavior), even on dynamic changing networks.

A fundamental objective of this research has been to model mass movement adoption behavior, and in doing so, several significant additional advantages have been gained. One contribution is the ability to model information propagation across multiple networks/spaces, and capture the propagation speed and possible propagation paths, as demonstrated in Chapter 2. Another benefit gained through enhancing the mass movement detection capability is the opportunity this provided to introduce a new group anomaly detection approach, as described in Chapter 3. The use of graph wavelets made it possible to develop a more appropriate definition of group anomaly that covers both bursts and absenteeism utilizing different scales, thereby increasing the probability of capturing protest behaviors. Another benefit is the ability to quantify compartment transition dynamics using the epidemic model SEIZ, thus facilitating the development of screening criteria that can distinguish real movements from rumors on Twitter, as demonstrated in Chapter 5.

Understanding information propagation over dynamic social networks is becoming a highly popular way to address real-world problems in social network analysis. The research reported in this dissertation analyzed several fundamental questions that underlie a broad range of propagation-like processes, focusing particularly on mass movement adoption and rumor transmission. These methodologies can easily be extended to applications such as infectious diseases, public health and marketing, among others.

**Future Directions**

Given the wide scope of this relatively new field, there are many opportunities for conducting potential fruitful research. For example, one very attractive area would be to continue to focus on social network analysis, specifically research into information propagation over dynamic rapidly changing social networks. Here, future research directions fall into two general categories, with one being to deepen the existing theories and algorithms by exploring them in more detail while the other is to broaden the current research by adopting a wider perspective. Examples of these and other research possibilities are presented below.

**Extend GBM model**

What happens to the geometric Brownian motion model if the underlying mention network changes over time? How can this this model be adopted or modified to apply to multiple networks? In addition to these theoretical questions, a number of practical applications are worth further investigation, for example would it be possible to introduce the GBM model into the infectious disease domain to model the spread of the zika virus? Assuming that bispace is composed of both a connection network and physical space, can we train the GBM model to estimate individuals’ infection probability based on their environment?

**Further studies involving graph wavelets**

It should be possible to extend the graph wavelet applications presented in Chapter 4 into other areas based on the two of the major distinguishing properties of graph wavelets. The ability to detect graph anomalies could be utilized to detect the wealth gap between rich and poor in a given region, identify brain neural network anomalies, or detect traffic congestion through road network analysis. Similarly, the ability to identify the central point of a subgraph could be employed to rank key players within networks, detect those spreading rumors in some cascade, or find the sources of infection for certain diseases.

**Broaden rumor detection scenarios**

Instead of determining whether a particular story is true or false, it may instead be more practical to predict how likely people are to believe it. Newspapers in particular would find it very useful, especially when it comes to a breaking news story that has not yet been officially confirmed. Before reporting to the public, they need to know how believable the story is. Also this offers a valuable way to decide whether vendors are cheating during online shopping.

**Deepen our understanding of personalized information propagation**

It would be useful to know how users' behavior either delays or boosts information propagation, especially when accompanied by strong sentiments. This may help formulate better advertisement strategies if a better way to manipulate the information flow can be identified. Further, it would be interesting to explore opportunities to extract personalized information spread patterns. What kind of news arouse the interest of individual consumers? What kind of role is he or she likely to play? and What kind of push strategy might stimulate their activity? This type of study would be invaluable for precision marketing or personalized recommendations, providing refined content filtering.

**Build an intelligent disaster detection system**

Another interesting option would be to build an intelligent system that is efficient at event detection, especially for disasters, protests, and other extraordinary events. Such a system would not only be capable of immediate reporting, but also able to track events and perform causality analyses and even provide future predictions. This type of system would require several critical building blocks: natural disaster detection could be based on the use of graph wavelets for detecting group anomalies, rumors or news detection could employ an epidemic SEIZ model, story causality analysis and event coding. Consider the Flint water crisis for example. First, it would need to identify the water crisis events from a social media analysis and confirm it to be a true story, after which it would trace all the relevant historical news to identify the causality, leading finally to the generation of a complete report using event coding.

**Combine social network analysis with physical data**

The advent of social media provides unprecedented opportunities to access vast quantities of information that could potentially benefit our research. This opens up new possibilities for reinvigorating traditional research in many areas, hopefully gaining new insights and leading to extraordinary discoveries. Take vaccines and the study of their potentially adverse effects, for example. Traditionally, vaccine research depends heavily on the raw data collected by the CDC, hospitals, patient reports, and vaccine adverse event reports. However, this data frequently suffers from problems such as lengthy time delays and incomplete information. Worse, most of the data is considered in isolation. If we can find a way to combine conventional statistical data with social media data such as tweets and Facebook posts, it may be possible to extract more information and create a complete picture showing what kind of people likely to be more vulnerable to a specific vaccine. In this way, it may be possible to better predict adverse events or even contribute to the design of new vaccine approaches that minimize or eliminate serious vaccine-related reactions. Given current advances in data mining, this is a revolutionary time for research in real-world applications using social network analysis.