# THE DYNAMICS OF INFORMATION DIFFUSION ON ON-LINE SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Shaomei Wu

August 2012

THE DYNAMICS OF INFORMATION DIFFUSION ON ON-LINE SOCIAL NETWORKS

Shaomei Wu, Ph.D.

Cornell University 2012

Your abstract goes here. Make sure it sits inside the brackets. If not, your bios-ketch page may not be roman numeral iii, as required by the graduate school.

## BIOGRAPHICAL SKETCH

Your biosketch goes here. Make sure it sits inside the brackets.

This document is dedicated to all Cornell graduate students.

## ACKNOWLEDGEMENTS

Your acknowledgements go here. Make sure it sits inside the brackets.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

CHAPTER 1

# INTRODUCTION

Understanding how information spreads in the society has attracted a growing interest from both practitioners and scholars. It is an essential element for many interesting problems, such as the diffusion of innovation, formation of public opinion, product adoption, and viral marketing.

Historically, one of the biggest challenges in the study of information diffusion is data collection. Given the difficulty of observing and tracing the diffusion process over all possible channels, most empirical studies in this field have been conducted manually by sociologists and communication researchers, through in-person interviews or standardized surveys of population samples. Therefore, these studies are largely confined by the scale and the data accuracy. As a result, existing theoretical models in the literature have relied on many assumptions about the underlying diffusion mechanism instead of empirical evidence. Being widely applied, these models are, however, hard to be verified or rebutted empirically on a large scale.

In recent years, the abundance of digital records of online interactions has provided us both the explicit network structure and detailed dynamics, supporting global-scale, quantitative study of diffusion in the real world. Using these large scale datasets collected from social media sites (i.e. Twitter, Orkut), this thesis addresses the following three questions: "who influences whom?", "how do different types of information spread?", and "how does the network structure impact the diffusion process?"

## 1.1 The Influencer Problem

Can we estimate person influence and pick out the influencers? Think of influence in context:

1. general influence: mass media
2. domain influence: masspersonal (celebrities, bloggers, organizations)
3. personal influence: tie strength, content, language, timing (accidental influencers)

### 1.1.1 Life of information in Social Media

How long should we expect a piece of information to live in the social media space [**?**]? Who are involved in each stage (production, circulation, and consumption) of lifecycle of information?

### 1.1.2 Modeling personal influence

### 1.1.3 The intersection between content and people

## 1.2 The role of content in diffusion process

Section 2 text.

### 1.2.1 Subsection heading goes here

Subsection 1 text

#### Subsubsection 1 heading goes here

Subsubsection 1 text

#### Subsubsection 2 heading goes here

Subsubsection 2 text

## 1.3 Diffusion and network structure

Section 3 text. The dielectric constant at the air-metal interface determines the resonance shift as absorption or capture occurs.

$$k_1 = \frac{\omega}{c(1/\varepsilon_m + 1/\varepsilon_i)^{1/2}} = k_2 = \frac{\omega sin(\theta)\varepsilon_{air}^{1/2}}{c} \tag{1.1}$$

where $\omega$ is the frequency of the plasmon, $c$ is the speed of light, $\varepsilon_m$ is the dielectric constant of the metal, $\varepsilon_i$ is the dielectric constant of neighboring insulator, and $\varepsilon_{air}$ is the dielectric constant of air.

CHAPTER 2

# BACKGROUND

A variety of methods have been applied by scholars to study the diffusion of information. From the theoretic perspective, sociologists and economists developed agent-based modeling (ABM) to explore the dynamics of diffusion in different networks and interactions [Centola and Macy 2007, Watts and Dodds 2007]. Computer scientists designed the algorithm to maximize the extent of diffusion by seeding the diffusion with specially-picked individuals [Liben-Nowell et al 2008].

On the other hand, modeling and predicting the propensity of real world diffusion going viral through WOM process is of central interest in recent empirical studies. Built on top of the cascade model, Gruhl et al [2004] tracked the diffusion of topics in blogspace to estimate the transmission probability with an expectation-maximization(EM)-like algorithm. Leskovec et al [2007] studied interpersonal recommendations on an e-commerce site to infer the adoption probability based on the category of products and invitation history. Cha et al [2008] studied the viral process of photos being marked as favorite on the Flickr social network. By applying the SIR model with infinite recovery time, they estimated the reproduction number with the mean degree of nodes at each step of contagion. Backstrom et al. [2006] modeled the probability of a user joining a community and the growth of communities on LiveJournal using decision trees with network structure features.

## 2.1 Social theories

## 2.2 Diffusion models

## 2.3 Temporal analysis

### 2.3.1 Temporal pattern detection

### 2.3.2 Trend detection

## 2.4 Lingusitic analysis

## 2.5 Collecting data from the web

## 2.6 MapReduce and parallel network algorithms

Introduction to MapReduce.

MapReduce is good with easy-parallelible tasks, hard for tasks that need certain global information (e.g., shortest path). However, many network problems are the second case.

Methods to convert a global problem to a series of mapreduce jobs [**?**].

CHAPTER 3

# THE INTERACTION BETWEEN PEOPLE AND INFORMATION

In [**?**], we studied the production, flow, and consumption of information in Twitter. As suggested in previous research in public communications, we classified users into 5 categories (celebrities, bloggers, mass media, organizations, and others) and found a striking concentration of attention on a small number of "elite" users on Twitter, as well as a significant homophily within categories. We also applied the classical "two-step flow" theory of communications in the context of social media sites such as Twitter. Our results confirmed that there are a large number of intermediary users who actively filter and disseminate information from media to the masses, and the composition of intermediaries is highly diverse. We also examined the lifespan and content of URLs broadcasted by different categories of users. We found that although content picked up by bloggers tends to stimulate a more persistent interest, the longevity of information is determined not by diffusion process, but by many different users independently rediscovering the same content.

CHAPTER 4

# THE ROLE OF CONTENT

Following up our previous work, in [**?**], we studied the relationship between content and the temporal dynamics of information on Twitter, focusing on the persistence of information. Our results demonstrated a strong association between the content and the temporal dynamics of information. For example, rapidly-fading information contains significantly more words related to negative emotion, actions, and more complicated cognitive processes, whereas persistent information contains more words related to positive emotion, leisure, and lifestyle.

CHAPTER 5

# NETWORK STRUCTURE

Arrival and Departure Dynamics in Online Social Networks (submitted to the International Conference on Web Wide Web, 2012). In this paper, we studied the dynamics of user arrival and departure in online social networks. We showed the network effect of the departure of friends on a user's tendency to leave a network. We also built machine learning models to predict the departure of users based on their local network properties.

CHAPTER 6

# SOCIAL MEDIA IN ARAB SPRING MOVEMENT

Joining social media data with real world events, we are able to study one of the most interesting (and also the most difficult) parts in media communication research (Lasswell's maxim): the effect of information. One of my ongoing projects is to study the role of social media in social movements, in order to understand how the propagation of information is leading or reflecting societal changes. We have collected a large number of tweets and twitter networks related to big social movements (i.e., Middle East Revolution, Occupy Wall-Street Movement). Using effective algorithms for community detection, hub detection, trend detection, and opinion mining, we will be able to identify the informal structure of massive communication networks for social movements and study the diffusion of ideology and behaviors within and across organizational/geographical boundaries.

## 6.1 Data

To collect a substantial set of users and their tweets from the Middle East area in the period of the recent social movements, we first identified a set of countries of interest, including Tunisia, Egypt, Libya, Bahrain, Iran, Iraq, Israel, Algeria, Morocco, Saudi Arabia, Kuwait, Yemen, United Arab Emirates, Palenstine, Quatar, Oman, Jordan, Cyprus, Syria, and Lebanon. For each country, we used Yahoo Maps APIs to get the list of cities and towns in that country, together with the geographical centroid point for each city/town, in the form of (latitude, longitude).

After we had the centroid points of cities/towns within a country, we used Twitter search APIs to retrieve all the recent tweets generated within 100 miles from every centroid point in that country. We then parsed these tweets and extracted the authors of these tweets.

Using these authors as "seeds", we crawled one degree out from the seeds, and retrieved the profiles of all the seeds, and all the neighbors (friends/followers) of the seeds. The size of graph grows rapidly in one-degree distance. In fact, the network induced by the seeds and their one-degree neighbors already cover over 3 millions distinct Twitter users.

We crawled the profiles of these 3M users, and tried to identify their country of origin in three ways:

1. look for the country name in their self-reported location in their profile;
2. if the time-zone city is specified in their profile, map the city to the corresponding country;
3. if the location-tracking service is turned on, get the tracked location in Twitter meta data.

After parsing the profiles for all 5M users, we were able to identify the country of origin for 260K of them.

In the end, we crawled the maximal available history of tweets generated by these $260K$ users, which is, up to 3200 tweets per user. In the end we collected in total $96,350,865$ tweets in this way. Among them, 36,857,387 were generated between Dec 1st, 2010 and March 31st, 2011, by 112,661 users from the countries listed above.

Here is a breakdown of the amount of data we collected from each country.

To compare the diffusion of protest and non-protest content on Twitter, we first identify protest-related tweets. We say a tweet is related to protest if it contains at least one protest hashtag . Protest hashtags are hand-picked by political scientists. However, as there are hundreds of thousands of hashtags in our dataset, it is not feasible for political scientists to manually label all the hashtags. To effectively identify protest-related hashtags while maintaining a high recall, we narrow the pool of hashtags to be examined based on two metrics: (1) the volume of tweets containing the hashtag; and (2) the bursty-ness (as defined by Kleinberg 2004 KDD) of the hashtag occurrence. In this way, we narrow the scope down to only the top 1000 most frequently used hashtags and the top 1000 most bursty hashtags. We then have the experts to only go through those 2000 hashtags, and are able to identify about 500 protest-related hashtags among them.

## 6.2 Method and Results

As shown in the previous section, there had been a substantial amount of protest content introduced by Egyptian users on Twitter, even before Jan 25, 2011, when the first big protests took place in Tahrir Square, Cairo. Who were V those foresighted users? Were they planning and organizing the protests? Were they qualitatively different than other users on Twitter? In this section, we will investigate these questions, focusing on the relationship between the status of users and their earliness at participating in the protest activities on Twitter.

To start, we first represent the earliness of a user by his mobilization day.

A user u's mobilization day d(u), is defined as the day when u first used any protest hashtag. We then quantify the status of user u on day t, by the number of Twitter followers u has on day t. In order to show the aggregated status of users who started to participate in the protest at different times, we group users by their mobilization day d, and calculate f(d), the median value of user status, for each group. In Figure 4, we plot f(d) for d between December 10, 2010 and January 25, 2011. Here we can see a clear trend of decreasing status as the mobilization day gets closer to the actual protest day.

## 6.3 Conclusion

By analyzing Twitter activity in Middle East area during the Arab Spring movement, we have shown that social media were used to both activate and reflect the on-goings of Middle East social movement. The relative weights of these two roles differed across countries. In particular, Egyptian users actively used Twitter to plan protests and call for a critical mass, and the users from Saudi Arabia or UAE mostly used Twitter to support or comment on on-going events. We also found that protest content travelled directionally from the central to the peripheral of the Twitter network: most protest memes were initiated by hub users and later picked up by the masses. At the individual level, we found that the adoption of protest content can be modeled by the complex contagion process - while the overall adoption rate of protest content is relatively low, people become significantly more likely to start tweeting about the protest when more than 2 friends already doing so.

Although our work is to our best knowledge the largest study of the role of

social media in social movements, we have to acknowledge that our dataset is rather disproportionate: 80% of the tweets we studied came from only 5 Middle East countries. Due to technical issues, we were not able to collect an equally large number of tweets from countries such as Libya, Tunisia, and Algeria, when dramatic societal changes were taking places in these countries.

For the future work, we want to extend our study to the diffusion of protest content among countries and communities through social media. Another interesting direction is to understand how mass media (newspaper, TV, radio) and social media interact and influence each other in social movements.

This work presents one of the largest studies on the role of social media in the Arab Spring movement. Using over 2 million tweets generated by 110 thousand users in 11 Middle East countries during early 2011, we depict the landscape of aggregated Twitter usage in those countries as the revolution unfolded. Our results suggest that social media has been used to both lead and reflect real world protest activities. Compared to non-protest-related content on Twitter, we find that protest-related content travels directionally from central users to peripheral users, and the adoption of protest-related content can be modeled by a complex contagion process.

# CHAPTER 7

# CONCLUSION

In summary, I am deeply intrigued by the developing characteristics of information diffusion in online social media. Thanks to the Internet and social media technologies, I believe that we are heading towards a more democratic era where revolutions can be started by ordinary people and the power to change is in the hands of the masses. As part of this process, social media sites such as Facebook and Twitter have also evolved from friendship networks to a much broader platform for organizing social/political changes and communicating with various communities. I hope my work can help understand this movement and foster the effective flow of information in the society.

APPENDIX A

# CHAPTER 1 OF APPENDIX

Appendix chapter 1 text goes here