CITY UNIVERSITY OF HONG KONG

香港城市大學

Jumping over the Network Threshold: Information Diffusion on Information Sharing Websites

跨越網絡的門檻：信息分享類網站上的信息擴散

Submitted to
Department of Media and Communication
媒體傳播系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

By

Wang Chengjun
王成軍

September 2014
二零一四年九月

Abstract

As a ubiquitous phenomenon, information diffusion has long been studied in communication research. Recently, accompanied by the rise of Web 2.0, information sharing websites (ISWs) have emerged on the World Wide Web as a new platform for information diffusion. ISWs can be defined as social media through which users submit, share, and comment on news or other types of information, which underscores the transformation of information diffusion as a participatory process. Technically, ISWs are usually based on social networking services (SNSs), information aggregators, and search engines. The evolution of information and communication technologies (ICTs) and their interplay with the public have transformed the underlying mechanisms of information diffusion, which have important implications for diffusion research and call for systematic investigations.

This research starts from a puzzlement in online information diffusion —contrary to the proposition of global diffusion, researchers keep finding that large-scale diffusion is rare and fragile. From the perspective of threshold models, the present dissertation aims at systematically investigating information diffusion on ISWs. Drawing on classic diffusion theories (e.g., news diffusion theory) and communication models of information flow (e.g., the ABX model), I extended the ABXC model to the ABXCT model as a general framework to incorporate and explain the emerging ICTs for information diffusion on ISWs, networked individuals, and the interplay between them.

The rise of digital media makes the “digital fingerprints” of human communication publicly available. This, in turn, grants us an opportunity to unobtrusively investigate the underlying diffusion mechanisms. Using the digital traces of information diffusion on Sina Weibo, Digg, and YouTube, three studies are carried out in order to reformulate the threshold

hypothesis of interpersonal effects, to compare collective gatekeeping with interpersonal effects, and to test the temporality hypothesis for online information diffusion research, respectively.

First, the study of Sina Weibo reformulates the threshold hypothesis for interpersonal effects in light of both threshold models and the J-curve model of information diffusion. Using the data of information diffusion on Sina Weibo, the findings confirm the threshold hypothesis of interpersonal effects. In particular, for information of public interest, interpersonal effects have a positive influence on diffusion size; for information of personal taste, interpersonal effects have a negative impact on diffusion size. The depth of diffusion networks is limited and the temporal diffusion curves are characterized by strong bursts. Thus, information diffusion on microblogs is both structurally and temporally constrained. To make information go viral means to break the bottleneck of local clusters and to spread information to diverse communities.

Second, the study of Digg primarily focuses on collective gatekeeping. ISWs provide a testing ground for the debate on the primacy of collective gatekeeping over interpersonal effects. Emphasizing interpersonal effects on ISWs, recent studies of online information diffusion generally overlook another player — information aggregators of ISWs. Through the process of news aggregation based on collaborative filtering, ISWs carry out a "collective gatekeeping" process that differs from news disseminations by mass media and news transmissions among individuals. Using the news voting data collected from a social news website (Digg), I compare the relative effects of interpersonal sources and collective gatekeepers on information diffusion, and find that collective gatekeeping is the primary driver for online news diffusion. Considered together, the nature of ISWs itself has a strong impact on the outcome of information diffusion. In the case of Digg, the mechanism of public voting rather than local social influence dominates

the size of information diffusion. In addition, the daily diffusion curves of news diffusion on Digg are also characterized by strong bursts.

Third, to understand the origin of bursts in public attention, I shift to a more general framework by proposing the hypothesis of temporality in the study of the diffusion of YouTube videos. Public attention is the extent to which individuals collectively allocate their attentions to cultural products across space and time. In addition to fragmentation and polarization, recent studies have shown that the temporal distribution of public attention is also characterized by strong bursts. Measuring bursts using peak fraction, this study aims at systematically studying how bursts of public attention are related to the popularity of YouTube videos (both lifetime and size of diffusion), and how it grows out of system recommendation, social influence, use of search engine, and mobile communication. The results show that, first, bursts are negatively related to the popularity of online videos; second, bursts are negatively related to searches, but positively related to system recommendations and the use of mobile devices; third, compared with videos in other categories (such as entertainment videos and education videos), news videos experience stronger bursts, while copyright-protected videos exhibit weaker bursts. In all, the findings suggest the temporality underlying various influential factors results in the bursts of public attention. Viral videos tend to seek continuous, incremental improvements, not great ups and downs. Thus, “going viral” means long-term growth rather than a temporary eruption. If the fragmentation of public attention implies “winner takes all”, the temporality of public attention implies that the winner grows steadily over time.

To summarize, the present dissertation contributes to the research of information diffusion in four aspects: first, it adopts the ABXCT model as a general theoretical framework to incorporate the emergence of ISWs, and to map concrete studies about information diffusion in a

systematic way; second, it formulates and confirms the threshold hypothesis of interpersonal effects; third, it conceptualizes collective gatekeeping as an alternative driving force of information diffusion; fourth, it investigates the burst of public attention in information diffusion, and formulates the temporality hypothesis. Overall, by illustrating the driving forces, the hidden patterns, and the underlying principles of online information diffusion, the present dissertation underscores the communication power of networked individuals and new communication technologies in the age of social media, and sheds light on the puzzlement about the limited size of online information diffusion.

Acknowledgement

Diffusion is a ubiquitous phenomenon widely existing in many different systems, including social systems. However, it is a great challenge for researchers to capture the underlying mechanisms and the hidden principles. Without a great deal of help from a number of individuals along the way, it would have been impossible for me to finish this dissertation. I would like to thank my teachers, friends, and family in following paragraphs of this acknowledgement. The "diffusion" of ideas and encouragements from them to me always inspires my passion for academic inquiry.

First and foremost, I am grateful for the guidance and support provided by my supervisor Jonathan J.H. Zhu. I have had much intellectual stimulation and enlightenment from him every week. His imagination in conceptualization and theory-construction, and his passions for cutting-edge research always invigorate me to walk along the way of academic research. Without his tolerance, patience, kindness, and stringent requirements, it would be impossible for me to finish this dissertation.

Second, I owe my thanks to the other faculty members in the Department of Media and Communication, City University of Hong Kong, especially Chin-Chuan Lee, Tsan-Kuo Chang, Xi-Gen Lee, Wan-Ying Lin, Crystal Li Jiang. They showed me the importance of personal interest, academic writings, and logical reasoning. Conversations with them gradually made me realize the boundaries of my research. Additionally, I would like to thank Robert Ackland and Lexing Xie, from whom I benefitted a great deal when I was an exchange student at Australia National University. Rob is a good role model in many ways, and his enthusiasms for network science excite me greatly. Lexing's constructive suggestions are very helpful for my study of the diffusion of YouTube videos.

Third, I would like to express my debts to a number of fascinating people that I have met in this small city. I would like to thank the members of the Web Mining Lab: Tai-Quan Winson Peng, Lun Zhang, Heng Lu, Lingfei Wu, Hai Liang, Jie Qin, Zhen-Zhen Wang, and He-Xin Chen. My special thanks go to Lingfei Wu. He persuaded me to use R and Mathematica, and we learned to use Python together. I would also like to thank my other friends in the department, especially Echo Liu, Pian-Pian Wang, and Xin-Zhi Zhang. Additionally, I would like to thank everyone who directly or indirectly offered his or her help on this dissertation. I want to thank Xiong Wang, whose lectures on "the four levels of scientific research" enlighten me to reflect on social scientific theories. I would also like to thank Wu Lin for his workshop on Python which paved the way for me to use Python. I also want to thank Fangjian Guo. As a visiting student from the University of Electronic Science and Technology of China, Fangjian helped me greatly in cleaning the huge data of Sina Weibo with Linux utilities, and in calculating the network threshold with C++.

Finally, I have to thank my family members, especially my father, my mother, my sister, and my wife, and I would like to dedicate this dissertation to them. Walking through the four years of Ph.D. training, I kept finding that I am increasingly ignorant about the outside world as well as about myself. Without their continually encouragements and understandings, I could not finish this dissertation.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Information diffusion is the spread of information among the members of a social system through certain channels over time. Individuals learn about the outside world, and even influence other people's decisions through information diffusion. As a fundamental communication phenomenon, information diffusion has great impacts on social life over a wide range, such as news consumption (Funkhouser & McCombs, 1971; Greenberg, 1964a, 1964b; Larsen & Hill, 1954; Miller, 1945), buying behavior (Ivković & Weisbenner, 2007), finding a job (Granovetter, 1973), collective action and political participation (González-Bailón, Borge-Holthoefer, Rivero, & Moreno, 2011; Lippe, 1994).

In the following sections of the introduction, I will first briefly discuss the research background which provides us a broad picture of online information diffusion. Next, I will explain the research objectives of this dissertation. Finally, I will introduce the structure of this dissertation.

1.1 Research Background

1.1.1 The Rise of Information Sharing Websites (ISWs)

Recently, accompanied by the rise of social media, especially social networking services and information aggregation services, information sharing websites (ISWs) emerged on the World Wide Web as a new platform for information diffusion. The evolution of information and communications technologies (ICTs), and their interplay with the public, has changed the underlying mechanisms of information diffusion, which has important implications for communication studies.

Information sharing websites (ISWs) can be defined as the social media through which users can submit, share, and comment upon information. Technically, ISWs are based on social

networking services (SNS) and information aggregation tools, which underscores the transformation of online news diffusion from mass communication to a participatory process. Firstly, ISWs allow people in online social networks to spread their interpersonal influence through the ties, branches, and communities of social networks. Secondly, ISWs help users find required information through a variety of the other information channels (Chowdhury & Landoni, 2006) in addition to interpersonal networks. For example, the collective gatekeeping function of information aggregators helps users directly learn information from the information system of ISWs.

Information diffusion has been widely studied since the very beginning of communication research. There are rich intellectual resources for diffusion research, including diverse theoretical perspectives, extensive empirical investigations, and many analytical models. Specificially, there are a number of lines of diffusion research, such as two-step flow (Katz, 1957; Katz & Lazarsfeld, 1955), news diffusion (Greenberg, 1964a, 1964b; Larsen & Hill, 1954; Miller, 1945), diffusion of innovations (Rogers, 1983), threshold models (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988; Valente, 1995), the Bass diffusion model (Bass, 1969, 2004), and epidemic models (Becker, 1992).

To date, information diffusion has evolved into one of the central concerns of social media research. It attracts many scholars from different disciplines to work on uncovering the hidden patterns of online information diffusion, as well as the underlying diffusion mechanisms. One of the most important diffusion mechanisms is the interpersonal effect based on social networks. The rise of ISWs revives diffusion research on interpersonal effect. However, in addition to interpersonal effects, there are many other influences, such as exogenous impacts,

individual attributes, features of information, as well as the mutual relationship between different information. To identify these driving forces has theoretical significance for diffusion research.

1.1.2 A Puzzlement of Online Information Diffusion

This research starts from a puzzlement of online information diffusion. The study of network epidemic models suggests that on complex networks, especially scale-free networks, information tends to infect a sizable proportion of the population (Pastor-Satorras & Vespignani, 2001). Given the long-tailed degree distribution of networks, "the threshold of generic epidemic models is vanishing or finite", and thus there tends to be a global cascade of information diffusion on the Internet (Pastor-Satorras & Vespignani, 2001). Global cascades denote sufficiently large diffusions. The existence of global cascades suggests that more than a fixed fraction of the large but finite networks have been activated to spread the information.

However, based on analytical models and numerical simulations, Watts (2002) contends that "the global cascade" is a *rare* and *fragile* nature of complex systems. "A system may appear stable for long periods of time and withstand many external shocks, then suddenly and apparently inexplicably exhibit a large cascade" (Watts, 2002). In another paper, Watts et al. (2007) posit that large cascades of influence are driven not by the "influential" but by a critical mass of easily influenced individuals. Therefore, there is a tension between theoretical models and social realities. However, epidemic models of this kind cannot ultimately resolve the debates.

To empirically study why some hit songs are much more successful than average, Salganik, Doggs, and Watts (2006) designed an artificial "music market" to observe how 14,341 participants downloaded previously unknown songs. They found that "increasing the strength of social influence increased both inequality and unpredictability of success", and the quality of music can only partly influence its success (Salganik, et al., 2006). Therefore, it is quite difficult

or even impossible to *predict* which information will go viral, which again confirms that "the global cascade" is a *fragile* nature of complex systems (Watts, 2002).

In the real world, large-scale information diffusion seems to be *rare* or *elusive*. Against the proposition of global diffusions, researchers keep finding that the diffusion size (e.g. cascade size) is limited (Bakshy, Hofman, Mason, & Watts, 2011; Lerman & Ghosh, 2010; Leskovec, Singh, & Kleinberg, 2006; Sun, Rosenn, Marlow, & Lento, 2009). The distributions of diffusion size are usually found to be very left-skewed or even scale-free. In other words, most information cannot reach a sizable proportion of people in the online social network.

The questions of how information spreads and goes viral have attracted much attention from many different disciplines, such as communication (Fu, 2012), marketing (Aral, Muchnik, & Sundararajan, 2009; Aral & Walker, 2011; Berger & Milkman, 2012), computer science (Lou & Tang, 2013; Steeg, Ghosh, & Lerman, 2011; S. Wu, Hofman, Mason, & Watts, 2011), and network science (González-Bailón, Borge-Holthoefer, Baños, & Moreno, 2013; Ugander, Backstrom, Marlow, & Kleinberg, 2012). Yet, what determines the success of information diffusion is still elusive (Weng, Menczer, & Ahn, 2013). As we have discussed above, the popular approach considers information diffusion as *simple contagion*, which assumes that information spreads like a disease, each exposure has the *same* probability to infect an individual, and employs the epidemic models (Pastor-Satorras & Vespignani, 2001; Watts, 2002; Watts & Dodds, 2007; F. Wu, Huberman, Adamic, & Tyler, 2004). However, recent studies reveal that diseases and information spread in differently. Therefore, the perspective of *complex contagion* which assumes that repeated exposures increase the possibility of spreading is more appropriate for modeling online information diffusion (Centola, 2010; Romero, Meeder, & Kleinberg, 2011).

Put together, our knowledge about the limited size of online information diffusion is still insufficient. Although there is an increasing interest in diffusion research in the past decade, the study of online information diffusion seems to be clogged by undigested facts. As a result, we are still ignorant about what makes global cascade possible, as well as what constrains the information from "infecting" a sizable population. Therefore, what stops the information going viral becomes a significant and specific research question, in terms of both theoretical contributions and practical applications.

1.2 Research Objectives

To uncover the puzzlement about the limited diffusion size for online information diffusion, this present dissertation aims to look into the relationship between diffusion size and diffusion m0echanisms for information diffusion on some representative information sharing websites (ISWs).

1.2.1 Research Questions

First, confidence in network epidemic models (Pastor-Satorras & Vespignani, 2001) rests on their most profound assumption that person-to-person influence has the potential to infect a nontrivial proportion of the population. Therefore, as the starting point of this study, the central concern is to identify the influence of interpersonal effects on information diffusion. Here, interpersonal effects are defined as the interpersonal influence on individuals' information exposure and sharing. This dissertation aims to assess to what extent an individual is influenced by his or her online friends to share the information, especially the effect of repeated exposure. Repeated exposure has been loosely defined as social reinforcement both in the study of diffusion of innovations occurring in the offline setting (Valente, 1993) and in the study of the

spread of online healthy behavior (Centola, 2010) and the diffusion of tweets (Weng, et al., 2013).

To quantify interpersonal effects in information diffusion, the present research is primarily based on the perspective of threshold models (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988). The benefit of employing network threshold to quantify interpersonal effects lies in its power in capturing the social reinforcement for repeated exposures. Most diffusion models are constructed based on the assumption of either simple contagion or complex contagion. To investigate the relationship between network threshold and the size of information diffusion (i.e., virality) can test and verify the hypothesis of complex contagion—multiple exposures contribute to the transmission of online information.

In addition to investing the reinforcement feature of interpersonal effects, this dissertation also investigates the non-linear characteristics of interpersonal effects. In the literature of diffusion research, the J-curve model is the first to address the curvilinear nature of interpersonal effect (Greenberg, 1964b). In the study of news diffusion, Greenberg found a non-linear relationship between diffusion size and interpersonal effects, which further suggests that the salience of information plays an important role. Consistent with the J-curve model, Onnela and Reed-Tsochas's research on the diffusion of Facebook applications reveals that there are also two regimes of social influence (2010). Once Facebook applications cross a particular threshold of popularity, social influence has significant influence on the diffusion of Facebook applications. Below this threshold, social influence appears to vanish. Therefore, it is also salient to test whether the two regime phenomenon of interpersonal effects exists in the social setting of online information diffusion. Consolidating the thoughts elaborated above, I raise the first research question:

RQ 1: How do interpersonal effects influence information diffusion on ISWs?

Second, information technology has been changing our daily lives by transforming the underlying mechanisms of human communication. As I have argued in the definition of information sharing websites (ISWs), social networking services, and information aggregators of ISWs are becoming important in online information diffusion. Their influence is not merely a technical impact. Humans are also interacting with each other and the diffusion systems. The first research question concerns the interpersonal effects which are closely relevant to the social network services of ISWs. As an alternative influence, information aggregators are also playing an important role in information dissemination (Lerman & Ghosh, 2010). For example, on Twitter, the popular tweets have been aggregated as trending tweets appearing on the sidebar of the front page, and on Digg, the most popular news is aggregated onto its webpage of information aggregation. If the direction of influence from interpersonal effects is bottom-up, the impact of information aggregators is top-down. Therefore, a second step should be taken to clarify the impacts of information aggregators, and then to conceptualize them within the framework of information diffusion.

RQ 2: How do information aggregators influence information diffusion on ISWs?

Third, the bursts of public attention describe the situation of a sudden and violent growth and decline of public attention. Recent research shows that online information diffusion is also characterized by bursts of public attention (Crane & Sornette, 2008). Bursts of online information diffusion underscore the time dimension of information diffusion, and concern the temporal patterns of online information diffusion, as well as their influences on the size of online information diffusion. S-shaped curves have been extensively employed to describe cumulative diffusion curves (Rogers, 2003). In addition to classic S-shaped diffusion curves, Chaffee (1975)

proposes that there are also three deviations: 1. Incomplete diffusion, where information fail in reaching 100 percent of the population; 2. Rapid diffusion, where there is a much more rapid acceleration than normal S-shaped diffusion. Rapid diffusion is usually stimulated by senders at the initial stage; 3. Slow diffusion, in the early stages, grows more slowly than normal S-shaped diffusion. However, since traditional classifications of diffusion curves are primarily based on the cumulative growth curves, they tends to neglect fluctuations in daily diffusion patterns, especially the temporality of diffusion which is reflected by bursts.

According to the temporal patterns of online information diffusion, especially the phenomenon of bursts, the most accepted explanation about the limited size of online information diffusion is the temporality of information diffusion. The term "temporality" emphasizes two aspects of information diffusion: First, the temporality of different driving forces of information diffusion, which concerns the duration of the antecedents (i.e., the determinant factor) of attention allocation. Second, temporality emphasizes the persistence of information diffusion, which concerns whether the information can continue infecting the public over time. The temporality of driving forces governs the rise and fall of public attention towards specific media products, e.g., according to prior studies, the burst of diffusion slows down the diffusion process (Karsai et al., 2011). In practice, the temporality of information diffusion can be indicated by the burst of diffusion curves. Thus, I formulate the research question about the burst of information diffusion to gauge the temporality of information diffusion:

RQ 3: How do bursts influence information diffusion on ISWs, and what is the origin of bursts in information diffusion on ISWs?

To summarize, this research starts with a puzzlement of online information diffusion—the absence of global cascade in online information diffusion. To answer this general question,

this dissertation focuses on the size and temporality of information diffusion, and examines the underlying diffusion mechanisms. The present dissertation is designed to conduct three studies about information diffusion on three kinds of information sharing websites (ISWs): microblogs (e.g., Twitter and Sina Weibo), social news websites (e.g., Digg), and video sharing websites (e.g., YouTube). Correspondingly, three data sets of information diffusion will be analyzed in this dissertation.

1.2.2 Theoretical Significance

First, this study aims to establish a theoretical framework, and to map the landscape of the research on information diffusion. Thus, one of the research objectives is to describe the overall picture of online information diffusion. The present dissertation draws on classic communication models of information flow (e.g., the Lasswell formula, the ABX model, and the Westley and Maclean's conceptual model of mass communication). Classic communication models start from the Lasswell formula (Lasswell, 1948) which formulates the process of communication as "who says what to whom in what channel and caused what effects" ("5W"). Newcomb (1953) extends this into a dynamic process by discussing the relationships among information sources (denoted by either A or B), information receivers (denoted by either A or B), and the information itself (denoted by X), which is the ABX model. Westley and Maclean (1957) later extend this model by adding the information channels (denoted by C).

With the rise of Web 2.0 and the other new technical features of computer-mediated communication, there are more and more driving forces of online information diffusion. Regarding social changes brought about by the advent of network society and especially social media, this study expands upon the adapted social roles of information sources, media, and audiences. Firstly, audiences have been connected with each other, and become networked

individuals. Secondly, networked individuals and organizational social media users can both act as information receiver and information source, which blurs the distinction between information source and information receiver. Thirdly, the social networking service and information aggregating functions of ISWs creates another kinds of information channels. Further, departing from the static style of traditional models, this new model calls for attention to the temporal dimension of information diffusion (denoted by T). Thus, this new model is named as the ABXCT model.

Second, this study aims to analyze interpersonal effects from the perspective of threshold models. The rise of online social networks underscores the significance of interpersonal effects within communication networks. The most significant feature of online information diffusion is the interdependence of the communication networks. The most fundamental assumption about interpersonal effects asserts that one person's information sharing behavior is contingent upon that of others. Based on the assumption of maximizing utility of cost and benefit from binary decision-making, threshold models deal with the extent to which an individual's behavior is based on the extent to which his or her local network has been activated. Therefore, threshold models can be employed to capture interpersonal effects. Theoretically linking the threshold of interpersonal effects and the size of information diffusion, the present dissertation attempts to conceptualize the threshold hypothesis of interpersonal effects.

Third, one of the new features of online information diffusion is introduced by the information aggregating function of ISWs. Thus, one central concern is to theorize the process of information aggregation. Actually, a great part of information aggregation is based on the collaborative filtering of information. In the present dissertation, I conceptualize collaborative information filtering and the subsequent information aggregation process as collective

gatekeeping. Collective gatekeeping make aggregated popularity publicly available, which has the potential to shape online information diffusion. From the perspective of information cascade theory, this dissertation compares the influence of collective gatekeeping with interpersonal effects in news diffusion on a social news website.

Fourth, the present dissertation aims to capture one of the novel patterns of diffusion curves—bursts of public attention in various test grounds of information diffusion (e.g., microblogs, social news websites, and video sharing websites). Based on the positive relationship between bursts of public attention and diffusion size, this study highlights the importance of identifying the origin of bursts. To fulfill this purpose, I propose the temporality hypothesis, which aims to find the origin of bursts from the temporal features of the driving forces of information diffusion.

Overall, this dissertation is expected to contribute to our understanding of information diffusion in various aspects. The insights derived from this study help advance our knowledge about diffusion mechanisms and hidden principles. In the digital ages, online information diffusion is increasingly shaping our daily lives to a great extent. The theoretical generalizations of this dissertation have the potential to help explain or even predict the popularity of information diffusion. The implications can be used to disseminate information quickly, widely, and sustainably. Taken together, the findings could contribute to our understandings of the limited size of online information diffusion.

1.3 Structure of the Dissertation

The present dissertation is composed of seven chapters, including this chapter of the Introduction (see Figure 1). Chapter 2 reviews an extensive body of literature about information

diffusion. Based on the relevant literature, I develop a conceptual framework about information diffusion on ISWs in Chapter 3.

Insert Figure 1 here

Three kinds of ISWs (social news websites, video sharing websites, and micro-blogs) will be investigated in Chapters 4, 5, and 6. The reason for investigating online information diffusion with three kinds of ISWs is to avoid the risk of limited generalizability (e.g., mistaking the part for the whole), and to demonstrate the diversity of diffusion mechanisms. In Chapter 4, I primarily focus on verifying the threshold hypothesis of interpersonal effects in the context of information diffusion on microblogs (Sina Weibo). Chapter 5 mainly focuses on news diffusion on a social news website (Digg), within which, I focus on comparing the influence of two driving factors of news diffusion—interpersonal effects and collective gatekeeping. The strong bursts of information diffusion lead the present dissertation to investigate the temporality of information diffusion in Chapter 6. In Chapter 6, I formally propose the temporality hypothesis for information diffusion, and verify it by gauging the origin of bursts for the diffusion of YouTube videos.

Finally, in Chapter 7, I discuss the overall findings of three empirical studies on information diffusion and summarize the theoretical generalizations, implications, and limitations of this dissertation.

Chapter 2. Diffusion Research: Theories and Models

There are many strands of research on information diffusion, including the seminal study of two-step flow (Katz, 1957; Katz & Lazarsfeld, 1955), news diffusion (Greenberg, 1964a, 1964b; Larsen & Hill, 1954; Miller, 1945), diffusion of innovations (Rogers, 1983), threshold models (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988; Valente, 1995), Bass diffusion models (Bass, 1969, 2004), and epidemic models (Becker, 1992). This chapter contains two sections. I would firstly review relevant theoretical perspectives of information diffusion, based on which, secondly, I will discuss a research agenda for the research of online information diffusion in this present dissertation.

2.1 Theoretical Perspectives of Diffusion Research

For the diffusion research in communication field, there are three related theories, including two-step flow (Katz, 1957; Katz & Lazarsfeld, 1955), news diffusion (Greenberg, 1964a, 1964b; Larsen & Hill, 1954; Miller, 1945), diffusion of innovations (Rogers, 1983). Rogers (2003) summarized the fields which have taken up the diffusion research approach: education, anthropology, public health, marketing, geography, sociology (including early sociology, rural sociology, and general sociology), political science, communication. Rogers's effort at integrating information-flow research with the theory of diffusion of innovations was so successful that information-flow theory became known as information diffusion theory (Baran & Davis, 2011). In addition to the diffusion theories, there are also several well established diffusion models. For example, threshold models (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988; Valente, 1995), Bass diffusion models (Bass, 1969, 2004), and epidemic models (Becker, 1992) have been extensively employed in economics, sociology, medicine, and physics, and the findings shed light into the conceptual framework of information diffusion.

In the following sections of this chapter, I review six theories and models of diffusion research related to the following questions: First, what are the assumptions of the theories? Second, what are the driving forces of the diffusion? And third, what are the underlying principles which shape the diffusion process?

2.1.1 Two-step Flow

Against the backdrop of magic bullet theory of media effect, the theory of two-step flow is the first attempt to capture the pattern of information flow in the society of mass media. It asserts that information flows first from the mass media to opinion leaders, and then flows from opinion leaders to their followers (Katz & Lazarsfeld, 1955). Lazarsfeld et al. (1945) found little evidence that media played an important role in directly influencing the voters; On the contrary, these voters were much more likely to report that they had been influenced by other people.

Based on the preliminary findings, Lazarsfeld formulated the proposition of opinion leaders who act as "gatekeepers" and mediate media's influence on the public. To empirically test this proposition, Lazarsfeld et al. carried out surveys with snow-ball sampling strategy in the city of Decatur. Eight years later, Katz and Lazarsfeld (1955) published their book of *Personal Influence* based on the Decatur data, in which they formally proposed the two-step flow theory (see Figure 2).

_____

Insert Figure 2 here

_____

The model of two-step flow assumes that individuals connect with each other, rather than being isolated. However, individuals are not equal in passing and receiving information. The

active roles (e.g., opinion leaders) in personal network are attentive to media information, and they transmit the media message to the other people. Thus, against the stimulus-response model of communication, mass media do not operate in a social vacuum. Mass media have to compete with the other driving forces (e.g., the other information sources, power, and knowledge) to influence individuals. Actually, to study the relative strength of media effect and interpersonal influence becomes a key question in later diffusion research (e.g., new diffusion research).

2.1.2 News Diffusion

Another line of research on information diffusion focuses on one specific category of information—news. News diffusion is the process of news flowing from news sources to audiences. Beginning in the 1940s, increasing during the 1960s-1980s, and decreasing in the 1990s, classic studies of news diffusion compared media effect and interpersonal effects by studying the diffusion channels, cumulative diffusion curves, and the determinants of news diffusion (Larsen & Hill, 1954). As the first research of news diffusion, Miller's study in 1945 established the tradition of assessing the impacts of various diffusion channels (Miller, 1945). Focusing on the diffusion of the news of Roosevelt's death among143 students of Kent State University, Miller (1945) found that 85% of the students learned the news by word of mouth.

As a common body of knowledge, news stories filter into community life and inform people through mass media and interpersonal channels. As a social process, news diffusion is the cumulative growth of the number of news knowers on the aggregate level, and on the individual level, news diffusion concerns when and how individuals learn about news events. In the diffusion process, social actors interact with the social structure, which makes news diffusion shaped by both agency and structure.

First, in the aspect of agency, conceptually, news diffusion is also a process of the flow of public attention. it is stimulated by individuals' needs and motivations. Viewing news as one kind of knowledge, and news diffusion as a process of learning, individuals' exposure to the news leads to awareness generated as a result of news-learning. In this process, two fundamental factors which will definitely influence the size of news diffusion should be highlighted: one is the limited capacity of individuals' attention (Kahneman, 1973), and another one is the selective attention (A. Treisman & Fearnley, 1969; A. M. Treisman, 1964).

Second, social structure also shapes the diffusion process. J-curve model of news diffusion states that there is a curvilinear relationship between percentage of news awareness and the percentage of interpersonal sources (DeFleur, 1987; Greenberg, 1964b). Greenberg (1964b) studied 18 news events in terms of size of diffusion, media differences, and found that the relationship between the extent of news diffusion and the percentage of interpersonal sources could be described as a J-curve: people learn breaking news primarily from interpersonal sources but other less important news from the mass media (see Figure 3).

Insert Figure 3 here

The other forms of interpersonal effects can also influence the size of news diffusion. For example, news diffusion can be stimulated by interpersonal discussions. Im et.al (2011) argues that "previous meaning of news diffusion was virtually equivalent to the increasing awareness level among its audiences, the implicit assumption was that the news remains unchanged" (Im, et al., 2011, p. 612). By contrast, online comments can usually modify the original news.

Interpersonal discussions or conversations deepen the depth of news diffusion (Budd, MacLean Jr, & Barnes, 1966). For instance, those informed by the interpersonal communication are more likely to talk to a greater number of other persons (Greenberg, 1964a; Larsen & Hill, 1954).

Third, prior research on news diffusion also concerns the physical location of news learning (e.g. home, work) (Greenberg, 1964a) as well as the time dimension of news diffusion in terms of social time (e.g. daily life routine, time of the day) (Budd, et al., 1966). Therefore, how public discussions and news sharing behavior promote the diffusion process in the aspects of both time and space deserves more attentions.

Fourth, since most of the studies on news diffusion are case studies, the attributes of the news are not fully explored, as suggested by many studies since 1954, the diffusion of news of various types to different segments of society should be studied to undercover the value structure within the society (Larsen & Hill, 1954).

Insert Figure 4 here

News diffusion research constitutes one of the most long-standing traditions in communication research (more than 60 years, see Figure 4), and the theoretical developments have been generally summarized as Information-flow theory (Baran & Davis, 2011). Yet, limited by methods and theoretical framework, research on news diffusions has been criticized for producing very few theoretical generalizations, and most of the accumulated contributions are debatable (DeFleur, 1987). For example, although news diffusion studies concern the impact of diffusion channels, the problem of whether interpersonal communication is an important channel for news diffusion has remained largely unresolved (DeFleur, 1987).

To expand the scope of research and make this line of research flourish again, Rogers (2000) suggests future studies to move from the traditional paradigm to more theoretically-driven areas, and considers the news diffusion function of the Internet a promising direction. Currently, accompanied by the thriving of news diffusion in the virtual community, a fast-growing body of literature is starting to concern online news diffusion.

2.1.3 Diffusion of Innovations

The theory of diffusion of innovations focuses on the process through which innovations (e.g., ideas, practices, products, technologies, and services) spread in social systems over time. Rogers's definition of diffusion is naturally relevant to innovations. For example, Rogers (2003) defined diffusion as a process in which (1) an innovation (2) is communicated through certain channels (3) over time (4) among the members of social system. There are four main concepts: the innovation, communication channels, time, and a social system. To be specific:

First, the characteristics of an innovation have impacts on its rate of diffusion. Five attributes have been summarized by Rogers in his influential book *Diffusion of Innovations*: (1) relative advantage, (2) compatibility, (3) complexity, (4) trialability, and (5) observability (Rogers, 2003). Given the importance of innovations' attributes, to modify or change the features of innovations has significant influence on users' adoption of innovations.

Second, communication channels are the means by which information flow from one person to another. Similar to the tradition of news diffusion research, the studies on diffusion of innovations also pay great attention to the communication channels, especially the relative importance of media channel to interpersonal channel. Firstly, media effect plays an important role in spreading innovations. Secondly, peers' attitudes and behaviors will influence individuals' decision of adoption, in terms of social influence. Thirdly, confounded with social influence, the

theory of homophily (and heterophily) argues that people with similar (or different) attributes tends to interact with each other (Rogers & Bhowmik, 1970). Homophily concerns the degree to which individuals who interact are similar in certain attributes. Thus, it is necessary to differentiate media effect, social influence, and homophily in the process of diffusion (Aral, et al., 2009).

Third, time is the third dimension in diffusion of innovations. The innovation-diffusion process can be conceptualized into five steps: knowledge, persuasion, decision, implementation, and confirmation (Rogers, 2003). It is necessary to note that, to reduce the uncertainty of adopting an innovation, individuals seek information at various stages of the innovation-diffusion process. Based on innovativeness, i.e., the degree to which adopters is relatively earlier in adopting new innovations than other members of a social system, adopters are classified into five categories: innovators, early adopters, early majority, late majority, and laggards (Rogers, 2003).

Fourth, diffusion occurs within a social system. The social structure, social norms, and opinion leaders of the social system constrain the diffusion process, making the human behavior regular, stable, and predicable. In addition, the decisions of adoption can be made by individuals, groups, or the collective (Rogers, 2003). Accordingly, innovation-decisions can be classified into three categories: optional innovation-decisions, collective innovation-decisions, and authority innovation-decisions (Rogers, 2003). Optional innovation decisions are choices to adopt or reject an innovation that are *made* by an individual independent of the decision of the other members of the system; collective innovation-decisions are choices of adoption made by consensus among the members of a system; and authority innovation-decisions are choices of adoption made by a group of elites in a social system.

Diffusion of innovations follows a ubiquitous pattern—an S-shaped curve. There are two ways to demonstrate the diffusion process (Rogers, 2003). On one hand, to draw the fraction of *new* adoption of innovation over time, the growth curve follows a normal, bell-shaped curve. On the other hand, to plot the cumulative growth of adopters over time, we can find a S-shaped curve of adoption (Ryan & Gross, 1943). As aforementioned, according to the continuum of innovativeness revealed by the sequential temporal diffusion curves, adopters can be partitioned into five adopter categories: innovators, early adopters, early majority, late majority, and laggards. Although S-shaped curve gained a dominant position in early 1960s, the limitations should be noticed. The S-shaped curve is innovation-specific and system-specific, and it is only applicable to the cases of successful innovation, in which an innovation diffuses to most potential adopters.

2.1.4 Bass Diffusion Model

Inspired by the original idea of diffusion of innovations, a tremendous expansion has emerged in the marketing literature since Frank M. Bass's work on Bass diffusion model in 1969. Bass diffusion model assumes that diffusion is motivated by two mechanisms: innovativeness and imitation (Bass, 1969). Innovative early adopters adopted the products first, and the followers will imitate early adopters and adopt the products later. Accordingly, there are two parameters in the Bass model: the coefficient (parameter) of innovation and the coefficient (parameter) of imitation. The probability of adopting by those who have not yet adopted is a linear function of those who had previously adopted. Note that, although Bass's 1969 paper used the term "imitation" to describe the interactive portion of adoption, the parameter of imitation could be "word of mouth", learning, contagion, interpersonal communication, and internal influence (Bass, 2004).

In the Bass model, the parameter representing the potential market is denoted by *M*. As the ultimate number of purchasers of the product, M is a constant. Time intervals are numbered sequentially with the first full time interval (usually a year) of sales, and the time interval is denoted by *t*. The portion (fraction) of the potential market that adopts at time t is *f*(*t*). Accordingly, the portion (fraction) of the potential market that has adopted up to and including time *t* is *F(t)*. The number of the potential market that adopts at time *t* is *a(t)*.The number of the potential market that has adopted up to and including time *t* is *A(t)*. The Bass model coefficient (parameter) of innovation is *p*. The Bass model coefficient (parameter) of imitation is *q*.

The model construction starts from the idea of hazard rate which is defined as the portion that adopts at *t* given that they have not yet adopted. Let *h(t)* denote the hazard rate at time *t*. The probability of adoption at time *t* is *f(t),* and the proportion that having not adopted the products by time *t* is *1- F(t)*. Thus, by definition of conditional probability, *h(t)* can be expressed by:

$$\mathbf{h}(\boldsymbol{t}) = \frac{\boldsymbol{f}(\boldsymbol{t})}{\mathbf{1} - \boldsymbol{F}(\mathbf{t})}$$

Bass (1969) posits that "the probability of adopting by those who have not yet adopted is a linear function of those who had previously adopted". Given the Bass model coefficient of innovation (P) and the coefficient of imitation (q), the probability of adopting a product at time t can be expressed as:

$$\mathbf{f}(\boldsymbol{t}) = \boldsymbol{p} + \boldsymbol{q}\boldsymbol{F}(\boldsymbol{t})$$

Combining the two equations together and solving the differential equation of bass diffusion model leads to the analytical formula for F(t) as a function of t, p, and q:

$$\mathbf{F}(\mathbf{t}) = \frac{\mathbf{1} - \mathbf{e}^{-(\mathbf{p}+\mathbf{q})\mathbf{t}}}{\mathbf{1} + \frac{\mathbf{q}}{\mathbf{p}}\mathbf{e}^{-(\mathbf{p}+\mathbf{q})\mathbf{t}}}$$

The most important contribution of Bass diffusion model lies in supplying a mathematical formula for diffusion research. Based on these assumptions, Frank M. Bass analytically constructs the mathematical relationship between the portion (fraction) of the potential market that has adopted up to and including time t and innovativeness and imitation. By solving the differential equation, Frank M. Bass successfully got the function of cumulative growth. The second important contribution of the Bass diffusion model is to predict the total number of adopters in each time period. Based on the mathematical formula, it is very convenient to forecast the adopters in the future by fitting the bass diffusion model. Note that such prediction deals with the market in an aggregate way, rather than the adoption or rejection on individual basis.

As one kind of S-shaped curve, Bass diffusion model is a special case of the Gamma/shifted Gompertz distribution (G/SG). There are two special cases of Bass diffusion model. When the parameter of imitation equals zero, the model reduces to exponential distribution; and when the parameter of innovation equals zero, the model reduces to logistic regression. If the coefficient of imitation is greater than the coefficient of innovation, the growth rate of cumulative growth curve rises to a peak and then declines, which is a classic form of S-shaped form (Bass, 2004).

2.1.5 Epidemic Models

The epidemic model is another intellectual source for information diffusion research. The first known mathematical model of epidemiology is formulated by Daniel Bernoulli (1760) when he studied the mortality rates in order to eradicate the smallpox. However, it was not until the early twentieth century that deterministic modeling of epidemiology started. Ross (1911) developed differential equation models of epidemics in 1911. Later, Kermack and McKendrick

(1927) found the epidemic threshold and they argued that the density of susceptible must exceed a critical value to make the outbreak of an epidemic happen.

The mathematical models developed by epidemic research help clarify assumptions, variables, and parameters for diffusion research, lead to useful concepts (e.g., threshold, reproduction number), supply an experimental tool for testing theoretical conjectures, and forecast epidemic spreading in the future (Hethcote, 2009). Although epidemic models are simplifications of reality, they help us refine our understandings about the logic of diffusion beneath social realities (disease transmission, information diffusion through networks, and adoption of new technologies or behaviors). To understand the epidemic models in a better way, I will briefly review the basic epidemic models: SI, SIR, SIS, and the applications in networks.

The SI model is the simplest possible model of infection. In the SI model, there are only two phases in the SI epidemic spreading process: Susceptible and Infectious. Let S be the proportion of the population that are susceptible. Let I be the proportion of the population that are infectious. At the initial time, the proportion of people who are infected is x_0, the proportion of people who are susceptible is S_0. β is the transmission rate, and it incorporates the encounter rate between susceptible and infectious individuals together with the probability of transmission. Consider a "closed population" with no births, deaths, or migrations, and assume the mixing is homogeneous (e.g., the susceptible individuals are uniformly spread in a geographic area, and the probability of contracting the infection is uniformly the same for all actors (T. G. Lewis, 2011)), yielding βSI as the transmission term. Thus, the equation for SI model is:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI$$

Insert Figure 5 here

Given every individual in the system must be either susceptible or infected, I + S = 1. Thus, the equations above can be transformed to:

$$\frac{\mathrm{dI}}{\mathrm{dt}} = \beta \mathrm{I}(1 - \mathrm{I})$$

To solve this differential equation, we can get the cumulative growth curve as a function of time:

$$I[t] = \frac{\mathrm{x}_0 e^{\beta t}}{1 - \mathrm{x}_0 + \mathrm{x}_0 e^{\beta t}}$$

Interestingly, this is a logistic growth featured by its S-shaped curve. The curve grows exponentially shortly after the system is infected, and then saturates as the number of susceptible shrinks which makes it harder to find the next victims. Thus, it could be used to model the classic diffusion of innovations.

In the naive model of SI, once one is infected, it is always infectious. However, this is not realistic for many situations of disease spreading. For many diseases, people recover after a certain time because their immune systems act to fight with the diseases. There is usually a status of recovery denoted by R. Let γ denote the removal or recovery rate[1].

[1] Usually, researchers are more interested in its reciprocal ($1/\gamma$) which determines the average infectious period.

There are two stages of the dynamics of the SIR model. In the first stage, susceptible individuals become infected by the infectious ones with who they contact. Similar to the SI model, β is the transmission rate between individuals; In the second stage, infected individuals recover at the average rate γ. Given the premise that underlying epidemiological rates are constant, the differential equations of simple SIR model (with no births, deaths, or migrations) are:

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

However, the differential equations above could not be analytically solved. In practice, researchers can evaluate SIR model numerically, as it is showed in the figure below.

Insert Figure 6 here

Another extension of the SI model is the one that allows for reinfection. If infected individuals are not immune to the diseases after their recovery, they can be infected more than once. The most simple model that captures this features is the SIS model. There are only two states: susceptible and infected, and infected individuals become susceptible after recovery. The differential equations for the simple SIS epidemic model are:

$$\frac{dS}{dt} = \gamma I - \beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

Insert Figure 7 here

Given S + I = 1, the differential equations have the solution:

$$I[t] = (1 - \frac{\gamma}{\beta}) \frac{Ce^{(\beta-\gamma)t}}{1 + Ce^{(\beta-\gamma)t}}$$

C is the integration constant in the form of $C = \frac{\beta x_0}{\beta - \gamma - \beta x_0}$.

One important contribution of epidemic models is the threshold phenomenon of epidemic diffusions existing in SIR model. The threshold of SIR model asks what factors determine whether an epidemic occur or fail. As the first step of analyzing the threshold of SIR model, the differential equation of SIR $\frac{dI}{dt} = \beta SI - \gamma I$ can be rewritten in the form:

$$\frac{dI}{dt} = (\beta S - \gamma) I$$

If dI/dt is smaller than 0, the contagion will soon wither and die out. Thus, as a boundary condition, $(\beta S - \gamma)$ should be larger than 0, and S should be larger than γ/β. This is the threshold phenomenon (Kermack & McKendarick, 1927). Based on the rationales above, if the initial fraction of susceptible (S(0)) is less than γ/β, the infection would not be able to start the invasion in the population. Here γ/β is defined as basic reproductive ratio R_0.

To summarize, "for an infectious disease with an average infectious period give by $1/\gamma$ and a transmission rate β, its basic reproductive ratio R_0 is determined by γ/β. In a closed population, an infection with a specified R_0 can invade only if there is a threshold fraction of susceptible greater than $1/\gamma$ " (Keeling & Rohani, 2011, p. 21).

In the section above, I mainly focus on the deterministic models of epidemics. However, despite the many advantages of deterministic models, it can be difficult to include realistic population networks, to incorporate realistic probability distributions for the time spent in the infectious period, and to assess the probability of an outbreak. Thus, the stochastic epidemic simulations, such as stochastic differential equations, Markov Chain Monte Carlo (MCMC), and agent based modeling, have been used to remedy the defect.

Network epidemic models have also been developed to investigate the widespread and rapid propagations (e.g., the contagion of computer virus) through a network. Typically, the network epidemic is brought about by adjacent nodes through propagations along one or more links. Network epidemic models consider the topology of the network as well as infection rate, death rate, and state transitions. This line of research is interested in the following questions: under what conditions will an initial outbreak spread to a nontrivial portion of the population? What percentage of the population will eventually become infected? What is the effect of immunization policies? For example, Pastor-Satorras et al. (2001) study the spreading of epidemics in complex networks using the mean-field method for network SIS model. Their findings indicate that in exponential networks (e.g., random graph network, small-world network), there is the usual epidemic threshold below which there is no prevalence of epidemic. Yet, on a wide range of scale-free networks, there is an absence of an epidemic threshold, which implies that scale-free networks are prone to the spreading of epidemics, as well as other

spreading phenomena, e.g., information diffusion. Based on this rationale of the absence of epidemic threshold, network scientists expect online information diffuse to a great proportion of the population.

2.1.6 Threshold Models

As a shift from traditional diffusion research to diffusion networks studies, threshold models supplies important intellectual resources for diffusion research. Threshold has been used to study interpersonal effects of collective behavior, e.g., residential segregation (Granovetter & Soong, 1988; Schelling, 1971), spiral of silence (Glynn & Park, 1997; Granovetter & Soong, 1988; Krassa, 1988), consumer demand (Granovetter & Soong, 1986), and diffusion of innovations (Valente, 1996).

One illustration of threshold could be easily given with a simple case of collective behavior. Suppose that one social system consists 10 people, and their individual thresholds are 0, 1, 1, 4, 4, 4, 5, 7, 8, and 9, respectively. The first one whose threshold is 0 initiates in engaging in the collective behavior without being influenced by others. Following the first actor, the second and third people whose thresholds are 1 spontaneously engage in the activity following the first people. Thus, so far there are three people engaging in the collective behavior. Given the smallest individual threshold of the other people is 4, nobody will engage in this activity. Thus, the diffusion stops with "infects" 3 people in the social system.

The significance of threshold lies in its determination of the diffusion curves. The threshold for different individuals in a social system varies. The innovators usually have very low thresholds, which makes them engage in the diffusion as soon as it is initiated. Later diffusers who are strongly against the innovation usually have very high thresholds. Their adoptions are only possible after many other individuals who have adopted the innovation in

their personal network. If the distribution is left-skewed, it means that the average threshold of the system is relatively small, thus the diffusion curve could take off relatively earlier and the diffusion can reach more people; if the distribution is right-skewed, it means that the average threshold is relatively big, and the diffusion stops earlier with a smaller diffusion size. If the distribution of the threshold is normally distributed, it is expected that the cumulative diffusion curve should be S-shaped.

Threshold models presume that individual behavior is based on the fraction of people in a social system who have already engaged in the behavior (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988). Thus, one person's threshold is the proportion of a community to engage in a behavior before the individual makes up his/her mind to do so. Individuals of lower thresholds tend to engage relatively earlier than those who have higher thresholds. The numeric value of threshold should be in the range of zero and one.

Threshold can be used to capture the impact of interpersonal effect. Based on the assumption of maximizing utility of cost and benefit from binary decision-making, threshold concerns the extent to which an individual's behavior is contingent upon that of others. The first diffusion investigation to gauge the nature of diffusion networks is the study of new drug's diffusion among medical doctors by Coleman et al. (1966). Valente (Valente, 1993, 1995) employed threshold models to reanalyzed the diffusion data of drug collected by Coleman et al. (1966). Each doctor has an individual threshold as the resistance to the medical innovation, which was gradually overcome by interpersonal effect. According to the utility-cost assumption of threshold models of human behavior, when the driving force equals the resistance of cost, individual starts to engage in the behavior. Based on the relationship between threshold and adoption of innovations, scholars of diffusion of innovations generalize that an individual is

more likely to adopt an innovation if more of the other individuals in his or her personal network have adopted previously (Rogers & Kincaid, 1981; Valente, 1995).

There is one extreme situation of threshold – the "zero threshold" (Valente, 1995, p. 64) , whose value is zero. Zero threshold implies that the people of zero threshold will engage in the collective behavior without being influenced by the other people in a social system. Thus, zero threshold itself implies the important of alternative influence in addition to interpersonal effect. Therefore, it could be used to capture the non-interpersonal effects in the diffusion process, for example, how many people have engaged in the diffusion without being influenced by their interpersonal networks.

Based on the mechanisms of diffusion, two basic types of diffusion are recognized: the chain-reaction snowballing of diffusion, and the diffusions among disconnected nodes. The chain-reaction contagion occurs because the members of a social system are closely linked by interpersonal networks. Prior research (Coleman, et al., 1966) shows that "the network interconnectedness of an individual in a social system is positively related to the individual's innovativeness" (Rogers, 2003, p. 330). In addition to interpersonal effect, external impacts also play an important role in diffusion process. For example, Burt (1987) argues that structural equivalence (e.g., two individual occupy the same position in the structure of the interpersonal network) shapes the adoption of new innovations, rather than social influence. Valente (1995) found that a combination of external influences (e.g., media influence) and the interpersonal effects can best explain the adoption of new drugs.

It's necessary to note the difference and connection between threshold and critical mass. Threshold concerns the tipping point for individuals, and critical mass is about the tipping point for the social system, groups and sub-groups. Critical mass occurs at the point at which enough

individuals in a system have adopted an innovation so that the innovation's further rate of adoption becomes self-sustaining. Another assumption of threshold models is the interdependence among individuals in social system. Interdependence is a situation in which an individual's behavior affects those of the others. According to the point of cost and benefit, there is a risk and uncertainty for people to engage in certain behaviors. Thus, if a group of people make decisions collectively, the interdependence of them help them reducing the uncertainty by making decisions based on the choices of others who they refer to, which implies that interdependence change individual threshold (Valente, 1995). Sequential interdependence happens when earlier diffusers' behavior influence later diffusers, and reciprocal interdependence exists when later diffusers provide benefits to earlier diffusers, especially in the context of interactive media. For example, reciprocal interdependence tends to spread telecommunication services by prevent preexisting adopters from stopping using them (Markus, 1987).

There are two ways to measure threshold (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988). First, to observe how an individual's behavior shifts with the other's decisions in terms of revealed preference (Granovetter & Soong, 1988), e.g., following this approach, Valente (1996) studied the diffusion of innovations. The second method to measure the threshold is by asking the respondents directly, e.g., Noelle-Neumann uses this method to measure the willingness to express (Noelle-Neumann, 1974, 1993). This second method is suspect, since its validity can't be tested. In this study, I adopt the first method and incorporate the ego-network information to unobtrusively gauge the collective behavior in news diffusion.

2.1.7 Summary of Diffusion Theories and Models

In the sections above, I briefly reviewed the most relevant diffusion theories and models developed in the past 60 years, this dissertation attempts to summarize and clarify some of the conceptual issues of information diffusion which serves as departure points for new investigations on information diffusion. In this section, I will briefly compare these major diffusion theories and models (see Table 1 and Table 2).

Insert Table 1 here

First, for diffusion theories, all of the three major diffusion theories (i.e., two-step flow, news diffusion, diffusion of innovations) I review here come from communication research. Two-step flow is an important phase of the research on media effect, and it opens the door of limited media effect and bringing interpersonal effects into the big picture of information diffusion research. Hereafter, distinguishing and comparing interpersonal effects and media effect becomes one central job of news diffusion research. Comparatively, the research on diffusion of innovations is more systematic, since it attempts to uncover how personal attributes, system features, innovation characteristic shapes the diffusion process. In terms of the main features of these theories, two-step flow is featured by the gatekeeping effect of opinion leaders; news diffusion is famous for the J-curve model which concerns the attributes of news (e.g., importance and relevance of news); and diffusion of innovations is well-known for the s-shaped diffusion curve. Accordingly, the most direct implications of two-step flow concerns the importance of diffusion networks; news diffusion emphasizes the salience of information; and diffusion of innovations underlines the categories of different adopters. Two-step flow is criticized for its simplistic, linear, and source-dominated communication model; news diffusion

research fails to consider utility or value of information for receivers; the most important limitation of diffusion of innovations is that it is a simplistic, linear, source-dominated theory, and it underestimates power of media (Baran & Davis, 2011, p. 163).

Insert Table 2 here

Second, in addition to three major theories, three diffusion models (Bass diffusion model, epidemic model, and threshold model) have also been reviewed. They have been developed in the disciplines of marketing, epidemiology, and sociology, respectively. In terms of driving forces, in Bass diffusion models, innovativeness and imitation are modeled with differential equation models; epidemic models mainly focus on the physical contacts (and demographic influences, e.g., birth, death); and threshold models also focus on interpersonal effects (however, zero threshold captures the extent of exogenous impact). All the three models (Bass diffusion models, epidemic models, and threshold models) are based on differential equation models which again reveal the power of mathematic in modeling human dynamics. Moreover, threshold models shift the perspective to utility and cost in decision-making, opening a new door to quantify interpersonal effects in empirical studies. However, it is also necessary to bear in mind the limitations of the approaches of mathematical modeling, e.g., Bass diffusion model presumes an accurate knowledge of the market size, and it is driven by the initial set of assumptions; Epidemic models mainly focus on social and physical contact, the other exogenous impact (e.g., media effect) are ignored. In the practical applications of threshold models, it is difficult to measure it empirically, and more importantly, the measurement of threshold usually confounds interpersonal effects with the other alternative influences (e.g., homophily).

2.2 Information Diffusion: A Research Agenda

The emergence of information sharing websites (ISWs) underlines the new characteristics of online information diffusion, which helps formulate new research questions and add new ingredients to the research agenda. In this section, I will illustrate the most important points of online information diffusion. Departing from classic diffusion theories and diffusion models, the research present in this dissertation is also designed following this research agenda (see chapter 3).

2.2.1 From Offline to Online

Online information diffusions are gaining increasing popularity and theoretical significance with the rise of network society (Castellano, Fortunato, & Loreto, 2009; Castells, 2007; Lerman & Ghosh, 2010; Shumate, 2010). The emergence of ISWs as one kind social media underlines the transformation of online information diffusion to a participatory process in which users can actively submit, comment, and disseminate information.

Online information diffusion advances traditional information diffusion in many crucial aspects. In the past, the scholars have to prepare in advance to study the sudden news events, which limits the design of study (Miller, 1945). Further, the self-report data got from survey or interview suffers from the memory problem brought by the time lag. Last but not least, the classic research on news diffusion are case studies, both the cases and the sample size are limited. While on the Internet, information diffusions happen in every moment, and scholars could trace the diffusion networks and the underlying social networks to get the unobtrusive data which could include thousands of pieces of information.

A fast growing body of literature about online information diffusion has been produced. However, by viewing information diffusion as a process of contagion (e.g. information cascade,

and social contagion) (Bakshy, Karrer, & Adamic, 2009), and epidemics (e.g. SIR model) (Becker, 1992), Most studies of online information diffusion merely extensively stress the interpersonal effects played social networks (Lerman & Ghosh, 2010; Shao, Havlin, & Stanley, 2009). Wu et al. (2011) examined the two-step flow model which asserts that information flows from mass media to opinion leaders first, and then flow to mass population (Kadushin, 2006; Lazarsfeld, Berelson, & Gaudet, 1944), for news diffusion on twitter to study who says what to whom, and find considerable support .

Online information is accompanied by new technical features. For example, social networking services, search engines, system recommendations are playing an important role. Although traditional study of information diffusion is well established, it may not be sufficient to explain the mechanism of online information diffusion, especially the role social media played in spreading information, and the networked individuals who actively share information to the ISWs.

2.2.2 From Information Exposure to Information Sharing

Information sharing is one necessary part of news diffusion to spread the information to a wider audience. ISWs users collaboratively filter information by making decision of sharing the information or not. Without information sharing, the information would be locked into the ego network of the information submitters. In the process of information diffusion, there are at least two kinds of participants: information knowers (those who have viewed the information), and information disseminators (the people who share the information, e.g., those who retweet information on Twitter, or those who share information on Digg.com). Such classification is consistent with prior research. For example, in the research of news diffusion, Larsen & Hill (1954) distinguished the news knowers and news diffusers. Accordingly, there are two ways to

measure the extent of information diffusion: the number of information knowers and the number of information disseminators.

Information sharing serves a new kind of online civil journalism or citizen journalism, and contributes to the democracy of society. “Online citizen journalism” refers to a range of web-based journalistic practices, including blogging, photo and video sharing, and posting eyewitness commentary. However, the definition of citizen journalism is still open for debate (Lasica, 2003). Citizens participate in the news process without necessarily acting as “content creators”. Reposting, linking, “tagging” (labeling with keywords), rating, modifying or commenting upon news materials posted by other users or by professional news outlets, which facilitate the news diffusion online is also defined as one kind of citizen journalism (Goode, 2009). This is also highly agree with the definition given by Shayne Bowman and Chris Willis (2003) about participatory journalism which describes the content and the intent of online communication that often occurs in news sharing website.

2.2.3 From Individuals to Network

The limitation of linear models became apparent in their applications to communication research. Early communication models, such as the Lasswell formula (Lasswell, 1948), the model of Shannon and Weaver (Shannon & Weaver, 1949), tend to be linear and neglect social interactions (e.g., feedbacks from audiences to mass media, the interconnections among the audiences) (Rogers & Kincaid, 1981). On the contrary, the convergence model for communication research (Kincaid, 1979; Rogers & Kincaid, 1981) argues that human communication is a dynamic, cyclical process over time featured by mutual causation and the interdependency among of the participants. One way to achieve this purpose is the approach of

communication network analysis. Communication network analysis is a method of research for identifying the communication structure in a social system.

Information diffusion happens within the communication networks. A communication network consists of interconnected individuals linked by patterned flows of information. Network data of information flows are analyzed with interpersonal interactions as the unit of analysis (Rogers, 2003). By analyzing the activity and popularity of news disseminators of news stories on news sharing website, future research should capture the structural characteristics of news transmission networks, both the topological patterns and nodes attributes, to judge the diversity of news and the equality of citizen participation.

2.2.4 From Information to Attention

First, information diffusion through social networks is simultaneously a process of attention allocation. In this perspective, information diffusion could be described as a process in which public attention selectively flows across a stream of information. The driving forces of information diffusion are also the antecedents of attention allocation. Therefore, the study of information can benefit a lot from the literature of public attention. The flow of public attention can be represent by the network of clickstream which shows how public attention flow through the nodes of information. One example is the clickstream network of 3553 Digg news stories in June 2009 (see Figure 8). The nodes denote news stories, and the links denote the flow of public attention between the nodes of news stories. The linear skeleton of the clickstream network indicates that public attention flows through 3553 news stories in the direction of time.

Insert Figure 8 here

Second, in terms of attention economy, information diffusion scholars should shift their perspective from information to attention, and pay more attention to the flow of public attention. Nowadays, we are living in an era with a heavy surplus of information, compared with our limited capacity of attention. Attention research lends itself to our understanding of the diffusion of information, especially in the era of attention economy. In the perspective of economics, the center of study originates in the assumption of scare resources. Although we have been familiar with the phrase of "information economy", information is not in short supply nowadays. On the contrary, human attention is the commodity in short supply (Goldhaber, 1997). Scholars even step further. For example, Davenport and Beck (2001) claims that in postindustrial societies, attention has become a more valuable currency than money in our bank accounts.

Based on this rationale, it is becoming obvious that it is the public attention rather than the exponentially growing information dominating the market of digital media (Webster, 2011). If this is the case, to investigate how public attention shapes the size of information diffusion is of great theoretical significance. For example, in light of the increasing supply of information and the limited size of public attention, there is a more intense competition among different information. Derived from the arguments above, this proposition would definitely forge the patterns of information diffusion nowadays.

Chapter 3. The Conceptual Framework of Online Information Diffusion

Information diffusion is a dynamic process driven by a number of driving forces, hidden mechanisms, and underlying principles. Based on the theoretical review in Chapter 2, in this chapter I develop a general conceptual framework in the form of graphical communication models to map the research of online information diffusion, and to organize different theoretical ingredients, such as perspectives, theories, models, patterns, driving forces, underlying principles of information diffusion, as well as the emerging new technical features of information sharing websites (ISWs). First, I will summarize relevant communication models about information flow. Second, I discuss the unique features of information sharing websites (ISWs). Finally, based on classic communication models of information flow and the technical features of ISWs, I develop a communication model for information diffusion in the era of digital media.

3.1 Theoretical Framework

3.1.1 Communication Models of Information Flow

Communication models in graphic form are defined as consciously simplified representations of a piece of reality (MaQuail & Windahl, 1993). They have advantages in (1) organizing diverse conceptual elements in a logical way to describe the big picture of social processes, (2) helping explain the social process in a heuristic manner, and even (3) making it possible to trace the course of events and to predict the outcomes.

Information diffusion is a process of information flow which has been well documented by various graphical communication models. For example, the Lasswell formula (Lasswell, 1948) demonstrates that communication concerns five elements: Who? What is said? In which channel? To whom? With what effect? In this way, Lasswell formula defines five types of communication

research: information sources, the content, media channels, audience, and effects (see Figure 9). Since Lasswell's model of communication is rather too simple, it has been widely criticized and modified (Braddock, 1958). However, we should not obscure the fact that it is a powerful model for communication research. For example, Wu et al. (2011) follow this line of thought to study who says what on Twitter. Their findings reveal that there is a striking concentration of attention on Twitter.

Insert Figure 9 here

In parallel to Lasswell's approach of modeling communication by the roles involved in the process, Shannon and Weaver (1949) developed a similar linear model. In Shannon and Weaver's mathematical model, an information source produces a message to be delivered by the transmitter. The transmitter translates the message into signals and sends it to the receiver. At the same time, noise would be produced by a noise source. Since the noise is mixed with the signal, the receiver receives the signal with a portion of noise. In the last step, the receiver translates the signal into a message, and sends it to the destination.

One main problem of these linear models of information flow is a lack of communication dynamics. In both Lasswell formula and Shannon et al.'s model, communication is described as a linear process without feedback. New communication models should be developed to overcome this shortcoming. To this end, Schramm and Osgood (1954) developed a circular model. However, the Schramm and Osgood model also has obvious weaknesses. They assume that two parties are equal in performing identical functions (e.g., encoding, decoding, and

interpreting). Thus, it conveys the feeling of equality in communication. Given the unequally distributed resources, power, and time, equality in communication is far from being true.

Another attempt to capture the interrelated features of communication is the Newcomb ABX model (1953). Against the backdrop of ignoring the interaction among individuals and the interplay between individuals and information, Heider (1946) proposes the balance theory and conceptualizes the cognitive consistency motive as a drive toward psychological balance. Following the balance theory, Newcomb (1953) formulates the Newcomb ABX model as "a very simple representation of the dynamics of communicative relationships between two individuals" (MaQuail & Windahl, 1993, p. 27). In the Newcomb ABX model, communication plays a key role in helping individuals maintain simultaneous orientations to an external object (e.g., the third person, or information). Two individuals (A and B) are orientated towards each other and an object (X). The inconsistency between A and B in their orientation towards X, under certain conditions, may stimulate communications, which in turn, restores the balance between A and B. Guided by similar intellectual sources, Festinger (1957) developed the cognitive dissonance theory, positing that simultaneously holding two or more conflicting information and choices could create psychological dissonance or distress for individuals. The dissonance will lead people to avoid contradictory information sources, and gives rise to selective exposure.

Insert Figure 10 here

However, it is important to bear in mind the weakness of this model. First, the ABX model focuses on communicative behaviors, but communication is only one way to resolving

uncertainty and discomfort caused by conflicting information. For example, on Twitter, if you dislike a tweets shared by one person, you may directly choose to end following him/her without paying any effort on initiating any communications with him/her. Second, the ABX model is primarily concerned with interpersonal interactions within social networks on the micro-level, while exogenous impacts on the macro-level are not sufficiently discussed.

Preserving the interactive feature of interpersonal communication, Westley and MacLean (1957) adapt the Newcomb ABX model to provide a representation of the much more complex situation of mass communication. They extend the Newcomb ABX model mainly in three aspects, to be specific:

_____

Insert Figure 11 here

_____

First, Westley and MacLean add mass media C into ABX model (see Figure 11). X is any social event or object of orientation in a social system, e.g., the information about U.S. election 2012; A stands for advocacy roles which are usually communicators who are the information sources, including elites, stars, organizations, etc.; B denotes behavioral roles who are the receivers of communication, e.g., audiences or the public who selectively seek information to satisfy their information needs and solve the cognitive dissonance; C are communication channels. As the agents of A, C selects and transmits information from A to B.

Second, in the Westley and MacLean's model, information selected by advocacy source A is encoded into X'; Similarly, information X' received by C will be decoded, interpreted,

encoded, and sending out to B, which converts the information from its original form and content to X''.

Third, Westley and MacLean add feedbacks to the ABX model. fBA is the feedback from the audience B to the original information source A (see Figure 11), e.g., a vote for a politician; fBC is the feedback from the public B to the mass media C, either by the way of audience research or audiences' letters; fCA is the feedback from mass media to information source. In addition, mass media C can directly get information without the aid of advocacy roles A.

It should be noted that, Westley and MacLean' work aims in supplying "a paradigm or model as a preliminary orientation to a theoretical system, rather than proposing a full-blown theory of mass communication" (Westley & MacLean, 1957, p. 31). It assumes that a minimum number of roles and processes are necessary in any mass communication. Thus, it may be misunderstood in various ways. For example, the communicators are not isolated from the social system, and they are always influenced by the political, social, cultural, and commercial forces. The political actors always have close relationship with mass media.

Another ingredient of communication study for constructing a general graphical model of information diffusion on information sharing websites (ISWs) comes from the diffusion model underlying the J-curve of news diffusion (Greenberg, 1964b). Based on the J-curve model, Greenberg (1964b) classified news events into five clusters (A, B, C, D, E) according to the awareness of news diffusion, which emphasizes the salience of news events (Gantz & Bradley, 2005). The other scholars suggest to classify news events into three types (A is in Type I; B, C, and D are in type II; E is in type III) according to the salience of the events (MaQuail & Windahl, 1993, p. 81).

Insert Figure 12 here

Type I are events which are of low public interest, but are very relevant to a special minority. For example, motivated by the relevance, residents in a community are very interested in the local news which reports social events of them. News stories of Type I usually will not get prominent media attentions. Since the information of a few events is very important for a group of audiences, they not only rely on the mass media as information source, they also tend to turn to their reference groups. Thus, the percentage of interpersonal source plays a more important role in spreading this kind of information.

Type II are average general news stories which are thought to be of general public importance and get a fair degree of prominence in mass media. Although this kind of information would be discussed by the public, they are not likely to be passed as information from person to person for a lacking of novelty (most people know about it), salience, and relevance.

Type III are events of very high significance, extreme urgency, and dramatic quality. The classic case is the Kennedy assassination in November 1963. Despite mass media allocate great attention to such news stories, interpersonal networks are also activated and play a more important role in spreading the information.

3.1.2 Information Sharing Websites (ISWs)

One unique feature of online information diffusion is the new characteristic of digital media. Information sharing websites (ISWs) refer to the social media through which users can

submit, share, and comment on information based on social network services (SNS), information aggregation tools, and search engines. By supplying diverse communication channels, ISWs underline the transformation of online information diffusion from mass communication to a participatory process. ISWs users spread information by both pushing and pulling information. People can "pull" information from the aggregators, the search engine, and their social networks. Meanwhile, people can also "push" information to the aggregators, search engine, and their social networks.

First, social network services are defined as web-based services that allow individuals to construct a personal profile (public or semi-public) within a bounded system, build social connections with a list of other users (either bidirectional friendships or directed following relationships), and present their lists of social connections to others within the system. After joining a social network site, users are prompted to identify others with whom they have a relationship.

Here, it is necessary to note the difference between ISWs and SNSs. The central concern of SNSs is to build and maintain social relationships. In contrast, ISWs are designed to spread information and facilitate democratic participation on the web. This research scope of information diffusion leads us to examine the information diffusion on Digg, YouTube, and Twitter, rather than that on Facebook.

Second, the information aggregator is a tool that demonstrates a list of popular information or recent information, which usually appears on the front page of the website. Based on users collaborative filtering of information diffusion, ISWs aggregate the popular information to the information aggregator. Therefore, the other users can browse the information directly from the webpage of aggregator, rather than from their friends in the social networks of the

website. Thus, the information aggregator supplies another alternative diffusion channel awaiting conceptual elaborations and empirical investigations.

_____

Insert Figure 13 here

_____

Third, the search engine is another important tool for information seeking on the Internet. Designed to search for information on the World Wide Web, search engine appears around 1993. Web search engines retrieve information by scraping the webpage on World Wide Web, and then they index the information and store the results. When users search a query (e.g., keyword) using the search engine, search engine examines its index according to the query supplied by users, and provides a list of best-matching web pages according to its criteria. Although, search engine came into being in the era of web 1.0, and usually is an independent website (e.g., Google, Yahoo, Bing, and Baidu). Now, search engines have been widely used in ISWs and become an important part for intentional information-seeking. For example, Twitter search becomes an important tool for users to find relevant tweets and people on Twitter.

3.1.3 A Communication Model of Information Diffusion—The ABXCT Model

Based on prior communication models of information flow and the change of online information system, especially the technical elements of information sharing websites (ISWs), I extended the ABXC model (Westley & MacLean, 1957) to theorize online information diffusion (see Figure 14). The main components of this new model follow the classifications of Westley and MacLean's conceptual model of mass communication (Westley & MacLean, 1957). There

are several modifications of this model according to the technical features. In addition to advocacy roles (A), behavioral roles (B, e.g., audiences), information (X), and communication channels (C), this new model includes the time (T) into the model.

Insert Figure 14 here

First, in this new model of information diffusion, the advocacy roles (A) are classified into two kinds—passive advocacy roles and institutional social media. Prior information sources heavily rely on mass media as diffusion channels, which grants mass media great power in the society. This agrees well with the notion of "the fourth estate"—Media as an independent social institution or a watchdog that ensures that other institutions (e.g., government, business, religion) serve the public. However, institution recently can use their social media account to release information to the general public. In this sense, Internet is supplying a competing communication channel for information diffusion. Thus, the development of Internet gradually frees the other institutions from their heavy rely on mass media for disseminating information to the public.

Second, in this new model, individuals can not only get information from the mass media, but also receive information from search engine, aggregators, and their online social networks. As noted above, information channels (C) in the information age have been extended by Internet. As a new kind of diffusion channel emerging in the information age, information sharing websites (ISWs) combine the functions of search engine, social network services, and information aggregators, and become "a new space of flows". The notion of "a space of flows" is coined by Manuel Castell (1996) to describe the space in which the time-sharing social practices

of material organization work through flows. E.g., Manuel Castell (1996) states that Internet is a space of flows, rather than a space of places. As "a new spatial form characteristic of social practices, the space of flows dominates and shapes the network society (Castells, 1996). If this is the case, as an important part of the place of flows, ISWs would have been shaping the quality of our information environment.

Third, the behavioral roles (B) or audiences have also evolved into two clusters of people. The first group is spectacle audiences who attentively follow the other diffusion channels and opinion leaders; the second group is opinion leaders who are social elites who attract public attention and shape the public opinion to a great extent. According to the classification of four estates, the clergy is the first estate, the nobility is the second estate, the commoners are the third estate, and the news media are the fourth estate. Following the logic of the fourth estate, Stephen D. Cooper (2006) suggests that bloggers serves as the watchdog of news media—the fifth estate. Further, William Dutton (2009) argues that the fifth estate is networked individuals enabled by the Internet in ways that can hold the other estates accountable, not simply the blogging community, nor an extension of the news media. Barry Wellman (2001) coined the term "networked individualism" to emphasizes the transformation of social networks from individual and place-based communities to personalized networking. Corresponding to the term "networked individualism" used by Wellman (2001), Dutton (2009) employs the notion of networked individuals to denote individuals linked together by Internet. Basically, this line of thought underlines the change from passive audiences to active and creative audiences.

Fourth, the attributes and interrelationships of information (X) also shape the diffusion of information. On one hand, there are many dimensions of the information, e.g., importance and relevance (Greenberg, 1964b). The J-curve model of news diffusion indicates that both the

information salience and diffusion channels have impacts on the size of diffusion. Thus, this model emphasizes the importance of information with the weight of lines. E.g., the bold line indicates the importance of the information. On the other hand, different information competes with each other or reinforces the diffusion of each other. Given a surplus of the supply of information and the limited capacity of public attention in the digital age, the competition of information is becoming extremely intense. However, the competition will never rule out the mutually beneficial relationship between some relevant information. For example, when one encounters one piece of information not well introduced in a Tweet, he/she would seek more relevant information.

Fifth, the last but not least element of this model is time (T). All the diffusion process happens within a social system over time. As it has been argued in classic diffusion models, is the temporal feature is one of the most important features of information diffusion. In Rogers' Book of Diffusion of Innovations, such temporal pattern has been well spelled as the S-shaped Curve. In the Bass diffusion model and the epidemic models, time has been explicitly recognized by employing the differential equation of time to describe the process of information diffusion. For information diffusion, what constrains the diffusion size is the limited capacity of public attention. Newcomb (1953) proposed the ABX model, and Westley & MacLean (1957) revised it by adding the information channels C. In general, the ABXC model illustrates the overall process of communication. Yet, it is weak in depicting the temporally evolving feature of information diffusion. To solve the theoretical puzzle of limited diffusion size, it's necessary to investigate the temporal patterns. First, the novelty of information decays along time. e.g., Dezsö et al. (2006) study the collective browsing activity to a news portal, and they find that access to most news items significantly decays after 36 hours of posting. According to the research of Karsai et al.

(2011), the burst of public attention to specific information tends to slow down its diffusion. Second, the time use pattern of individuals shape the diffusion process. E.g., by using a branching model which accurately describes the experiment of email communication, Iribarren et al. (2009) show that the large heterogeneity found in the response time of email communication is responsible for the slow dynamics of information diffusion for emails at the collective level. Further, collective attention is also featured by the burst phenomena. E.g., a news document's visitation peaks after a few hours (Vázquez, et al., 2006). That is to say, there is an apparent burst of public attention allocated to a specific piece of information.

Put together, I extend the classic dynamic communication model of ABX to ABXCT model by involving communication channels (C) and time (T). According to this ABXCT model, there is a stream of information flowing into the social system through communication channels, such as exogenous outlets (e.g., media channel, aggregators), social networks, search engines, to the public. With the aid of this model, I intend to summarize the driving forces of information diffusion on different levels of analysis in this dissertation in terms of multi-level perspective. The focus of this dissertation is the information diffusion on information sharing websites (ISWs). Thus, I will concentrate on the forces that drive ISWs users to spread information come from at least four aspects: exogenous impacts, interpersonal effects based on social networks, individual attributes, and content features.

Viewing graphic communication models as aids to thought, media scholars should be aware of the drawbacks of communication models. Although communication models point to areas of strength and weakness in our knowledge, it must not be viewed as a theory but as a preliminary step to stimulate further investigations. Although this ABXCT model captures the primary elements of the information diffusion process on ISWs, and pictures a generally

conceptual map for information diffusion research, it is not sufficient for depicting diffusion mechanisms in details.

3.2 Research Design

The driving force in information diffusion refers to the influence that promotes the spreading of information. As aforementioned, there are at least four kinds of driving forces for information diffusion: exogenous impacts, internal influence (e.g., interpersonal effects based on social networks), individual attributes, and information characteristics (e.g., the category of information). To deepen our understandings about the size of information diffusion, this study will start from testing the threshold hypothesis of interpersonal effect, and then conceptualize the influence of aggregators, as well as the temporal patterns of information diffusion, and finally probe the burst patterns of information diffusion, which constitute the skeletons of three subsequent studies of information diffusion on information sharing websites (ISWs).

First, for a specific piece of information tweeted by microblog users, the main diffusion channels are social networks based on the social networking service supplied by microblogs. Second, for a news story on Digg, there are two diffusion channels: the information aggregator and the social networks. Third, compared with Digg and Twitter, the diffusion channels on YouTube are even more diverse. For a video on YouTube, in addition to social networks and the information aggregator, recommendations from related videos and external sources also play an important role in spreading YouTube videos.

Following the logic above, I organize the following chapters as follows: Chapter 4 focuses on the information diffusion on Sina Weibo in order to test the threshold hypothesis of interpersonal effect. Chapter 5 conceptualizes the interaction between users and information aggregators as a process of collective gatekeeping, and analyzes the underlying diffusion

mechanisms with the perspective of information cascade theory, adopting the network data of news diffusion on a social news website (Digg). The analysis in both chapter 4 and chapter 5 reveals the extremely high temporality of information diffusion on ISWs featured by strong bursts. Thus, Chapter 6 devotes to the analysis of the origin of burst for public attention, employing the diffusion data of YouTube videos. Ultimately, the dissertation attempts to offer general understandings about how these particular studies fits into a broader theoretical lens, and how the framework I propose here travels to further investigations.

Chapter 4. Information Diffusion on Microblogs: Testing Threshold Hypothesis of Interpersonal Effects

4.1 Introduction

Microblogs gain great popularities in the past years. Created in 2006, Twitter generates over 340 million tweets daily. By 2012, there are 500 million registered users on Twitter, and 200 million of them are active users, helping Twitter gain worldwide popularity. In China, there are 400 million registered users on Sina Weibo by February 2012, and 300 million of them are active[2].

According to the ABXCT model, the most prominent change of communication process is the participation of networked individuals into the online information diffusion process. Contrary to the role of passive audience, networked individuals actively use microblogs to share information and express their opinion. e.g., Twitter has been extensively used for political deliberation (Conover et al., 2011). On microblog, individuals can build up their social connections by following any other they are interested in. The interpersonal effects played an important role in shaping the diffusion size. Additionally, different from the other social media, microblog makes elites more reachable online. To date, political leaders, stars, journalists, and commoners widely adopt microblog to release information and communicate with the public. To promote their products, business organizations maintain their official accounts on Twitter. Therefore, microblogs play an increasingly important role for information diffusion, and they become ideal test grounds to study the interpersonal effects in online information diffusion.

Guided by the ABXCT model, especially its emphasis on networked individuals, the theoretical relevance of this study presented here is framed within the threshold hypothesis of

[2] http://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_users

interpersonal effects (Rogers & Kincaid, 1981; Valente, 1995). Scholars of diffusion of innovations generalize that an individual is more likely to adopt an innovation if more of the other individuals in his or her personal network have adopted previously. In prior studies of information diffusion on microblogs, it is assumed that interpersonal influence is crucial for information diffusion. Yet little is known about to what extent interpersonal effects shape individuals' behavior of information diffusion, and what the driving forces are. To gauge the role interpersonal networks play in online information diffusion, this study focuses on how information is shared on microblogs (e.g., Twitter and Weibo).

In the light of the threshold hypothesis of interpersonal effects, I claim that the network threshold of interpersonal effects shapes the size of information diffusion on microblogs. To frame and test this central concern, this study is organized as following: First, the next two sections will provide the theoretical underpinnings of this study by discussing the literature on diffusion mechanisms on microblogs and the specific features of microblog. Second, specific research hypotheses are formulated. And third, the research hypotheses will be examined by two datasets of information diffusion on microblogs. Finally, the theoretical generalizations will be summarized and discussed.

4.2 Information Diffusion on Microblogs

4.2.1 Microblogs and Information Diffusion

Microblog is a broadcast medium in the form of blogging (Microblogging, 2013). In general, the information on microblogs can be classified into three categories: retweets, conversations (e.g., replies or comments), and monologues. There are two other special features for the information spreading on microblogs: hashtags and URLs. A hashtag is a word or a phrase prefixed with the symbol #. Created organically by Twitter users, hashtags are employed

to tag the information of the same categories or topics. For example, #ows denotes the tweets of Occupy Wall Street on Twitter. Accordingly, three general kinds of information can be tracked on Microblog (e.g., Twitter, Sina Weibo). The first one is the diffusion of a specific tweet which is disseminated in the form of retweets or reposts. In this case, for one piece of information, there is only one submitter who initially writes the original information. The second one is the diffusion of URLs. In this case, there may be more than one original submitter who independently shares the information containing a specific URL. The third one is the diffusion of hashtags. It should also necessary to note that one tweet may also contain both hashtags and URLs. However, this is not a problem if we are clear about which kind of diffusion we want to study.

Technically speaking, microblogs are also featured by its always-on persistence (e.g., Twitter maintains an updated information stream), light-weight scripting (e.g., a tweet is a short message containing no more than 140 characters), open infrastructural base, and portable back-end interface. Sina Weibo and Tencent Weibo are two typical microblogs in China and they implement basic features of Twitter. However, as the hybrids of Twitter and Facebook, they also allow users to comment to others' posts. Note the other leading social networking websites, such as Facebook, MySpace, LinkedIn, Yahoo Pulse, Google+, also have their own microblogging features, e.g., users of Facebook can update their status like microblog users.

In addition to technical advantages, microblogs have several other advantages for information diffusion. First, as an important form of information sharing websites (ISWs), social connections (i.e., following and followed relationships) on microblog are mostly weak ties. According to the theory of weak ties (Granovetter, 1973), Twitter is for suitable for information dissemination. Second, the social relationship on microblogs needs little or even no emotional

attachment, which reduces the cost of information diffusion, and makes microblog easier to maintain than the other social media. Third, people from different places around the world are connected together to become a networked public. Thus, in the digital space of microblog, no one is really isolated. Everyone can make his/her effort in spreading information to the other users they want to. Microblog can be viewed as an information market (Tumasjan, Sprenger, Sandner, & Welpe, 2011).

4.2.2 Diffusion Mechanisms on Microblog

To understand the size of diffusion on microblogs, it is necessary to look into the driving forces of information diffusion. Basically, there are four kinds of influential factors: exogenous impacts, interpersonal influences, individual attributes, and the features of information. In this dissertation, the distinction between exogenous impacts and endogenous influence are based on the boundary of online social networks. The impacts outside of online social networks can be classified as exogenous impacts. The influences within social networks are endogenous influences, such as interpersonal influences, personal attributes, and information characteristics.

First, exogenous factors play an important role in information diffusion. Firstly, information sharing behaviors are influenced by external events and media coverage. Lehmann et al. (2012) track the propagation of hashtags in the Twitter social network and find that epidemic spreading plays a minor role in hashtag popularity, whereas the diffusion of hashtags is mostly driven by exogenous factors. Secondly, information sharing websites (ISWs) are information systems with information aggregation mechanisms. In addition to interpersonal effects solely based on ego networks, aggregated information may also impact the size of information diffusion. Microblogs (e.g., Twitter) seem only have an implicit mechanism for aggregating information. The size of the followership and the rate of retweets may represent the

Twitter's "currency" and provide it with its own kind of a "price system". Thus, microblog users tend to be influenced by the aggregated popularity which manifests itself as trending topics and items returned by search engines of microblogs.

Second, interpersonal effects have impacts on the size of information diffusion. Interpersonal influences in diffusion research concern the extent of accessing information through social networks. According to this definition, interpersonal effects for one particular piece of information can be roughly measured by the proportion of people learning information through social networks. The J-curve model of news diffusion indicates that there is a nonlinear relationship between percentage of interpersonal source and the size of diffusion (Greenberg, 1964b). For the information of public interest, interpersonal effects have a positive relationship with diffusion size. While for the information of local relevance, interpersonal effects have negative relationship with diffusion size. Therefore, there is a critical point for interpersonal effects on information diffusion. In general, the turning point could be decided by the overall popularity of information.

Third, individual attributes also influence the process of information diffusion. Ardon et al. (2011) investigate the effect of initiators on the popularity of topics, and find that users with a high number of followers have a strong impact on popularity. Tonkin et al. (2012) find that Tweets offered by well-known and popular individuals were more likely to be retweeted. Lou and Tang (2013) provides evidence for the theory of structural holes, e.g., 1% of Twitter users who span structural holes control 25% of the information diffusion on Twitter. In contrast to the number of followers, Huberman (2009) show that the number of friends is the actual driver of Twitter user's activity. Further, Gonçalves et al. (2011) find users can entertain a maximum of 100–200 stable relationships. Thus, individual attention in the online world is also limited by

cognitive and biological constraints, as predicted by Dunbar's theory (Dunbar, 1992; Gonçalves, et al., 2011).

Fourth, the attributes of information also has close connections with the diffusion size. Wu et al. (2011) show that there exists a strong association between the content and the temporal dynamics of information. Wu et al. (2011) find the information containing more words related to positive emotion, leisure, and lifestyle can survive for a longer time. Conover et al. (2011) find that political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users.

4.3 Threshold Hypothesis of Interpersonal Effects

The most important assumption of threshold theory is the utility-cost assumption. Usually, individuals have to overcome their resistance of making a decision on binary decisions. Resistance could be described as the cost minuses the utility. If the utility of the engagement is larger than the cost of engagement, individuals tend to make up their decisions of engagement. One major driving force to reduce the resistance of engagement is the interpersonal effect. Threshold models conceive that the utility of engagement is a function of the fraction of people in a social system who have already engaged in the behavior (Granovetter, 1978; Granovetter & Soong, 1983, 1986, 1988). With the gradual growth of the fraction of people who engage in the dynamic behaviors, individuals' resistance gradually diminishes. When interpersonal effects are larger than the cost of engagement, individuals tend to make up their decisions of engagement. Threshold is the proportion of people engaging in the activity when their cost is equal to their utility of engagement. If the individual equilibrium of utility and cost is reached, its threshold can be measured. As a measurement of interpersonal effects, threshold is featured by its duality: A larger threshold suggests a larger local pressure as well as a larger obstacle or cost.

Insert Figure 15 here

In the perspective of communication networks, later research shifts their focuses to network threshold. "Network provide a certain degree of structure and stability in the predictability of human behavior" (Rogers, 2003, p. 363). Everyone has his/her personal networks or, in term of network analysis, the ego network. For any individual, when his/her utility of engagement equals the cost of engagement, network threshold is the proportion of people in its personal network who have engaged in the behavior. The theoretical significance of network threshold calculated based on individuals' ego network captures the local influence of social networks. The relationship between local influence and global influence has been the interest of network studies. Thus, network threshold contributes to our understanding about the local and global driving forces.

Thus, threshold theory grants us a new lens to gauge interpersonal effects on information diffusion. For example, prior studies on diffusion of innovations generalize that an individual is more likely to adopt and innovation if more of the other individuals in his or her personal network have adopted previously (Rogers, 2003; Rogers & Kincaid, 1981). Thinking about the situation of information diffusion, each individual has to make a decision of spread a specific piece of information or not when he/she is exposed to the information. This present study focuses on the diffusion of information which is the aggregated outcome of individual behaviors. Aggregate all the information diffusers' threshold, the mean threshold can be calculated. The measurement of network threshold captures the repeated exposures to specific information in local networks. Concerning the activation state of information diffusion in the local environment,

network thresholds gauge the local influence of ego networks. Mean threshold for one information diffusion actually captures to which extent the local network has been activated to spread the information. Based on this rationale of threshold models, the stronger the threshold, the larger probability individuals engage in disseminating the information, and therefore, the information could spread further. Therefore, the hypothesis of interpersonal effects for information diffusion can be inferred:

H1a: Mean threshold of interpersonal effects significantly influences the size of diffusion

However, as the J-curve model illustrates, there is a nonlinear relationship between interpersonal effects and diffusion size (Greenberg, 1964b; MaQuail & Windahl, 1993). Thus, the direction of mean threshold' influence on diffusion size deserves more attentions here. Firstly, a larger threshold implies larger interpersonal effects in the local social networks. Thus, it could be inferred that there is a positive relationship between threshold and diffusion size. Secondly, the information of minority interest will not get prominent treatment in the exogenous sources, and they mainly infect those who are relevant to the information. Since a lack of knowledge, they tend to be selectively noticed by the relevant group. For example, a microblog user tweets about his/her daily life, those who responds to him/her tends to be his/her friends who are familiar with him/her or relevant to the events discussed in such personal talk. Most of the information is tweeted by grassroots who have a significantly smaller size of followers, and their community tends to be closely connected. With the growth of salience and the decline of direct relevance, there is a tendency that the information diffusion relies less heavily on interpersonal effects. Therefore, there is a negative relationship between threshold and diffusion size. In all, the J-curve model integrates those two points together: for the information of less importance, there is a negative relationship between interpersonal effects and diffusion size; while for those

information of larger salience, there is a positive relationship between interpersonal effects and diffusion size.

H1b: the relationship between mean threshold and diffusion size is curvilinear.

For a specific tweet submitted by a concrete microblog user, the branching process of information diffusion is anticipated to be related to the attributes (e.g., popularity, activity) of the submitter. The popularity not only includes the number of followers, but also whether the user has been verified by Sina Weibo. Similar to Twitter, there are two types of user accounts on Sina Weibo, regular user accounts and verified user accounts. A verified user account typically represents a famous public figure (e.g., celebrities, sports stars, journalists) or a well-known organization (e.g., Government, companies, media). According to the theory of opinion leader and the theory of social influence, I derive the following hypotheses:

H2a: the popularity of submitter is positively related to the size of information diffusion.

H2b: the activity of submitter is positively related to the size of information diffusion.

Another important influence comes from audience interest. One indicator of the audience interest is the number of comments received by the tweets. As a measurement of revealed preference, the number of comments is expected to be positively related to the size of information diffusion.

H3: the number of comments is positively correlated with the size of information diffusion.

Audiences can be divided or even fragmented into several clusters, according to different interest to different genres of information. The content plays a crucial role for information diffusion. Given the difficulty of the diversity and complexity, it is almost impossible to model the impact of content on information diffusion in an explicit way. One practical way is to control

the category of information. The mission of microblog in nowadays is believed to spread information and express personal opinions. Thus, researchers are interested in whether opinion expression occupies a salient position in the use of microblogs. To translate these concerns into the language of information diffusion, this study formulate the following research question and hypotheses:

H4: the category of information has significant influence on the size of information diffusion.

RQ1: whether tweets of opinion expression can spread to a wider audience than the other categories of information?

Further, in addition to the category of information, Sina Weibo enable users to insert rich media like images, videos & music, emoticons, which greatly increase the information-carrying capacity of microblogs, and make the information attractive. Given this features of microblog, it is necessary to control this influence in the study of information diffusion by proposing the following hypothesis:

H5: the information embedded with rich media (e.g., url, image, video, and emotion) tends to increase the size of information diffusion.

The flow of information penetrates in the social graph constructed by friendship or following/followed relationships. Threshold concerns interpersonal effects or to what extent the local networks (i.e., individuals' ego networks) are activated. However, information diffusion through social networks is a chain-reaction process. The contagion may " infect" the local community effectively, and stop trigger further diffusions, which "signals the existence of structural bottlenecks" (González-Bailón, et al., 2013). In terms of network threshold, the structural bottlenecks are those "structural holes" of higher network threshold. The theory of

structural holes implies that the nodes in the middle have an information advantage by acting as gatekeepers who filter the information and decide whether to spread the information to a further depth. The depth of diffusion highlights to what extent information can penetrate social circles. If the bridging nodes have a higher resistance to retweet the information, the diffusion has a higher propensity to stop with informing a limited size of audiences. Thus, information diffusion as a branching process heavily relies on the bridging nodes to spread the information to the deeper depth, and a larger diffusion depth tends to make the information spread further.

H6: The depth of information diffusion has positive influence on the size of information diffusion.

Last but not least, the lifetime of diffusion concerns the temporality of public attention. By definition, if the public attention is sustainable rather than temporal, the information can diffuse to a wider audience, and if the information can easily go through different social circles, it tends to get viral. Based on these simple but strong rationales, the following hypotheses can be derived:

H7: Lifetime of information diffusion is positively related to the size of information diffusion.

4.4 Method

4.4.1 Data Collection & Sampling

As the largest microblog in China, Sina Weibo gains great popularity in the past years We randomly generate 300 million Weibo user ids in June 2012, and detect the existence of these users, which yields a random sample of 62316 Weibo users, and all of their tweets (N = 5036596) are crawled. Among these tweets, 3185574 of them are retweets (63.2%), and the other 1851022 tweets are original tweets (36.8%) tweeted by these 62 thousand users. For the retweets,

27158 original tweets retweeted by these 62 thousand Weibo users have been deleted (0.85%). Besides, 2186849 tweets have been retweeted by these Weibo users.

_____

Insert Figure 16 here

_____

To quantify the degee distribution of information diffusion. 300000 tweets tweeted by these 62 thousand Weibo users, as well as 300000 tweets retweeted by those users, are sampled from the original data. Through the API of Sina Weibo[3], I collected the number of retweets of these 600000 tweets. In this sample, the largest diffusion size is 417116.[4]

Further, to study the information diffusion with regression models, I randomly sample 2500 pieces of information whose diffusion size is larger than 100, and 1000 pieces of information whose diffusion size is smaller than 100.

4.4.2 Measure

Size of Diffusion. Here, the dependent variable –size of information diffusion is measured by the times of retweets. According to the data of retweeted times for 600 thousand tweets, the distribution is highly unequal. If the distribution of diffusion size (*Mdn* = 335, *M* = 1255.8, *SD* = 76156) is plot in a log-log plot, a scaling relationship can be identified (See Figure 17). Such inequality of diffusion is featured by a long tail, in accordance with prior studies on scale-free distribution and the underlying mechanisms of preferential attachment.

[3] http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3_V2/en
[4] http://weibo.com/1686659730/ygxCRrnAO

Lifetime of Diffusion & Depth of Diffusion. In addition to the size of diffusion, the other attributes of information diffusion, such as the lifetime of diffusion, and the depth of diffusion, have also been calculated. First, the lifetime of diffusion is measured with the one day as a unit of time (*Mdn* = 9 days, *M* = 18.7 days, *SD* = 29.0). Second, the depth of diffusion is calculated from the information of diffusion network. Diffusion depth is the largest network distance from retweets to the original tweet. I calculate the distance from each node to the submitter, and get the largest network distance—diffusion depth (*Mdn* = 4, *M* = 4.9, *SD* = 2.9).

Insert Figure 17 here

Network Threshold. For different information, the network threshold for each individual is calculated by tracking the diffusion process. Note that since the distribution of network threshold is usually skewed. A logarithm transformation is necessary. Assume that the distribution of x is lognormal, which implies that the distribution of x is highly skewed, however, it can be normalized by taking the logarithm form x': $x \to x' = \log(x)$, Thus the mean value of x' is $mean(x') = \frac{\sum Log(x)}{n}$. Note that, n is the number of variable x. In this way, we can get the mean value of x: $mean(x) = Exp(\frac{\sum Log(x)}{n})$.

After the individual network thresholds have been calculated for each piece of information, aggregate them to the information level, the mean-threshold for a piece of information can be calculated (*M* = 0.15, *SD* = 0.04). Mean-threshold therefore measures to what extent diffusers' local information environment has activated.

It's necessary to note that, there are cases of zero-threshold which has been discussed in prior literature (Valente, 1995). Thus, the ratio of zero-threshold can also be calculated (M = 0.23, SD = 0.19). In contrary to interpersonal effects about repeated and multiple exposures, zero-threshold indicates that individuals spread the information without direct exposure to the information through his/her interpersonal networks.

The Category of Information. The information retweeted on Sina Weibo has been humanly coded by two coders into nine categories: fun, life encyclopedia, star, movie, news, mood, hobby, and product (Cohen's kappa = 0.987). The following Figure 18 shows the distribution of different categories. For the salient information which has been retweeted more than 100 times, the most popular information come from Fun (32.1%) and Life encyclopedia (23.3%); for the trivial information which has infected less than 100 people, the most popular information are mood (27.3%) and opinion (17.9%). This is consistent with prior observations. E.g., Yu et al. (2011) find that in China, the trends are created almost entirely due to retweets of media content such as jokes, images and videos, whereas on Twitter, the trends tend to have more to do with current global events and news stories. The other features of information have also been coded, such as whether the information contains pictures (Cohen's kappa = 0.968), videos (Cohen's kappa = 0.964), URLs (Cohen's kappa = 1), the icon of comment labeled with the symbol of "@" (Cohen's kappa = 0.962), and the emotion icon (Cohen's kappa = 0.872).

Insert Figure 18 and Table 3 here

4.5 Results

The main concern of this study is the threshold hypothesis of interpersonal effects, which claims that the aggregated network threshold has significant influence on the size of information diffusion. The dependent variable is the size of information diffusion. By testing the hypotheses, this study aims to shed light on our understanding about the interpersonal effects in information diffusion. To test the hypotheses, two random samples of information have been used to run linear regressions. Model 1 focuses on the sample whose diffusion size is larger than 100, and model 2 is based on the sample in which diffusion size is smaller than 100. The results have been demonstrated in Table 4.

Insert Figure 19 here

H1 asserts that mean threshold of interpersonal effects has significant effect on diffusion size, and such influence is curvilinear. First, as I have introduced in the measurement section, the mean value of zero-threshold ratio is 0.23, which suggests that for the diffusion of specific information, 77% of information diffusion are based on direct interpersonal sources. In model 1, there is a positive influence (Beta = 0.06, sig < 0.001); and in model 2, there exists a negative influence (Beta = -0.46, sig < 0.001). To visually demonstrate such nonlinear relationship, following the tradition of J-curve model proposed by B. S. Greenberg (1964b), I plot the scatters of mean threshold of interpersonal effects against diffusion size, and draw two smooth curves using the method of Lowess (Cleveland, 1981).

As Figure 19a illustrates, the relationship between mean threshold and diffusion size could be described as a U-shaped curve which is an approximation of the original idea of J-curve. However, it is not a perfect J-curve, given a relatively small mean threshold for popular information (e.g., diffusion size > 100). Thus, H1a and H1b are confirmed.

H2 concerns the attributes of submitters (e.g., seed nodes). In this study, I look into whether the popularity and activity of the submitters could trigger a larger diffusion. Two measurement of popularity have been adopted: first, whether the submitter's social identity has been verified as a celebrity, second, the number of followers. Both model 1(Beta = -0.09, sig < 0.001) and model 2 (Beta = -0.002, sig > 0.05) indicate that a verified identity as celebrity has no positive influence on diffusion size. Model 1 shows that the number of followers has no significant influence on diffusion size (Beta = -0.023, sig > 0.05), and model 2 shows the number of followers has positive influence on diffusion size (Beta = -0.178, sig < 0.001). Thus, for H2a, only the influence of number of followers has been confirmed. On the contrary, both model 1 (Beta = 0.143, sig < 0.001) and model 2 (Beta = 0.069, sig < 0.001) indicate that the activity of submitter is positively related to the size of information diffusion, and therefore, H2b is confirmed.

H3 argues that the number of comments is positively related with the size of information diffusion. Both model 1 (Beta = 0.529, sig < 0.001) and model 2 (Beta = 0.177, sig < 0.001) indicates that there is a scaling relationship between the number of comments and diffusion size, i.e., the logarithmic number of comments is found to be proportional to the logarithm of the diffusion size (see Figure 20). Thus, H3 is confirmed.

Insert Figure 20 and Table 4 here

H4 concerns the influence of information categories. In the regression models, the information categories are dummy variables with the category of opinion expression as the baseline group. Model 1 demonstrates that information of opinion expression can be significantly diffused to a wider audience than the information about star, product, and life encyclopedia. Model 2 also indicates the information of opinion expression is more popular than that of star. However, the diffusion size of the information in the category of opinion is not significantly different from the information of fun, movie, mood, hobby, and news. Thus, H4 is partly confirmed. Based on this finding, I can also answer research question about whether tweets of opinion expression can spread to a wider audience than the other categories of information.

H5 concerns whether the information embedded with rich media (e.g., URLs, images, videos, and emotional icons) could diffuse further. The results from model 1 indicate that only the information with URLs can significantly get more popular than information without URLs. Thus, this hypothesis is basically rejected.

H6 and H7 concern the influence of diffusion depth and duration on diffusion size. Both model 1 and model 2 confirm the strong, significant, and positive influences of diffusion depth and lifetime. Thus, H6 and H7 are both confirmed.

Further, although the nonlinear relationship can be captured by both two separated regression models and the scatter plot, the results appears to be a bit noisy. One explanation is the interpersonal effect is modified by the topological structure of diffusion networks. In this

study, I measure the penetration power of information diffusion using diffusion depth. It is measured by diffusion depth. As it will be shown that diffusion depth has significant influence on the diffusion size. To probe the moderation effect of diffusion depth, I build up the third model (see Table 4) by adding in an interaction term—Diffusion depth * Mean threshold (Log), which is negative and significant (Beta = -0.01, sig < 0.01). To visualize the moderation effect, we present Figure 19b here, and it demonstrates that with the increase of diffusion depth, the negative relationship between interpersonal effects and diffusion size is reversed, and the larger diffusion depth is, the stronger interpersonal effects on diffusion sizes are.

4.6 Conclusion and Discussion

Put together, this study offers an empirical test for the threshold hypothesis of interpersonal effects. The analysis confirms the threshold hypothesis of interpersonal effects. For the information of modest importance, smaller interpersonal effects are needed to trigger a larger diffusion; for the information of strong importance, larger interpersonal effects are necessary to make the diffusion spread to a wider audience. These findings render support for the classic J-curve model (Greenberg, 1964b) in the digital world. Thus, this study supplies a baseline to understand social networking services in social media.

The distribution of diffusion size, featured by a long-tail, is highly unequal. A few fraction of information diffuses to a large proportion of people, while most information could not spread to more than 100 people. The findings of J-curve model of interpersonal effects indicate that interpersonal effects tend to *intensify* the inequality of information popularity in terms of retweets. E.g., the findings confirm that for the information of low audience interest, the strength of interpersonal effects is negatively related to diffusion size. For the information of high audience interest, the strength of interpersonal effects is positively related to diffusion size.

In addition to the unequal distribution of diffusion sizes and its relationship with interpersonal effects, the depth of diffusion has positive influence on the diffusion size. Compared with diffusion sizes, the depth of diffusion is extremely small. E.g., the largest diffusion depth is 26, and the largest diffusion size is 417116. Information diffusion on microblogs is a branch process, and its width is much bigger than its depth. Although the massive online social network of social media is usually a small world, an average diffusion depth of 4 is not enough.

The limited diffusion depth implies that most information are trapped in the local community, and therefore it underscores the structural trapping which has been emphasized in prior studies. The proposition of structural trapping suggests that dense communities with few outgoing links naturally trap information flow (J-P Onnela et al., 2007; Weng, et al., 2013). In the seminal work on weak ties, Granovetter (Granovetter, 1973) proposes that weak ties pass information to a larger number of people thus information can traverse a longer social distance (path distance). In Onnela et al.'s study of mobile communication networks, they find that mobile social networks are "robust to the removal of strong ties" but fall apart if the weak ties are removed, "resulting in a dynamic trapping of information in communities" (J-P Onnela, et al., 2007). In Weng et al.'s study about the information diffusion on Twitter, they also find such structural trapping in terms of community concentration (Weng, et al., 2013). The findings of the positive relationship between diffusion depth and diffusion size suggest that the viral information tends not to be trapped by the local community, and it can relatively easily permeate through more communities.

The limited value of diffusion depth implies the potential of online information diffusion lies in the percolation power of the information. Information diffusion on microblogs decays

quickly with the depth of information diffusion, which may suggest that individuals of the same audience interest tends to cluster together, and thus the boundaries among social circles of different tastes block the information from propagating within the online social networks of microblogs.

Insert Figure 21 here

The pattern of short diffusion depth also illustrates the weakness of the information diffusion on microblogs. The diffusion mechanism for a specific piece of information is mainly based on interpersonal networks. Therefore, the information diffusion for concrete information is a *bottom-up* process. To make effective diffusion, a *top-down* diffusion process is necessary. As it has been mentioned in the literature section of microblogs, aggregated popularity manifests itself on microblogs in the form of information aggregators and search engineers on both Twitter and Sina Weibo [5].

To understand the limited diffusion size for information diffusion on social media, especially the temporality of information diffusion, an analysis of the temporal patterns of information diffusion is also necessary. To do so, I sample 8000 tweets from Sina Weibo, and analyze the temporal patterns of their information diffusion. First, I find the lifetime for most of the information is quite limited. I calculate the lifetime using one day as the unit of analysis, and find that the average lifetime is 8.78 days which is only a bit longer than one week, and the median lifetime is 2 days (See Figure 22).

---

[5] http://huati.weibo.com/

Insert Figure 22 and Figure 23 here

Further, the information of such small lifetime is usually accompanied by a drastic burst of public attention, and then dies out soon. To demonstrate the burst phenomena of public attention allocated to the information, I calculate the peak fraction which is the proportion of the daily information diffusion happens in the peak day. The results indicate that the average peak fraction is 0.827 (See Figure 23), which suggests most of the information diffusion for one specific piece of information (e.g., a tweet) happens within one day. Consistent with this finding, recent research also demonstrates a strong burst phenomenon in the diffusion of hashtags. For example, Glasgow and Fink (2013) study the Hashtag lifespan for the tweets of London riot. They find that the half-life span is less than 24 hours. Half-life is the time required for a quantity to fall to half its value as measured at the beginning of the time period. For example, the hashtag of #riotcleanup last for 52 days (lifespan), but its half-life is less than 20 hours.

Put together, this study confirms the threshold models of interpersonal effects for online information diffusion on microblogs. Further, the patterns of limited depth of diffusion networks and short lifetimes with strong and early bursts illustrate the potential importance of analyzing communication networks as well as the temporal patterns. Thus, further attention should be put to the aggregated popularity and the bursts of online information diffusion.

Chapter 5. Collective Gatekeeping of Information Diffusion on Social News Website

5.1 Introduction

Research on news diffusion reveals that people learn news largely from two sources: interpersonal contacts and news media (DeFleur, 1987; Funkhouser & McCombs, 1971; Greenberg, 1964a, 1964b; Larsen & Hill, 1954; Miller, 1945). However, scholars have long debated the relative importance of the two sources. Recent studies of the emerging social news websites (SNWs) have portrayed interpersonal sources as the primary driver of news diffusion in cyberspace (Lerman & Ghosh, 2010; Shao, et al., 2009; S. Wu, et al., 2011). I contend that, while the open and participatory platforms of ISWs may indeed bring people back to the center stage of news diffusion, these studies are likely to inflate the importance of interpersonal sources because they overlook the collective gatekeeping function performed by SNWs as an effective driver of news diffusion.

Informed by the ABXCT model, the technical features of social media has profound impact on the diffusion process. In addition to the social networking services, social media also supplies powerful information aggregation services. News aggregation tools usually appear on the front page, often as sidebars on the right hand side of the page (i.e., the webpage of news aggregation). The list of popular news items is selected from news stories contributed by users of the website, with the selection process usually automatically based on some measure of "popularity" (e.g., number of clicks, comments, or forwards that each of the stories receives up to the time point).

I conceptualize the information aggregation process as "collective gatekeeping" and the participants (i.e., users) as "collective gatekeeper". Collective gatekeeping refers to news aggregation tools provided by SNWs (e.g., list of top/hot stories, list of popular keywords, and

summary statistics of stories shared) based on collaborative filtering of news by users through voting, comments, recommendations, and other means. Collective gatekeeping is collective because, firstly, on the news level, news stories compete with each other for public attention; and secondly, on the individual level, everyone has the chance to act as a gatekeeper by making decisions about sharing news or not.

It should be noted that traditional news media also provide listings of hot news (e.g., headline stories) on their websites. However, headline news on news media websites (e.g., cnn.com) is compiled by professional editors/reporters based on their perceived timeliness, importance, or other criteria of news values (Gans, 1979), whereas collective gatekeeping comes from the wisdom of the crowd. Given this, I consider the former a *top-down* process of news production and dissemination and the latter a combination of both *top-down* (because the news aggregators, i.e., SNWs, act similarly to traditional media websites in news dissemination) and *bottom-up* (because production/selection of news is contributed by users) processes. Therefore, collective gatekeeping is governed not only by the traditional determinants of news diffusion such as source credibility (Gans, 1979; Hovland & Weiss, 1951) but also by the new forces of news diffusion such as citizen journalism (Goode, 2009).

News aggregated on SNWs also differs from news shared among friends on social network sites (SNSs) such as Facebook. The difference lies in the fact that news shared on SNSs comes from friends whom the receivers know on an individual basis, whereas the list of aggregated news comes from the aggregated community with the identity of original sources disappearing after the aggregation process. In social network terminology, the former (i.e., news from friends) involves production, dissemination, and consumption within a *local* community, in which everyone knows each other, whereas the latter occurs on a *global* scale, among members

who are largely strangers (Kossinets & Watts, 2006; Scott, 1988). Therefore, comparing the relative importance between the two competing sources will help not only determine the "true" (or more accurate) contribution made by interpersonal source to news diffusion, but also shed insights into a central question in social network research, i.e., the relative importance between local and global influences on individuals (Robins, Pattison, & Woolcock, 2005; Watts, 2004). Thus, I formulate the main research question of this study: Which is the primary driving force of news diffusion on SNWs, interpersonal sources or collective gatekeepers?

5.2 News Diffusion and Social News Websites

5.2.1 Traditions and Limitations of News Diffusion Research

News diffusion is the process of news flowing from news sources to audiences. As a common body of knowledge of the community, news stories filter into community life and inform relevant people through mass media and interpersonal channels. On the aggregate level, news diffusion is the cumulative growth of the number of news knowers; on the individual level, news diffusion concerns when and how individuals learn about news events. Beginning in the 1940s, increasing during the 1960s-1980s, and decreasing in the 1990s, classic studies of news diffusion compared media effect and interpersonal effects by studying the diffusion channels, cumulative diffusion curves, and the determinants of news diffusion (Larsen & Hill, 1954).

Concerned with estimating media effects, interpersonal effects, and the size of news diffusion, classic research on new diffusion generally found that people learn breaking news first from interpersonal sources but other less important news first from the mass media. This is characterized as the J-curve model (DeFleur, 1987; Greenberg, 1964b). As the first research of news diffusion, Miller's study in 1945 established the tradition of assessing the impacts of various diffusion channels (Miller, 1945). Focusing on the diffusion of the news of Roosevelt's

death among143 students of Kent State University, Miller (1945) found that 85% of the students learned the news by word of mouth. Later, Greenberg (1964b) studied 18 news events in terms of size of diffusion, media differences, and found that the relationship between the extent of news diffusion and the percentage of interpersonal sources could be described as a J-curve. Based on the J-curve model, Greenberg (1964b) classified news events into four categories, which emphasizes the salience of news events (Gantz & Bradley, 2005).

News diffusion research constitutes one of the most long-standing traditions in communication research (more than 60 years), yet, limited by methods and theoretical framework, research on news diffusions has been criticized for producing very few theoretical generalizations, and most of the accumulated contributions are debatable (DeFleur, 1987). For example, although news diffusion studies concern the impact of diffusion channels, the problem of whether interpersonal communication is an important channel for news diffusion has remained largely unresolved (DeFleur, 1987).

5.2.2 Social News Websites (SNWs)

Online news diffusions are gaining increasing popularity and theoretical significance with the rise of network society (Castells, 2007; Shumate, 2010). SNWs are democratic systems that foster online citizen journalism. Online citizen journalism refers to a range of web-based practices in which users collaboratively engage in journalistic practices, including blogging, photo and video sharing, and posting eyewitness commentary about current events. However, the definition of citizen journalism is still open for debate: citizens' participation without necessarily acting as content creators (e.g., retweeting, linking, tagging, rating, modifying or commenting upon news materials posted by other users or professional news outlets) is also included in definitions of citizen journalism (Domingo et al., 2008; Goode, 2009).

Departing from classic studies of news diffusion, research on online news diffusion has advanced the landscape of research on news diffusions in many ways. Digital traces of human online behavior grant us an opportunity to look into the emerging mechanisms of online news diffusion and address long-standing puzzles of traditional research. In the past, scholars had to prepare in advance to study whatever news appeared suddenly, which limited the research design (DeFleur, 1987; Miller, 1945). Further, self-report data got from survey or interview suffers from a memory problem because of the time lag (DeFleur & Cronin, 1991; Funkhouser & McCombs, 1971). Finally, most academic inquiries into news diffusion rely on case studies, where both cases and sample size are limited (Funkhouser & McCombs, 1971). However, on the Internet news diffusion occurs at any time, and thus scholars can track both the diffusion process and the underlying social networks, which carry thousands of news items through millions of diffusers, through digital traces, in an unobtrusive way (Lazer et al., 2009).

Viewing news diffusion as contagion (Bakshy, et al., 2009), in according to an epidemics model (Becker, 1992), most studies of online news diffusion merely extensively stress the interpersonal effects played by social networks (Lerman & Ghosh, 2010; Shao, et al., 2009; S. Wu, et al., 2011). Although classic news diffusion studies are not sufficient to interpret the mechanism of online news diffusion, recent research overlooks other factors, especially the role played by collective gatekeepers, which may account for the crucial problem of recent studies (E.g., different from the conjecture of epidemic models, the size of online news diffusion is limited).

The puzzlement of online news diffusion lies in the limited diffusion size, compared with the prediction of epidemic models (Pastor-Satorras & Vespignani, 2001). Lerman et al. (2010) studied news diffusion on Digg (Digg is a leading social news website, see the method part for

more information), and found that most news ends up "infecting" only 0.1% of the entire network. This conclusion holds for other SNWs. Bakshy et al. (2011) studied the interpersonal influence of Twitter users by tracking 74 million diffusions, which showed that a large "cascade" of diffusion is relatively rare. This observation also holds for other social media. For example, Leskovec et al. (2006) found that the diffusions on a product recommendation network also tended to be shallow. By bringing collective gatekeeping into consideration, this study attempts to shed lights on this puzzlement.

5.3 Collective Gatekeeping vs. Interpersonal Effects

I study news diffusion on SNWs from the perspective of news sharing behavior. News sharing behavior on SNWs tends to stimulate more news viewing. First, individuals' news learning is influenced by the other people's news sharing behaviors (or news transmissions). Second, those who do not spread the news have no impact on their online friends (i.e., those who follow them on SNWs). Thus news sharing is the central part of online news diffusion studies, especially for the understanding of collective gatekeeping and interpersonal effects.

In terms of news sharing behavior, online news diffusion is boosted both by interpersonal effects and collective gatekeeping. First, interpersonal effects are based on social contact established through social networks. Two-step flow model posits that news flow from mass media to opinion leaders, and from them to a wider population (Lazarsfeld, et al., 1944). Wu et al. (2011) examined the two-step flow model for news diffusion on Twitter to study who says what to whom, and finds 50% of URLs consumed are generated by just 20K elite users, where the media produces the most information, but celebrities are the most followed. Those informed by interpersonal communications are more likely to influence others (Greenberg, 1964a; Larsen & Hill, 1954), and interpersonal conversations deepen the depth of news diffusion (Budd, et al.,

1966). Second, I emphasize the collective gatekeeping of SNWs. Gatekeeping is traditionally the process of selecting, writing, editing, positioning, scheduling, repeating and otherwise massaging information to become news, which is usually conducted by journalists and editors (Shoemaker, Vos, & Reese, 2008). However, online news is changing the mechanisms of gatekeeping, since readers can communicate with other readers, thereby extending the usual gatekeeping process, so that the selections may not reflect the news values of the journalists and editors (Shoemaker, Johnson, Seo, & Wang, 2010).

Insert Table 5 and Figure 24 here

Collective gatekeeping and interpersonal effects are different in both process and characteristics (see Table 5). Interpersonal effects are based on the ego networks of the diffusers. SNWs users selectively follow each other based on their preferences, and the information flows to individuals through social ties. Learning news from their ego networks, individuals' decision-making about diffusing news or not from bottom up are shaped both by personal preference and local wisdoms (see phases I, II, V, and VI in Figure 24). Collective gatekeeping, on the other hand, is based on both collaborative filtering and the news aggregation function of SNWs (see phases III and IV in Figure 24). The most popular news will be selected as the trending news or popular news, and further be aggregated on the webpage of news aggregation (see phase III in Figure 24), which is a bottom up process. Hereafter, the news will be exposed to larger and more diverse audiences, which is a top down process, and will greatly magnify the size of news diffusion (see phase IV in Figure 24). it is also necessary to note that the collective gatekeepers

of the news are invisible in this phase, i.e., the majority learn news directly from the webpage of news aggregation, but have no idea about who have participated in filtering and transmitting the news. Thus collective gatekeeping is based on both collective intelligence and empowered authority (Weiss, 2005; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Besides, compared with the long-term memory of interpersonal effects, the effects of collective gatekeeping is temporary (Sornette & Helmstetter, 2003).

I claim that collective gatekeeping is more important than interpersonal effects for online news diffusion, following the principle of least effort (Zipf, 1949). As part of the continuum of the pain-pleasure principle hypothesized by Jeremy Bentham in the perspective of economic man, George Zipf (1949) asserts that human behaviors follow the principle of least effort, which has been extensively confirmed by various studies of language usage. Mandelbrot (1953) further developed Zipf's idea with the explanation of minimum cost. Within the context of information seeking, predicted by the principle of least effort, information seekers will minimize the effort of information seeking, especially in choosing information channels (Allen, 1977; Rosenberg, 1967). Compared with interpersonal sources, collective gatekeeping needs less effort. First, collective gatekeeping is more efficient because it has the advantage of collective intelligence. According to Woolley et al.'s study (2010), the collective intelligence factor explains a group's performance in a wide variety of tasks. For Digg users, obtaining news from the front page is more efficient than reading news from their online friends. Second, compared with interpersonal effects, collective gatekeeping is more likely to minimize production cost (Lorenz, Rauhut, Schweitzer, & Helbing, 2011) and reduce audiences' costs of allocating limited attentions. Third, since collective gatekeeping is based on collective collaboration, it has higher credibility than information forwarded by online friends.

Based on the insights gained through the above discussion, I draw the following hypotheses, primarily focusing on how widely news diffuses on SNWs, interpersonal effects, collective gatekeeping, and the categories of news. Compared with interpersonal sources, I emphasize the primary role of collective gatekeeping in news diffusion, thus I propose H1. Additionally, consistent with the arguments above, I propose H1a and H1b for collective gatekeeping, and H1c and H1d for interpersonal effects.

H1: Collective gatekeeping has greater influence on how widely news diffuses on SNWs than interpersonal effects.

H1a: Collective gatekeepers have positive influence on the size of news diffusion.

H1b: The earlier news is promoted to SNWs' webpages for news aggregation, the more widely it is diffused.

H1c: Interpersonal sources have positive influence on the size of news diffusion.

H1d: News submitters' popularity has positive influence on how widely news diffuses.

Classic research on news diffusion focuses on the reach rate of news, which is measured by the number of knowers of news (i.e., the extent of exposures to news). This study primarily focuses on the diffusers' news sharing behavior at the individual level and the accumulated number of shares (e.g., transmissions) on the news level. To control the influence of exposures, I propose H2.

H2: The more news is exposed to audiences, the more widely news diffuses.

News diffusion is content-specific. It is necessary to control the influence of news content, and one way to do this is to control the category of news. Nowadays, audiences increasingly prefer entertainment news and become less likely to learn about political news (Prior, 2005).

Thus entertainment news may dominate the news diffusion on SNWs, which will hurt SNWs' function for online citizen journalism, and leads to H3.

H3: Entertainment news is more widely disseminated than the news of other categories.

5.4 Method

5.4.1 Data

I employ the Digg 2009 data collected by Kristina Lerman (2010) to test and verify the hypotheses. Digg (www.digg.com) is a social news website designed for news sharing. With over 3 million registered users at the time data was collected (i.e., June 2009), Digg allowed users to submit and rate news stories. Digg users use their friend interfaces to read news diffused by their online friends, and if a news story has accumulated enough votes (e.g., 100 votes, about 800 views), it is promoted to the webpage for news aggregation, from which it can directly reach other users.

Insert Figure 25 here

The data consist of two parts: voting records and social graph. The voting records were made by 139,409 distinct users for 3553 news stories in June 2009. The social graph contains 1,731,658 friendships of 71,367 users. I fit the distribution of node degree and test the goodness of fit using the method proposed by Clauset et al. (2009), which indicates that the degree distribution is scale-free (*Alpha* = 1.75, Kolmogorov-Smirnov statistic $D = 0.02$), while the clustering coefficient is relatively small ($C = 0.06$).

5.4.2 Measure

Size of Diffusion. The dependent variable is the size of news diffusion (*M* = 849, *SD* = 987). It is measured by the number of diffusers/votes for the news. The least popular news only has 122 votes, while the most popular news story, about the death of Michael Jackson, accumulated more than 24,000 votes. The distribution of the size of news diffusion is log-normal, and by taking the logarithm of the number of votes, the distribution can be easily normalized (*M* = 6.42, *SD* = 0.72).

Network Threshold. To quantify the interpersonal effects and collective gatekeeping, I employ the measurement of "network threshold". Further, to demonstrate the evolving friend network and the dynamic diffusion behavior of the individuals, I construct the original data into a directed network format with two time dimensions: link time, diffusion time. E.g., A → B, B is the person followed by A (i.e., A is a fan of B). The link time is the time at which node A follows node B; diffusion time of node A is the time at which node A votes for the news story, and similarly, diffusion time of node B is the time node B votes for the news story.

For a specific news story, there are n individuals participating in diffusion. Let i represent the i^{th} individual, and i votes for a piece of news at time t_i. Then the Network threshold of i is:

$$\text{Threshold}_i = \frac{n_{it_i}}{N_{it_i}} \qquad (1)$$

$N_{\text{i t}_i}$ is the total number of people who are followed by i at time t, and $n_{\text{i t}_i}$ is the number of people N_{it_i} who also have diffused the news by time t_i. E.g., if, by the time person i decides to vote for a specific news item, he/she has followed 100 people (i.e. $N_{it_i} = 100$), and among those 100 people, only 20 people have already shared the news (i.e., $n_{it_i} = 20$), the network threshold of person i is 0.2.

Interpersonal Effects. On the individual level, interpersonal effects are measured by the network threshold. Average threshold is the mean value of all individuals' thresholds for specific news; thus, it measures the extent to which news is diffused by interpersonal networks. Aggregating network threshold from individual level to the news level, I can get average threshold (M = 0.03, SD= 0.02). The number of followers of submitters (M = 3324, SD = 3163) is also included to control the personal influence of submitters.

Collective Gatekeeping. Collective gatekeeping is the process of individuals collectively voting the information and promote the information to aggregators. The promotion time is the time elapsed from the information appeared to the time the information is promoted to the aggregator. For a specific news story, the zero-threshold ratio is the ratio of the number of news disseminators whose network threshold is zero to the total number of news disseminators. Theoretically, on the news level, average threshold measures the extent to what news diffusion is influenced by interpersonal effects, while zero-threshold ratio of news measures to what extent news is not diffused by interpersonal network. In our case, I adopt zero-threshold ratio (M = 0.68, SD = 0.13) to measure the influence of collective gatekeeping. Additionally, promotion time (*M = 12.56 hours, SD = 8.64 hours*)—the time elapsed from the news has been submitted to SNWs by submitters to the time it is promoted to the aggregator—is also adopted to investigate the impact of collective gatekeepers.

5.5 Results

OLS regression is used to test the hypotheses. The dependent variable is how widely news diffuses, while the key independent variables are interpersonal effects and collective gatekeeping. Essentially, I need to compare the impact of zero-threshold ratio and average threshold to test H1.

H1 asserts that, compared with interpersonal effects, collective gatekeeping has more influence on how widely news diffuses on SNWs. Table 6 indicates that the influence of the zero-threshold ratio (Beta = 0.86, sig < 0.001) is two times greater than that of average threshold (Beta = 0.41, sig < 0.001). Thus H1, H1a, H1c are confirmed.

_____

Insert Table 6 here

_____

H1b hypothesizes that the earlier news is promoted by means of an SNW webpage for news aggregation, the more widely news diffuses. The model demonstrates that the coefficient of promotion time is negative and significant (Beta = -0.03, sig < 0.001), which confirms H1b.

H1d asserts that news submitters' popularity has influence on how widely news diffuses on SNWs. In the regression model (see Table 6), submitters' popularity is measured by the number of their fans. The model reveals that the influence of submitters' popularity is significant (Beta = 0.02, sig < 0.01), which supports H1d well.

H2 states that the more the news is exposed to SNWs users, the more widely news diffuses on SNWs. The extent of news exposure is measured by the number of views of a particular news item. The model indicates that the more a news story is exposed to SNWs users, the farther the news story diffuses (Beta = 0.47, sig < 0.001). Thus I confirm H2.

To test H3 which posits that entertainment news is more widely spread than other news, news categories are coded using dummy coding, and entertainment news is treated as the baseline group. The model demonstrates that, except for world & business news (Beta = 0.07, sig <0.001) and technology news (Beta = 0.03, sig < 0.01), entertainment news diffuses farther than

other kinds of news (e.g., sports news (Beta = -0.02, sig < 0.01) and lifestyle news (Beta = -0.02, sig < 0.01). However, the differences between entertainment news and gaming news, offbeat news, and science news are not significant. Thus I only partly confirm H3.

5.6 Conclusion and Discussion

By bringing collective gatekeepers into the theoretical framework of news diffusion on SNWs, and moving from awareness of news to news sharing, this study has aimed to identify and compare the influence of interpersonal effects versus collective gatekeeping. The mean value of the zero-threshold ratio (0.68) implies that almost 70% of news diffusions are not influenced by interpersonal sources, and that news "jumps" in the social network rather than "infects" individuals one by one. The findings above have important implications for audience fragmentation, diffusion mechanisms, and the limited size of online news diffusion.

First, an understanding of collective gatekeepers as the primary driver for news diffusion on SNWs contributes to our knowledge of audience fragmentation. New media have eroded the centralized nature of mass communication, which fosters audience fragmentation. The fragmentation hypothesis rests on the assumption of diversified information channels and different content (Shaw & Hamm, 1997). The rise of social media intensifies this tendency, since individuals reinforce their preferences in news by following specific kinds of people and content (Prior, 2005; Tewksbury, 2005). However, on SNWs, only news above a threshold is promoted to the webpage for news aggregation; unpopular news soon dies out. For collective gatekeeping, the early diffusers, whose active news sharing activities promote news stories to the webpage for news aggregation, may act as "invisible" opinion leaders, since their preferences may dominate the visibility of news and shape the public's news learning. Thus news learning based on

collective gatekeeping may enhance the probability of audience homogeneity, which serves as a competing argument against the fragmentation hypothesis and asks for future research.

Second, against the backdrop of epidemic models, which overestimate online friends' influence, the strong collective gatekeeping effect deepens our understanding of the diffusion mechanism. Increasingly, epidemic models of information diffusion often starts with the assumption that a few nodes trigger long reactions through underlying social networks and result in large-scale information cascades (Lerman & Ghosh, 2010; Steeg, et al., 2011; S. Wu, et al., 2011). While reasonable under some conditions, the chain reaction assumption of epidemic models may not hold for SNWs. For example, on Digg, 70 % of the news sharing is not influenced by interpersonal effects, and the news diffusion is scattered and discontinuous. Similarly, although the diffusion chains in Facebook are very long, they are not primarily caused by a single chain-reaction event (Sun, et al., 2009).

This study modestly reminds us not to overestimate interpersonal effects in news diffusion for SNWs. However, it is necessary to note that the findings don't rule out the influence of interpersonal effects which, although not overwhelming, is strong and significant (Beta = 0.41, sig<0.001). To explain the limited size of news diffusion, prior literature speculates that although more exposures leads to larger news diffusion, repeated exposure to the same news doesn't make it more likely to be diffused (Steeg, et al., 2011). This proposition is rejected by our findings. Our results are consistent with Romero et al.'s study (2011), in which it is found that repeated exposure to a hashtag (i.e., the # symbol, called a hashtag, is used to mark keywords or topics in a Tweet) on Twitter has significant marginal effects.

Third, collective gatekeeping is featured by temporality due to the decay of novelty (F. Wu & Huberman, 2007) , which sheds light in understanding the limited size of information

diffusion. In terms of the zero-sum feature of public attention, the rise of an issue is at the expense of other issues (McCombs & Zhu, 1995b; Zhu, 1992). Different news stories compete with each other to stay on the webpage of news aggregation. As the result of the fierce competition, recent news on the webpage of news aggregation will be replaced by some other news. Thus the fast update of news on the webpage of news aggregation also restricts the size of news diffusion. According to Wu and Huberman's study (2007), the half-lifetime of Digg news is 69 minutes, which is consistent with the fact that a news story usually lives on the webpage for news aggregation for a period between 1 and 2 hours. In our case, although most news diffuses for an average of 14 days (M= 14.20, SD= 8.04), more than 70% of the information sharing happens within one day (M= 0.71, SD=0.12) (see Figure 26). Thus there are strong bursts of news sharing, which, according to recent research, will slow down the diffusion and limit the size of news diffusion (Karsai, et al., 2011). It is necessary to note that the relatively weak interpersonal effects also hamper the local community in accumulating enough diffusers to overcome individuals' network thresholds, which also hinders online news diffusion.

Insert Figure 26 here

The overall evidences indicates that most SNWs users are primarily influenced to share news by collective gatekeepers rather than interpersonal sources, which is support for the idea that the principle of least effort underlying news seeking behaviors shapes the diffusion channels, both online and offline (Zipf, 1949). Prior studies find a competing relationship between interpersonal effects and mass media (Bordogna & Albano, 2007; DeFleur, 1987). Although

there are many cases in which interpersonal networks override the influence of mass media, classic studies of news diffusion in general support a strong mass media effect (DeFleur, 1987). However, on SNWs, both collective gatekeeping and interpersonal effects eliminate the direct impact of mainstream news sources and allow the public to decide what is news. Although mass media and collective gatekeeping are different, they are similar in including a top-down process, which is more efficient, low-cost, and credible than interpersonal effects.

This study is not without limitations. First, it is necessary to improve the measurement and investigate more complicated diffusion mechanisms in future research. Second, according to the literature, interpersonal effects and homophily are generically confounded (Shalizi & Thomas, 2011; VanderWeele, 2011), and thus in this study the influence of interpersonal effects is still overestimated, which asks for further attempts to distinguish them from each other.

Chapter 6. The Origin of Bursts in Public Attention: The Temporality Hypothesis for the Diffusion of YouTube Videos

6.1 Introduction

Public attention is the extent to which individuals collectively allocate cognitive resources to cultural products and social events across space and time. Recently, accompanied by the rise of digital media, public attention has evolved to be one of the central concerns of Internet study, and attracts many interdisciplinary scholars to uncover the hidden patterns and underlying mechanisms. In addition to fragmentation and polarization (Webster, 2005, 2011; Webster & Ksiazek, 2012), recent research shows public attention is also characterized by the burst phenomena (Crane & Sornette, 2008). As it has been illustrated in the Model of ABXCT, temporal patterns of information diffusion influenced the diffusion process. One of the most important aspects of temporal changes comes from the burst of diffusion curves.

The burst of public attention describes the situation of a sudden and violent growth and decline of public attention, which can offer deep insights into the rhythm of how we collectively do everything. For example, Figure 27 demonstrates the Google search queries for "Harry Potter" and "hurricane" in 2011. Driven by the unsuspected and sudden natural disaster, there is a much stronger burst in the public attention for the hurricane compared to Harry Potter; Driven by the more enduring audience interest in Harry Potter, the burst of the public attention for Harry Potter decays one month slowly than that for the hurricane. To analyze such daily temporal patterns of public attention opens a new door to the hidden pattern of public attention.

Insert Figure 27 here

The origin of burst in *individual* behavior has been extensively investigated by prior studies (Oliveira & Barabási, 2005; Vázquez, et al., 2006). The interevent times of human communication follows a power law distribution rather than a Poisson distribution. Barabasi (2005) confirms that emailing behavior is characterized by bursts of rapidly occurring events separated by long periods of inactivity. Vázquez et al. (2006) find empirical evidence of the non-Poisson statistics of the timing in letter based communications, web browsing, library visits and stock trading. To explain the origin of the burst phenomena for individuals, a queuing model of individual decision making has been proposed by Barabasi (2005). Barabasi (2005) asserts that when individuals execute tasks based on personal priority, the timing of the tasks will be heavy tailed.

For burst in *collective* behaviors, based on the non-Poisson timing of individual behaviors, prior research suggests both exogenous shocks and endogenous perturbations as the sources (Sornette & Helmstetter, 2003). Exogenous shock is the impact from the source outside of social networks (e.g., occurrence of news event, and recommendation of the website); endogenous origin is the self-organized interactions among audiences, which is also conceptualized as social influence or interpersonal effects. Based on temporal patterns of bursts, prior research focus on distinguishing exogenous shocks and endogenous perturbations (Sornette, Deschâres, Gilbert, & Ageon, 2004; Sornette & Helmstetter, 2003). Following this line of thought, Crane and Sornette (2008) propose a model of epidemic spreading on a social network to model how the public allocate their attentions to YouTube videos.

However, this line of research is not without limitations. First, for YouTube videos, although hundreds of thousands of videos in which a burst is followed by a ubiquitous power-

law relaxation governing the timing of views, most activities (90%) can be described accurately as a Poisson process (Crane & Sornette, 2008). Second, prior studies mainly focus on the exogenous shocks and internal perturbations (e.g., social influence) which are the factors on the system level and network level, whereas the other factors on the individual level and content level are overlooked. Third, how burst of public attention is related to the popularity of media products, diffusion channels, audiences' interest, and the category of YouTube videos still remains a question. Last and the most important, the general principle governing these underlying factors is not clear.

To gauge the underlying principle of these influential factors for bursts of public attention, this study draws on the literature of public attention in communication research, especially the principle of the limited-capacity for public attention. We claim that temporality is the determinant factor underlying the different impacts that govern the rise and fall of public attention. Here the term of temporality is defined as the duration of the driving forces for public attention, as well as its manifestation on daily diffusion curves—burst (i.e., the sudden rise and decline of public attention). Thus, the concept of temporality touches the origin of self-sustainability for media products. E.g., YouTube videos with large bursts usually imply driving forces have a short life-span. In the later section, first, we review related communication studies on public attention. Second, we derive the temporality hypothesis from the influences of information diffusion, based on which, we formulate specific hypotheses. Third, we test our hypotheses with the diffusion data of YouTube videos. Fourth, we discuss the theoretical generalizations and implications.

6.2 Public Attention: Antecedents, Patterns and Principles

Since the very beginning, studies on public attention have been the traditions of communication research. First, in terms of the stages of public attention, the research on public attention in communication can be separated into two branches: one focuses on the allocation of public attention, or the *antecedent* of public attention. This line of research draws on the driving forces of public attention, e.g., information flow (Greenberg, 1964b; Larsen & Hill, 1954; Lazarsfeld, et al., 1944), agenda-setting (McCombs, 1981; McCombs & Shaw, 1972), and audience flow (Headen, Klompmaker, & Rust, 1979; Webster, 1985); the other branch of intellectual investigations into individual attention focuses on the *effect* of attention allocation, for instance, how selective attention generates short-term and long-run memory, e.g., the limited capacity model of mediated messages (Lang, 2006). Since this study is primarily about the antecedents of attention allocation, rather than the effect of attention allocation, this study primarily focuses on the first branch of study, especially the driving forces, the explicit patterns, and the underlying principles of public attention.

First, the way in which different antecedents shape the allocation of public attention has been extensively discussed by diffusion scholars, since the seminal study of two-step flow (Lazarsfeld, et al., 1944). Against the backdrop of magic bullet theory of media effect, the theory of two-step flow suggests that social network has the potential to influence the allocation of public attention (Katz, 1957; Lazarsfeld, et al., 1944). In the following investigations, the line of research on news diffusion has systematically documented how individuals learn news driven by both interpersonal networks and mass media (Greenberg, 1964b; Larsen & Hill, 1954). Thus, both exogenous impacts (e.g., media effect, inheritance effects of adjacent media products) and endogenous influences (e.g., social networks) determine the allocation of public attention.

Second, shifting the perspective from events to issues, researchers of agenda-settings measured the relative attention given to public issues and the media coverage, which provided positive evidences for media agenda-setting, i.e., media coverage determines issue salience, and thus, media attention shapes public attention (McCombs, 1981; McCombs & Shaw, 1972). Therefore, agenda-setting research again demonstrates the power of exogenous impacts on the allocation of public attention.

Third, scholars of TV ratings have extensively studied the audience flow in the television viewing situation. The pattern of " inheritance effects" confirms the tendency of people who watch one program to stay tuned to the next (Headen, et al., 1979; Webster, 1985). Further, modeling the growth of attention to public issues, Neuman (1990) demonstrates that the evolution of public attention on a set of political issues varies dramatically for different types of issues. Similarly, the J-curve model of news diffusion also indicates the attributes of information itself have an important influence on the allocation of public attention (Greenberg, 1964b). Therefore, both the content of media product and the mutual relationship between media products (e.g., inheritance effects of adjacent or similar media products) as the antecedents shape the allocation of public attention.

Fourth, fragmentation and polarization are two most obvious patterns for public attention (Webster, 2005, 2011; Webster & Ksiazek, 2012). Fragmentation is concerned with how total attendance is distributed across media or media products. Polarization describes a tendency for subsets of the public to concentrate their attention on a homogeneous assortment of media products. Measuring the allocation of public attention to a wide range of media products with digital traces, fragmentation and polarization have been extensively confirmed by recent research. The highly skewed distribution of public attention is featured by a long tail or a heavy tail

(Webster & Ksiazek, 2012; F. Wu & Huberman, 2007). Therefore, although most media products receive a negligible amount of attention, a few items pass the attention threshold and become extremely popular and persist as public trends for a long time (Huberman, 2008; Klamer & Van Dalen, 2002). If fragmentation and long-tail distribution indicate the allocation of public attention may also be a process of preferential attachment (Barabási & Albert, 1999), polarization demonstrates that each audience settles into its preferred niche (Webster, 2011). Thus, both fragmentation and polarization suggest people tend to selectively allocate their attention to some media products. Therefore, the individual attribute, especially the audience interest, is an important antecedent of attention allocation.

Drawing on Giddens' theory of structuration, Webster (2011) supplies a general framework in order to understand the forces that shape public attention. In light of the duality of media, Webster (2011) asserts that the interplay between agents and structure determines the patterns of public attention. Structure and agency are mutually constituted in a continuous process of reciprocal causation. In particular, he states that both market information regimes (e.g., TV rating companies) and user information regimes (i.e., search and recommendation systems, e.g., Google, Twitter) have impacts on the duality of media. This makes "a triumph of convergence" between the desires of agents and social structures possible. If this is the case, "giving people what they want" tends to govern the public attention in digital media systems, and understanding the audience interest would become the central concern.

Underlying various influences of the flow of public attention, the most important principle on public attention is the limited capacity for information processing (Kahneman, 1973; Norman & Bobrow, 1975). The principle of limited capacity manifests itself to various stages of information processing: attention, perception, memory, and recall. In this study, I primarily focus

on the attention stage of information processing. The capacity models of attention rests on two assumptions: First, attention and mental resources are strongly correlated. Second, mental resources are limited. Thus, the limitation of attention occurs when resources are exhausted (Kahneman, 1973).

In later studies, the limited-capacity principle has been widely used in various areas of media use, e.g., mediated message processing (Lang, 2006), television viewing (Lang, 1995; Lang, Geiger, Strickwerda, & Sumner, 1993; Lang, Newhagen, & Reeves, 1996)., and the limited capacity principle of attention. Based on but not limited to the theoretical perspective of limited capacity, scholars of agenda-setting argue that the allocation of public attention is governed by the zero-sum principle (McCombs & Zhu, 1995b; Zhu, 1992). Issues compete for the limited public attention, however, the total amount of public attention as the computing capacity to process information is a constant (McCombs & Zhu, 1995b), thus one issue rises in the public attention is at the expense of the fall of the other issues. In addition to the limited capacity argument which is a psychological constraint, they argued that there are also physical limits on the time and access (Zhu, 1992). For example, time is a constant resource for everyone. In most cases, individuals can't view multiple videos at the same time slot. In the age of web 2.0, an enormous amount of information competes for the limited public attention. Therefore, the principle of limited capacity for public attention is becoming more visible in human communication.

6.3 Temporality Hypothesis of Influences

Repositioning research questions into the big picture of attention research helps us understand the origin of burst. Burst is about the temporal pattern of public attention towards specific media products (e.g., these media products are YouTube videos in our empirical

inquiries, either user-generated or not). Based on the driving forces of public attention and the underlying principles, one speculation is the limited duration of driving forces lead to the temporal patterns of public attention, which generates burst in public attention. Thus, I propose the temporality hypothesis: the temporality of different impacts is the determinant factor governing the rise and fall of public attention towards specific media products. Here the term of temporality is defined as to what extent the specific media products (e.g., YouTube videos) could sustainably attract the public attention, and it concerns the self-sustainability for the diffusion of media products. Thus, YouTube videos of large bursts imply a short duration of driving forces and usually a short life time.

According to the general framework of ABXCT model, the temporal patterns, especially bursts, are bursts, are associated with the overall diffusion process. Therefore, it is natural to reason the relationship between bursts and diffusion size, communication channels, audience, and information information features. According to the levels of analysis and the temporality of various influences, I influences, I analyze the burst of public attention in four levels (for a brief summary, see Table 6 OLS regression of interpersonal effect, collective gatekeeping, and news category on the size of news diffusion

	Size of News Diffusion (Log)		
	Beta	SE	
Number of views	0.47	0.00	***
Collective gatekeeping			
Zero-threshold ratio	0.86	0.08	***
Promotion time	-0.03	0.00	***
Interpersonal effect			
Average threshold	0.41	0.54	***

Submitter's fans	0.02	0.00	**
News Category			
Gaming News	0.00	0.03	Ns
Lifestyle news	-0.02	0.02	**
Offbeat news	-0.01	0.02	Ns
Science news	-0.01	0.02	Ns
Sports news	-0.02	0.02	**
Technology news	0.03	0.02	**
World & Business news	0.07	0.02	***
(Intercept)	-----	0.07	***
R Square (%)	84.8***		

Note: Sample size N = 3553. The baseline group of news category is entertainment. The dependent variable is the logarithm form of the size of news diffusion (i.e. how widely the news diffuses). * P < 0.05. ** P < 0.01. *** P< 0.001.

Table 7): (1) the exogenous shocks, e.g., news events and recommendations (Cha, Kwak, Rodriguez, Ahn, & Moon, 2007; Crane & Sornette, 2008); (2) social influence through social networks (Crane & Sornette, 2008; Sornette & Helmstetter, 2003); (3) the interest of audiences (Funkhouser & McCombs, 1971); and (4) the novelty of the media content (F. Wu & Huberman, 2007). In this study, I identify the burst in public attention for YouTube videos, and form the temporality hypothesis based on the literature. To be specific:

Insert Table 6 OLS regression of interpersonal effect, collective gatekeeping, and news category on the size of news diffusion

	Size of News Diffusion (Log)		
	Beta	SE	
Number of views	0.47	0.00	***
Collective gatekeeping			
Zero-threshold ratio	0.86	0.08	***
Promotion time	-0.03	0.00	***
Interpersonal effect			
Average threshold	0.41	0.54	***
Submitter's fans	0.02	0.00	**
News Category			
Gaming News	0.00	0.03	Ns
Lifestyle news	-0.02	0.02	**
Offbeat news	-0.01	0.02	Ns
Science news	-0.01	0.02	Ns
Sports news	-0.02	0.02	**
Technology news	0.03	0.02	**
World & Business news	0.07	0.02	***

(Intercept)	-----	0.07	***
R Square (%)	84.8***		

Note: Sample size N = 3553. The baseline group of news category is entertainment. The dependent variable is the logarithm form of the size of news diffusion (i.e. how widely the news diffuses). * P < 0.05. ** P < 0.01. *** P< 0.001.

Table 7 here

First, there is a time range of influences for either exogenous shocks or endogenous perturbations (see factors on level 1 and level 2 in Table 7). External shock is a strong *instantaneous* perturbation that is sufficient in itself to change the trajectory of the time evolution of the complex system, while endogenous shock is the result of the *cumulative* effect of many small perturbations which can add up coherently due to the long-range memory of the dynamical process (Sornette & Helmstetter, 2003). Dynamic shift of public attentions is shaped by both exogenous shocks and endogenous perturbations (Figueiredo, Benevenuto, & Almeida, 2011; Pollmann, 1998; Roehner, Sornette, & Andersen, 2004).

Based on this line of thought, the response function of exogenous and endogenous shocks can be captured by an epidemic branching process. After a large fluctuation, the temporal evolution of the system as a long memory process makes specific signatures in the precursory and the relaxation of a profound shock (Sornette & Helmstetter, 2003). Thus, the influence of exogenous and endogenous shocks on the popularity of YouTube videos could be distinguished (Cha, et al., 2007; Crane & Sornette, 2008; Figueiredo, et al., 2011). However, only exogenous shocks and endogenous perturbations (in their model, the social influence based on social networks) are not enough, since it at least overlooks the factors on the audience level and the content level.

Second, structured by the social structure, audience interest is relatively stable over time (see the factors of level 3 in Table 7). E.g., Kevin Lewis et al. (2012) find little evidence for the diffusion of tastes among Facebook friends—except for tastes in classical/jazz music. Defined as a general term, audience interest here includes: preferences, motivations, innovativeness, and curiosity. Research on attention to political news also confirms that “there is a reasonably stable

stratum of the population that pays attention to public affairs" (Price, 1992, p. 39). Audience interest has definite influence on information diffusion and attention flow (Funkhouser & McCombs, 1971). In terms of selective exposure, individuals generally tend to expose themselves to ideas that are in accordance with their interests, needs, or existing attitudes; consciously or unconsciously, individuals avoid information that is in conflict with their existing predispositions (Frey, 1986; Rogers, 1983). Thus, those who have interest in the online videos tend to allocate their attentions to the videos even after the bursts caused by exogenous shocks and social influences.

Third, according to Wu and Huberman (2007), the novelty of media content diminishes over time (see the factors of level 4 in Table 7). The decay of novelty is primarily caused by the competition of different media products for public attentions. Therefore, the popularity of different YouTube videos reflects how audiences collectively allocate their attentions. However, the capacity of audience attentions is remarkably limited. The rise of an issue on public agenda is at the expense of other issues, in terms of the zero-sum feature of public agenda (McCombs & Zhu, 1995b; Zhu, 1992). The decay of novelty in the long run has been studies in terms of forgetting curves. E.g., Pollmann (1998) finds that there is a diminishing attention to the phenomena of the past, and he explains it with the cognitive distance between past and present. For YouTube videos of different categories, the influence of temporality of novelty will significantly shape the diffusion process, which will be further analyzed in the following part of hypotheses formation.

Further, the temporality of public attention towards specific media products is determined by the nature of content. For instance, news propagates through a social network until its novelty fades away within hours (F. Wu & Huberman, 2007), while in the case of knowledge networks

like Wikipedia, the popularity of a document can remain vital for months or even years (Ratkiewicz, Fortunato, Flammini, Menczer, & Vespignani, 2010). Similarly, Cha et al. (2009) demonstrate how propagation patterns depend on content category. For example, the latest political news video immediately catches the attention of bloggers and fades away after a week, while a music video propagates slowly over a long time.

Put together, all the influential factors on the four levels of analysis are featured by different properties of temporality. Although the temporality is caused by different mechanisms, it has important impacts on the response functions of public attention. Based on this line of thought, we argue that the property of temporality underlying these factors lead to the burst of public attention. In the later sections, I will elaborate more about the theoretical link between specific influential factors and their properties of temporality, and formulate the research hypotheses.

6.3.1 Burst of Public Attention and the Popularity of YouTube Videos

First, although public attention towards most videos experiences a burst and then dies out, for the popular videos, there is a smaller burst. According to the bandwagon effect of aggregated popularity, when a video displays a high view count, its popularity over that of the other videos snowballs further (Fu, 2012; Fu & Sim, 2011). Therefore the aggregated popularity makes public attention on popular videos perish more slowly than that of those less popular videos, which mitigates the degree of bursts. Thus, there is a negative relationship between popularity and burst. The best example should be the most popular YouTube video Charlie bit my finger–again! [6]. The total views of the video are over 450 million. The number has kept growing linearly since it was uploaded to YouTube in 2007 without any burst.

[6] http://youtu.be/_OBlgSz8sSM

Second, burst tends to limit the popularity of information diffusion (Karsai, et al., 2011). Based on the non-Poisson distribution of inter-event time, the revealed burst implies that the inter-event time will be much longer after the burst. Thus the information would propagate slowly after the burst, which limits the popularity of information diffusion. Based upon the arguments above, we expect that the burst will be negatively related with the popularity of videos, especially the total number of views and the length of lifetimes. Note that there exists a mutually reinforced negative relationship between burst and popularity. Therefore, here we merely aim to investigate the correlations between them, rather than to gauge the causal relationships between them.

H1: bursts are negatively associated with the popularity (e.g., views and lifetimes) of videos.

6.3.2 Burst of Public Attention and Diffusion Channels

Basically, two general mechanisms of information diffusion are considered: push and pull. Diffusion rate is a function of push and pull forces (Lyytinen & Damsgaard, 2001; Rogers, 1983). Push factors include features of technology and channels of communication, e.g., system recommendation; pull is determined by adopter's rational choices, e.g., audiences' searching behaviors. There are various diffusion channels for YouTube videos which can be categorized as: (1) system recommendations, (2) social influences, (3) searching engines, and (4) mobile communications. Both external recommendations and internal recommendations are exogenous origins of bursts.

Recency is an important criterion for recommender systems. Due to the keen competition of media products (Zhu, 1992) and the decay of novelty (F. Wu & Huberman, 2007), system recommenders frequently update the content over time. Thus, the strength of system

recommendations for a specific media product generally decreases over time. For example, YouTube displays an auto-generated list of popular videos, and more than 90% of the recommended videos are uploaded within a short period of days[7]. Based on the temporality of recommendations, I propose the respective hypothesis:

H2a: Recommendation has positive impacts on the bursts of public attention.

Social influence coming from the interaction of the agents in the social network is the internal factor influencing the burst of public attention. As it has been argued that the recovery after an endogenous shock is in general *slower* at early times. Consistent with this argument, prior studies of information flow states that the information flows driven by social networks is relatively slow compared with the information flow driven by system recommendations, e.g., Lerman et al. (2010) found that information diffuses faster on Digg relative to Twitter, since system recommendation plays a more important role on Digg. Thus, I expect that a slow relaxation process after endogenous influences tends to decrease the burst of public attention.

H2b: Social influence has negative impact on the burst of public attention.

Although use of search engines may be motivated by both exogenous shocks and endogenous factors, searching is primarily driven by audience interest. Since the long-lasting feature of audience interest, viewing by searching makes the allocation of public attention relatively stable along time, i.e., use of search engines tends to slow down the precursory before the burst and recovery after the burst. Thus, searching would reduce the odds of finding a burst. Note that popularity is one of the important criteria for search engines to rank results, so popular items show up in priority positions when searched (Fu & Sim, 2011). Therefore, the pull mechanism based on using search engines is not neutral. Thus, popular items viewed by search

[7] Check the list http://www.youtube.com/channel/HC4qRk91tndwg

engines are more stable than less popular items, which support the negative association between the burst and popularity (H1).

H2c: Searching has negative influence on the burst of public attention.

Different from traditional diffusion channels, the World Wide Web is becoming more accessible by portable and wireless devices. E.g., more and more traffic flowing to YouTube comes from smartphones. However, mobile communication is restrained by both devices and services, especially the limited storage of the mobile, the speed of mobile Internet, the size of screen, the difficulty of navigation and displaying multiple windows in the same screen (Finamore, Mellia, Munafò, Torres, & Rao, 2011). Therefore, it is not convenient to search YouTube videos and navigate freely on YouTube via mobile phones; On the contrary, users of mobile phone tend to surf from the front page of YouTube, which implies that recommendation plays a more important role for mobile Internet surfers. Thus, viewing YouTube video via mobile tends to be featured by the temporality.

H2d: Surfing YouTube by mobile tends to strengthen the burst of public attention.

6.3.3 Burst of Public Attention and Audience Interest

Audience interest, as one form of revealed preference, guarantees a long-term growth. Audience interest is revealed by the user-generated evaluations. The theory of revealed preference assumes that consumers' preferences can be revealed by their behavioral habits. Audiences' evaluations are given after viewing the videos. e.g., rating is the result of voting by clicking the 'like' button or the 'dislike' button. In terms of attention economy, to reveal personal interest online requires individuals to invest their scarce time and energy. If audiences' revealed preference indicates they like the video, the video would have a smaller temporality and

maintain to spread for a longer time, which gives rise to a smaller burst and leads to the following hypothesis:

H3: burst of public attention is negatively related to audience interest, e.g., ratings, comments, and favorites.

6.3.4 Burst of Public Attention and Category of YouTube Videos

There are various categories of YouTube Videos. As one important part of the video content, category has important impact on the burst of public attention during the diffusion process. Since media products are content-specific, it is necessary to partial out the influence of the content, and one way is to control the influence by the categories of the content.

The central concern of this study is the temporality of different influential factors. For the category of YouTube videos, the most important one is the videos of news, which are greatly influenced by the external news events. Hence, videos of news have a larger temporality relative to the other videos, and thus, it is reasonable to formulate the following hypothesis:

H4a: compared with the videos of the other categories, those videos of the news category will experience a stronger burst of public attention.

On the contrary, since the UGC nature of YouTube videos, many videos uploaded to the system breaking copyright or intruding upon privacy. Most these videos have larger attractions to audiences, and therefore a smaller temporality. Thus, it is sufficient to deduce the following hypothesis according to the temporality assumption:

H4b: compared with the videos of the other categories, the videos of copyright-protected have a weaker burst of public attention.

6.4 Method

6.4.1 Data and Sampling

We employ the YouTube data collected by Figueiredo et al. (2011) to test and verify our hypotheses. YouTube maintains several top lists (e.g., most viewed and most commented videos) for per country and world-wide, and each top list contains 100 videos. By crawling all world-wide top lists available on YouTube for different time scales (i.e., top of the day, week, month, and top of all time) in a single day of April 2010, the information about diffusion curves, audience evaluations, diffusion channels of 14978 videos were collected.

To study the influence of video categories, I collect the metadata of these YouTube videos in April, 2012. The interval time of data collections grants us an opportunity to unobstrusively check how many videos have been set as private or deleted due to copyright infringement (3592 videos of 14978 videos are set as private or deleted).

The growth curve of YouTube videos could be crawled since YouTube supplies the cumulative growth of the number of views for each video using the Google charts API. By providing the requested URL, YouTube requests the Google charts API and get one hundred pairs of (x,y) values. By collecting these requested URL, the (x,y) values have been collected. It is worth to note that the limitation of only 100 data points might introduce distortions in this analysis. It could be expected that for those videos lasting more than 100 days, the peak fraction calculated may overestimate the proportion of burst. Thus, it is necessary to interpolate the data points according to the growth curve.

6.4.2 Measure

Burst of public attention. The burst of collective viewing behavior, on the video level, can be measured by the peak fraction of daily growth curve. E.g., Riley Crane and Didier Sornette (2008) propose to distinguish YouTube videos by the proportion of burst which is measured by peak fraction (*F*). Peak Fraction is the fraction of views observed on the peak day compared with the total cumulative views. it is ranging from 0 to 1. According to the origin of burst (e.g., exogenous, endogenous) and the extent to which social network has been activated (e.g., critical, subcritical), they propose to practically classify YouTube videos into three categories: exogenous subcritical (80% < *F* < 100%), exogenous critical (20% < *F* < 80%), endogenous critical (0% < *F* < 20%). In this study, I also measured the burst of public attention with peak fraction (*M* = .330, *SD* = .199).

Popularity of YouTube videos. The popularity of videos can be measured in two dimensions: the time dimension and the size dimension. The time dimension concerns how long the videos can live until they attracts no attentions, and it is measured by the lifetime of videos (*M* = 149.9; *SD* = 300.1); and the size dimension concerns how many people have viewed the videos, and it is measured by the number of total views (*M* = 934300, *SD* = 4536706). it is necessary to note that, the popularity of YouTube videos are highly skewed which is also the reason of high values of standard deviations. Thus, I need to take the logarithmic form to normalize it. For example, I transform the skewed distribution of total views to a normal distribution by taking the logarithmic form (*M* = 10.3, *SD* = 2.7). Such transformations have been used in the regression model for the measurements of popularity, diffusion channels, and audience interest.

Diffusion channels. Compared with the other social systems, the diffusion channels of YouTube are very diverse. By collecting these requested URL using Google Chart API, a set of

referrers that is, links used by users to access the video, along with the number of views for which each referrer is responsible, are also retrieved.

Compared with the other online social systems (e.g., social news website), the diffusion channels of YouTube are very diverse, including system recommendations (i.e., internal recommendations, and external recommendations) (M = 186000, SD = 1040476), social influence (i.e., subscribing a Channel page, subscribing a user) (M = 24340, SD = 115942.4), and mobile communications (M = 17350, SD = 123103.5), and audiences' searching behaviors (i.e., searching with search engine, either on YouTube or elsewhere) (M = 69790, SD = 881392.8) (see Figure 28). Similar to the distribution of the popularity of YouTube videos, the distribution for the diffusion channels is also highly skewed, and I need to normalize it by taking the logarithmic form.

Insert Figure 28 here

Audience interest. Audience interest is measured by the number of total comments (including both positive and negative comments) (M = 3631, SD = 20931), the number of ratings (i.e., the number of likes plus the number of dislikes) (M = 3234, SD = 12017.9), and the number of favorites (i.e., a special playlist for your favorite videos appearing on the right side of the users' channel page) (M = 3239, SD = 16829.5). Just as popularities and diffusion channels, these measurements will also be normalized by taking the logarithmic form.

Categories of YouTube videos. According to the category of videos given by YouTube, the categories of YouTube videos are recoded. To be specific, nine categories are identified,

including: copyright-protected (*N* = 3592), hobbies (*N* = 4548, videos about animals, autos, games, people, travel, sports), broadcast (*N* = 2057, videos about comedy, film, movies, music, shows, and trailers), education (*N* = 766), entertainment (*N* = 720), how-to (*N* = 835); nonprofit (*N* =783), technology (*N* = 881), and news (*N* = 796).

6.5 Results

To investigate the features of bursts, clustering analysis has been employed. Based on two attributes of bursts—peak fraction (*M* = 33%, *SD* =20%) and the time of peak fraction (*M*= 29% of lifetime, *SD* = 26% of lifetime), YouTube videos can be classified into four categories (see Figure 29): (1). bursts of large peak fraction (*M* = .47, *SD* = .12) and early peak time (*M* = .20, *SD* = .12), (2). bursts of small peak fraction (*M* = .20, *SD* = .09) and early peak time (*M* = .12, *SD* = .09), (3). bursts of small peak fraction (*M* = .12, *SD* = .10) and late peak time (*M* = .68, *SD* = .18), (4). bursts of large peak fraction (*M* = .58, *SD* = .14) and late peak time (*M* = .69, *SD* = .13).

OLS regression is used to test the hypothesis (see Table 8). H1 concerns the association between burst and the popularity of YouTube videos. The results indicate that peak fraction is negatively related with the views (β = -.20, p < .001), and the lifetime of videos (β = -.60, p < .001), thus I confirm H1.

Insert Table 8 and Figure 29 here

H2 concerns the relationship between diffusion channels and the burst of public attention. The results reveal recommendation has positive impact on the burst of public attention (β = .14, p < .001); social influence has negative impact on the burst of public attention (β = -.02, p < .05); searching has negative impact on the burst of public attention (β = -.01, p < .001); and mobile has positive impact on the burst of public attention (β = .04, p < .001). Thus I confirm H2a-H2d.

Further, for the audience interest towards YouTube videos, peak fraction is negatively related to the number of favorites of YouTube videos (β = -.36, p < .001). However, it is positively related to the total number of ratings (β = .25, p < .001) and the total number comments (β = .02, p < .1). Thus, I fail in fully confirming H3.

H4 concerns the relationship between categories of videos and the burst of public attention. All the categories of the YouTube videos are coded using dummy coding, and the baseline group is the videos of news. The findings suggest the standard coefficients for all the categories of videos are negative, and the coefficient of the copyright- protected videos is the smallest (β = -.17, p < .001), which means: (1) compared with the videos of the other categories, those videos of the news category will experience a stronger burst of public attention; (2) compared with the videos of the other categories, the copyright protected videos have a weaker burst of public attention. Thus, I confirm H4a and H4b.

6.6 Conclusion and Discussion

In this present research, I capture the burst of public attention to YouTube videos with peak fraction, and look into how burst of public attention is related to the popularity of videos, audience interest, diffusion channels, and different categories of YouTube videos. Our findings demonstrate that bursts of public attention are strongly related to the popularity of media

products, especially the lifetime; Favorites, social influence, and searching behavior have negative impact on the burst of public attention; Recommendations and mobile communication have positive impact on bursts; Videos of news have the strongest burst, whereas videos of copyright-protected the weakest burst.

In summary, the temporality underlying different influences of information diffusion leads to the burst of public attention. The flow of public attention is also the process of the diffusion of online videos. Employing the digital traces of the diffusion of YouTube videos, this study serves as the starting point to empirically study the origin of burst of public attention. In the past, diffusion scholars primarily focus on the *cumulative* diffusion curves, e.g., the s-curve of diffusions (Rogers, 1983), while the line of research about burst of public attention shifts the perspective to the *stepwise* diffusion curves, which opens a new door to the hidden pattern of public attentions. This line of inquiry focuses on how exogenous and endogenous shocks shape the response or the memory of the systems (Sornette & Helmstetter, 2003). In the context of online videos, public attention responds to the online videos (e.g., YouTube videos), and gives rise to the burst of viewing behaviors, which sheds light to our understanding about the temporal patterns of information diffusions. Departure from the generalized fluctuation-dissipation theorem (Crane & Sornette, 2008; Nyquist, 1928; Sornette, et al., 2004; Weber, 1956), this study steps further to detect how influential factors on audience level and content level shape the response function of the social system (e.g., the bursts in public attention), in addition to exogenous shocks and social influences. The findings deepen our understanding of the hypothesis of temporality. As a theoretical generalization for the origin of burst in public attention, the temporality hypothesis argues that temporality could be used to different aspects,

either the lifetime of YouTube videos, or the diffusion channels, or even the content of the videos.

Bursts of public attention underline the temporality of public attention, which contributes to our knowledge about the dynamic evolutions and the underlying mechanisms of public attention. As the "species" living in the ecological environment, media products are given birth by the antecedents of public attention, and they live on public attention. However, since the principle of limited capacity for public attention, media products have to fiercely compete with each other to survive. One natural result of the principle of limited capacity for public attention is the temporality of media products. Newborn media products suddenly leap into prominence, remain there for a short time, and then fade away from the center of public attention (although still largely unnoticed).

Theoretically, the findings of this study render support for the theorization of temporality hypothesis of public attention. Conceptually, there are two aspects of temporality in public attention: the temporality of "latent" antecedents and the temporality of "manifest" temporal patterns (e.g., bursts). First, the "latent" antecedents are featured by their properties of temporality. E.g., exogenous shocks are very instantaneous. Second, the "manifest" temporality concerns bursts, precursors before bursts, and recoveries after bursts. The temporality hypothesis claims that the temporality underlying different antecedents of public attention leads to the bursts in public attention. In light of the temporality hypothesis, this present study steps further to detect how the temporality of latent antecedents shapes the manifest burst in public attention. Employing the digital traces of the diffusion of YouTube videos, and positioning it in communication research on public attention, this study serves as a starting point to empirically study the origin of bursts in public attention.

As a theoretical generalization for the origin of bursts in public attention, the temporality hypothesis exhibits itself in many concrete aspects, including popularity of YouTube videos, the diffusion channels, and the content of the videos. First, bursts are negatively *associated* with views and lifetime. Those videos of higher popularity, their bursts are relatively small. Put it in another way, the popularity of viral videos persists over time. Viral videos tend to seek continuous, incremental improvements, not great ups and downs. Thus, going viral means long-term growth rather than a temporary eruption. If the fragmentation of public attention implies "winner takes all", the temporality of public attention implies that the winner grows steadily over time. For example, the videos of news are mostly driven by exogenous shocks, and are featured by their temporality. Thus, the lifetimes of YouTube videos of news are significantly short. On the contrary, there are many videos which can easily live out the other videos, since they are of the public interests. Thus, they can continuously attract public attentions, and therefore they have very little temporality.

Second, the temporality of public attention also extends our understanding about the channels of information diffusion. Our study shows, at least for video sharing website, system recommendation is still the most important channel (53%) to "push" media products to audiences. However, since recommender systems update the content of recommendation fast, it is featured by temporality. On the contrary, audience interest revealed by users' searching behavior before viewing the video and evaluations after viewing the video promise a long-run diffusion of YouTube videos. Similarly, online social networking services make the flow of public attention through social networks successively and incrementally. Thus, social networks have the potential to spread information in a longer term.

Third, the temporality of public attention informs the classification of YouTube videos and their survival strategies. According to the origin of bursts and the extent to which social networks have been activated, Crane and Sornette (2008) practically classify YouTube videos into three categories: viral videos, quality videos, and junk videos. Viral videos ($0\% \leq F \leq 20\%$) are videos fueled by word-of-mouth, which primarily spread through social networks. Quality videos ($20\% < F < 80\%$) are similar to viral videos, but experience a sudden burst, and subsequently trigger a cascade through social networks. Junk videos ($80\% \leq F \leq 100\ \%$) experience a burst of activity for exogenous reason but fail to spread through social networks. Further, considering both the peak fraction and peak time of bursts, YouTube videos can be categorized into four kinds: large & early bursts, small & early bursts, small & late bursts, and large & late bursts.

Fourth, although this study focuses on the diffusion of online videos, the conclusions can be easily generalized to the other forms of information diffusion. For example, the conclusion can be applied to explain the limited diffusion size of online news (Steeg, et al., 2011). As I have demonstrated, of all the categories of information, news has the largest size of temporality. Therefore, news experiences strong bursts, and thus could not continuously diffuse so widespread as the prediction of epidemic models (Bogun á, Pastor-Satorras, & Vespignani, 2003; Pastor-Satorras & Vespignani, 2001).

It's also worth to note the limitations. Different from the hypotheses about audience interest and burst, I find ratings have positive influence on the burst of public attention. Ratings invoked by the content of YouTube videos tend to spark a temporary eruption of views, rather than a long-term growth of views. This may be caused by our measurement of ratings. In this study, rating is measured by the number of "likes" plus the number of "dislikes", rather than only

the number of "likes" or the number of "likes" minus the number of "dislikes". Similarly, comments also include both positive comments and negative comments. Thus, ratings and comments fail in capturing the real audience interest of YouTube videos. Negative evaluations (e.g., click the "dislike" button, and post a negative comment against a specific YouTube videos) distort the directions of influences on the burst of public attentions. On the contrary, favorites measure merely how many people set the video as "favorites". In the context of YouTube, favorites are a special playlist for YouTube videos which appears on the right side of the user's channel page. Compared with the measure of ratings and comments, the number of favorites is a better indicator for the audience interest of YouTube videos. Consistent with the hypothesis, I find negative relationship between favorites and burst of public attention.

Following the logic of the temporality hypothesis, two directions should be considered for future research: First, although we generalize temporality underlying various factors as the driver of the burst in public attention, further work is needed to both analytically model the diffusion mechanism and empirically operationalize the temporality of public attention, e.g., measuring the size and duration for both precursor before bursts and the recovery after burst. Second, as a temporal phenomenon, bursts of public attention have potentials to impact on both fragmentation and polarization of public attention. Therefore, it is necessary to investigate their mutual relationships.

Chapter 7. Conclusion and Discussion

The motivation of this study is to deal with the puzzlement of the limited size for online information diffusion. By looking into the competing and complementary diffusion mechanisms, this dissertation presents threshold models for information diffusion on information sharing websites (ISWs), and demonstrates how interpersonal effects, collective gatekeeping, burst of public attention, and the features of information determine the spread of information on ISWs (e.g., Sina Weibo, Digg, and YouTube).

Based on the results of three studies presented in Chapter 4, 5, and 6, this chapter synthesizes the findings and offers overall discussions and conclusions. First, the empirical findings are summarized, interpreted and discussed to answer the research questions. Second, theoretical contributions and implications are discussed. Finally, the limitations of this study as well as the recommendations for future research are presented.

7.1 Summary of Findings

In this section, I summarize and compare the overall findings of three studies, to get a full picture about interpersonal effects, collective gatekeeping, burst in public attention, and the limited size of information diffusion (see Table 9 for a brief summary).

Insert Table 9 here

7.1.1 Interpersonal Effects

The theoretical perspective of diffusion research shifts from individuals to networks. Information flows within social networks, along the ties, branches, and communities. Online social networks are increasingly supported by social networking services supplied by various kinds of social media, which underlines the salience of interpersonal influence, especially for online information diffusion. However, how does interpersonal effect influence information diffusion on ISWs is still an empirical question. The measurement of zero-threshold helps to estimate the percentage of interpersonal sources in information diffusion.

The findings about the percentage of interpersonal source vary with the information of different kinds and the ISWs of different diffusion mechanisms: for the diffusion of specific information on microblogs (e.g., the diffusion of a tweet), the percentage of interpersonal source is 77%. While for the diffusion of URLs on Twitter, the percentage of interpersonal source is only 16% (for more information, see Appendix A of this present dissertation); for the diffusion of news on Digg, the percentage of interpersonal source is 32%; for the diffusion of videos on YouTube, the percentage of interpersonal source is only 7%.

Therefore, the key point is whether there are alternative diffusion channels instead of interpersonal networks on the ISWs. For information diffusion on microblogs, the diffusion of specific tweet mainly relies on interpersonal networks, while the diffusion of URLs on microblog can both rely on exogenous impact and interpersonal networks, and compared with individuals' active sharing URLs to Twitter, the retweeting of URLs in microblog not so frequent.

Further, the threshold hypothesis of interpersonal effects asserts that there is a curvilinear relationship between interpersonal effects measured by mean threshold and diffusion size, which is exactly the situation described by the J-curve model of news diffusion (Greenberg, 1964b). For the information of minority interest, it tends to infect only a small size of relevant people.

Accordingly, information sharing is more active in the local network, and there is a negative relationship between interpersonal effects and diffusion sizes; while for the information of majority interest, the larger interpersonal effects are, and there is a positive relationship between interpersonal effects and diffusion size. In all, there is a nonlinear relationship.

There is a locked-in pattern for the information of minority interest. The information of less public interest is predestined to infect a few relevant people, since they are bound to be "locked" in local networks. Although there is a huge network which connects all of us on ISWs, the information of the daily life for the public can only spread in our small social circles. It rarely has the chance to spread to a non-trivial proportion of the nodes in the network. As a small world, we can reach any person in the network directly or with merely a few steps. However, if we are not popular or active enough, the information sharing by us would not be widely diffused. Thus, it helps us to understand the limited size of online information diffusion.

For the information of public interest, the stronger interpersonal effects are, the wider it could spreads. This pattern is firstly confirmed by the diffusion data of Sina Weibo, and then I find it again in the diffusion data of Digg. Interestingly, both of the cases, for the information with diffusion size larger than one hundred, interpersonal effects have a positive effect on diffusion size. The cutting point may not be one hundred in other cases, however, the basic logic holds. Although the concrete number of cutting point may be contingent upon many other factors, its existence clearly implies that two regimes of social influence exist for the information diffusion on information sharing websites. Thus, the claim of J-curve models still works well for online information diffusion. Bearing in mind that, in this case, interpersonal effects help to spread the information further, then why the overall size for the popular information diffusion is still much smaller than the expected size? I will try to elaborate it more in the later subsections.

7.1.2 Collective Gatekeeping

Collective gatekeeping is another important diffusion mechanism. Based on the collaborative filtering behaviors of ISWs users and the information aggregating service of ISWs, collective gatekeeping is argued to be more powerful than interpersonal effects for news diffusion on Digg. The findings confirm that only 32% of the people expose to information through their friends followed by them. Thus, most people learn news directly from the webpage of information aggregation.

Further, the findings also confirm that collective gatekeeping has positive influence on diffusion size. There is a question emerging naturally: when both collective gatekeeping and interpersonal effects are available for information diffusion, why collective gatekeeping is more powerful than interpersonal effects? In the specific study of news diffusion on Digg, I argue that, people tend to jump onto the bandwagon of collective gatekeepers in the news diffusion on SNW, which is consistent with the bandwagon heuristics hypothesis (Fu, 2012; Fu & Sim, 2011; Sundar, Oeldorf-Hirsch, & Xu, 2008).

Instead of comprehensively considering all the information available, individuals tend to rely on mental shortcuts to make decisions, which fits well in the principle of least effort proposed by George Zipf (1949). The principle of least effort postulates that human behaviors naturally choose the path of least resistance (Zipf, 1949). For example, information seekers tend to minimize their effort of information seeking, especially in choosing information channels (Allen, 1977; Rosenberg, 1967). Since popular news on the bandwagon of collective gatekeepers serves as the mental shortcuts for aggregated popularity (i.e., popularity heuristics), collective gatekeeping can effectively minimize searching cost (Lorenz, et al., 2011). Thus, for SNW users,

obtaining news from information aggregators is more efficient than reading news from their online friends.

7.1.3 Bursts of Public Attention

In addition to the diffusion mechanisms, the time dimension is also one concern of this study. In terms of temporality of influences, this study looks into the burst of public attention—the time stamp/fingerprint of temporality on daily diffusion curves. Following prior research (Crane & Sornette, 2008), I measure burst with peak fraction—the proportion of diffusion in the peak day. The results indicate that the average peak fraction is 0.827 for information diffusion on Sina Weibo, 0.71 for news diffusion on Digg, and 0.33 for video diffusion on YouTube. In addition to burst, lifetime is another indicator of temporality. The average lifetime is 18.7 days for information diffusion on Sina Weibo (*Mdn* = 9 days, *M* = 18.7 days, *SD* = 29.0), 14 days for news diffusion on Digg, and (*M*= 14.20, *SD*= 8.04), and a much larger 150 days on YouTube (*M* = 149.9; *SD* = 300.1). Thus, it is clear that video as one form of information has less temporality, compared with the information in the form of text.

It's also interesting to note that burst and lifetime are negatively correlated. The correlation coefficient is -0.37 on Sina Weibo ($r(3498) = -0.37$, $p < 0.001$), -0.1 on Digg ($r(3551) = -0.1$, $p < 0.001$), respectively. For the diffusion of videos on YouTube, the regression coefficient is -0.61 (Beta = -0.61, sig < 0.001). Further, lifetime is found to be positively related to diffusion size. The regression coefficient is 0.39 on Sina Weibo (Beta = 0.39, sig < 0.001), 0.02 on Digg (Beta = 0.02, sig < 0.01). Therefore, the logic is quite straightforward, if the information can survive for a longer time, it can reach a larger proportion of people.

7.1.4 The Limited Size of Information Diffusion

After the summary of the findings on interpersonal effects, collective gatekeeping, and bursts of public attention, I can now get back to the original research question of the present dissertation—the puzzlement of the limited diffusion size. The tension between theoretical propositions and empirical findings directs the present dissertation to look into the empirical reasons of limited diffusion size for online information diffusion.

In this dissertation, three studies are carried out to investigate the size of information diffusion and its determinants, especially the driving forces and bursts. All of the three studies demonstrate that the distribution of diffusion size is featured by a long tail or heavy tail. It could be either log-normal or power law. To be specific, for the popular information which has been shared for more than 100 times, the distribution is usually log-normal. While for the information covering a full range of diffusion size, the distribution is usually an obvious power law. This finding suggests that global diffusion is possible but rare, which is consistent with the proposition of Watts (2002).

However, the findings suggest that the absolute values of diffusion size are still limited and vary from different ISWs. The diffusion size for information diffusion on Sina Weibo and Digg is measured by the frequency of sharing (e.g., retweets, diggs). The maximum number of retweets on Sina Weibo is 422500, and the largest diggs on Digg is 24100. The maximum number of views for news on Digg is 47260, and the maximum number of views for YouTube videos is 176 million. If the size of information diffusion on microblogs and social news website is limited compared with the number of registered users, the largest diffusion size of YouTube videos is not trivial. Thus, as the distribution of long tail implies that, although global contagion is rare, it still exists for special cases in special forms. E.g., YouTube videos have relatively more

chance to go viral. However, for most information, the size of diffusion is very small. In terms of competition for public attention, the winner of popular information tends to take all.

First, interpersonal effects intensify the unequal distribution of diffusion size. Due to the highly skewed distribution of diffusion sizes, most information could not spread to a sizable fraction of people. The information is usually of small importance and high relevance for the users in the local network. According to the J-curve model of interpersonal effects, interpersonal effects have negative impacts on the diffusion size. Thus, such information tends to be trapped in the local social network. Since the proportion of such information is usually larger (more than 80%), therefore the probability of observing a global diffusion is very small. However, for the information of large importance, interpersonal effects tend to increase the diffusion size.

Second and similarly, collective gatekeeping also tends to intensify the unequal distribution of online information diffusion. The findings for popular news indicate that collective gatekeeping has positive influences on diffusion size. Collective gatekeeping contributes to the results of aggregated popularity. According to the technical features of ISWs, the aggregated popularity always manifests itself according to both recency and popularity. Thus, the winners take all, and the gap between popular information and non-popular information will be enlarged. For example, according to the bandwagon effect of aggregated popularity, when a video displays a high view count, its popularity over that of the other videos snowballs further (Fu, 2012; Fu & Sim, 2011).

Third, the limited diffusion depth explains the limited diffusion size in another way. Strikingly, although the diffusion size could be large than hundreds and thousands, the depth of information diffusion is extremely small (Mdn = 4, M = 4.9, SD = 2.9). Thus, the online information diffusion seems to confront with much higher network thresholds after a few steps of

information propagations, which is consistent with prior assertion about the weakness of long ties (Centola & Macy, 2007). One possible explanation is ISWs users of similar audience interest or taste tends to cluster together. Thus, there are fragmented social circles within the online social networks of ISWs, and therefore it is hard for information spread within one social circles to cross the boundary between this social circle and the other adjacent social circles (Ugander, et al., 2012; Zhang, Liu, Tang, Chen, & Li, 2013). However, this proposition about the weakness of structural holes and long ties still awaits further investigations.

Fourth, the temporal patterns of online information diffusion tend to limit the diffusion size. The temporal patterns of online information diffusion primarily presents itself in two aspects of the temporal patterns of online information diffusion—short lifetimes and strong bursts. In the first place, a short on-shelf time can definitely shorten the size of information diffusion. In the second place, such short lifetimes are often related with the strong bursts of public attention. As we have discussed before, bursts of public attention are negatively associated with the popularity of information diffusion. For example, news diffusion on SNW is featured by temporality due to the decay of novelty (F. Wu & Huberman, 2007) , which help explain the limited size of news diffusion (Steeg, et al., 2011). In light of the zero-sum feature of public attention, the rise of an issue is at the expense of other issues (McCombs & Zhu, 1995a; Zhu, 1992). News stories compete with each other for public attention. As a result, the fast update of news on the webpage of news aggregation can effectively restricts the size of news diffusion.

7.2 Contributions and Implications

Drawing on classic theories of diffusion research, the present dissertation contributes to our understanding in three aspects: first, I outlined a general model to depict the whole picture of

information diffusion; second, I looked into specific mechanisms of information diffusion, especially the threshold hypothesis of interpersonal effects, and the conceptualization of collective gatekeeping; and third, I articulated the temporality hypothesis of driving forces for information diffusion. Put together, the insights gained from the present work help us understand online information diffusion. The overall theoretical generalizations shed light on the mechanisms underlying information diffusion, and pave the way for more realistic diffusion models.

7.2.1 ABXCT Model of Information Diffusion

Drawing on classic communication models, I uses the ABXCT model as a conceptual framework to *map* and *integrate* current developments of information and communication technologies (ICTs) and to *understand* information diffusion on ISWs. As a schematic representation of the social reality, ABXCT helps us understand the relationships between components and ensures that we do not overlook whole areas.

First, as a development of classic communication models, ABXCT model attempts to synthesize both theoretical advancement and the emergence of new ICTs. For example, the most important technical development during the past decades is the emergence of web 2.0, especially the social media. For one thing, social networking service empowers individuals to build their connections online with each other with ease. To date, social media with social networking services have dominated the market of Internet. E.g., Facebook, Twitter, and Google+ have become the most popular websites. Thus, the theoretical significance of interpersonal effects on information diffusion is emphasized. Correspondingly, social ties among the networked individuals are always assumed or taken for granted as the channels for online information diffusion. For another, although bottom-up information diffusion along social connections is

very powerful, we should not ignore the other top-down diffusion channels. One example of the top-down diffusion channel is information aggregators and search engines used in ISWs. As we have demonstrated in Chapter 5, information aggregation based on collaborative news filtering makes the information jump within the social graph, which indicates the strong effect of collective gatekeeping on news diffusion on Digg.

Second, ABXCT model serves as a map to position concrete studies on information diffusion in a systematic way. Diffusion researchers can put their studies into this general map to compare their research with others. For example, this dissertation primarily deals with the relationship between networked individuals and ISWs, especially how social networking services and information aggregators facilitate information diffusion. Thus, scholars can easily find out what has been done and what has not. E.g., both political and business organizations recently realize the importance of releasing information directly from their social media account. Therefore, they do not totally rely on mass media to communicate with the public. It is worth to study how organizations use social media and its impact on the logic of media production today (Waters, Burnett, Lamm, & Lucas, 2009).

Third, ABCXT model highlights the importance of information itself (X) and temporal patterns (T). ABCXT models attempts to integrate the theoretical ingredient of the J-curve model. Note that, not only the attributes of information, but also the inter-relationship between different information. In terms of attention economy (Davenport & Beck, 2001; Goldhaber, 1997), different messages compete for the public attention, especially when information is surplus and attention is scarce, which intensifies the temporality of information diffusion. Further, to capture the temporal patterns of information diffusion, the time dimension is spelled out and added into the model. It is necessary to note that, time dimension of ABXCT model leads this dissertation to

investigate the temporal patterns of information diffusion, especially the temporality of influences.

7.2.2 Threshold Hypothesis of Interpersonal Effects

There have been many ways to capture interpersonal effects. Prior research primarily focuses on to what extent interpersonal channels serve as information sources. Therefore, classic research of news diffusion investigates the proportion of people who learn the news from their interpersonal networks. One of the most important generalizations derived from this line of research is the J-curve model (Greenberg, 1964b). J-curve model asserts that the relevance and importance interact with each other to influence interpersonal effects on diffusion size. First, if the information is of minority interest and high relevance, the interpersonal effects are stronger, while the diffusion size is smaller. Therefore, there is a negative relationship between interpersonal effects and diffusion size; second, if the information is of larger importance, there is a positive relationship between interpersonal effects and diffusion size. Put together, there is a curvilinear relationship between interpersonal effects and diffusion size. However, capturing interpersonal effects by measuring the overall proportion of learning information from interpersonal sources, such studies fail in fully capturing the strength of networks.

Threshold models supply another perspective to gauge to what extent interpersonal effects influence diffusion size. Threshold models concerns to what extent individuals' local network has been activated when he/she make up his/her decisions on information sharing. Thus, threshold models shift the perspective of analysis from individuals to ego networks, and attempt to tackle the tension between local activation and global diffusion. In this dissertation, I focus on how the aggregated network threshold (e.g., mean threshold) is related with the overall diffusion size.

This dissertation contributes to our understanding of information diffusion by proposing the threshold hypothesis of interpersonal effects. In terms of J-curve models, for information of public interest, mean threshold of interpersonal effects has positive impact on diffusion size; for information of local interest, mean threshold of interpersonal effects has negative impact on diffusion size. Consistent with the J-curve, one research reveals that there are two regimes of social influence in the diffusion of Facebook applications (Jukka-Pekka Onnela & Reed-Tsochas, 2010). Once Facebook applications cross a particular threshold of popularity, social influence has significant influence on the diffusion of Facebook applications; below this threshold, social influence appears to vanish. Thus, the study of information diffusion on Sina Weibo confirms the two regimes of interpersonal effects, which deepens our understanding about the role of interpersonal effect on information diffusion.

7.2.3 Collective Gatekeeping

In the process of information sharing, everyone acts as the gatekeeper for his/her followers. One problem of information diffusion driven merely by interpersonal effects lies in path-dependent diffusion process. As a result, if the individuals occupying the structural holes reject to spread the information, the size of information diffusion can be significantly reduced. Thus, the diffusion is fragile, as it has been found in the study of information cascade theory (Chamley, 2004; Easley & Kleinberg, 2010). Similarly, just as the study of information diffusion on Sina Weibo has shown, information tends to be trapped within the local area spanning only a few steps.

Information aggregator supplies one way to overcome the path-dependent problem of information diffusion, and makes collective gatekeeping possible. Collective gatekeeping consist two processes: the first phase is collaborative filtering of early diffusers as the effort of collective

gatekeeping; and the second phase is the system recommendation as the outcome of collective gatekeeping. To date, recommendation systems are extensively used by various kinds of websites, and recommendation by aggregated popularity has been widely used to capture the trends of information diffusion on social media. E.g., on Twitter, the most popular information within a time period will be aggregated as a list of trending topic. On Digg, the most popular information within certain time period were also aggregated and demonstrated on the webpage of information aggregator.

Collective gatekeeping as a conceptualization contributes to the well-established theory of gatekeeping (Lewin, 1947; Shoemaker, 1991). Gatekeeping is "the process by which the billions of messages that are available in the world get cut down and transformed into the hundreds of messages that reach a given person on a given day" (Shoemaker & Vos, 2009). Kurt Lewin (1947) coins the word called "gate keeping" to describe housewives as gatekeepers for the food of the family. Later, gatekeeping is used to describe the news selection process within a media (White, 1950). The major concern of gatekeeping as a communication process is how the gatekeepers select news. There are some limitations of gatekeeping theory. E.g., there is only one main "gate area", and only a few professional individuals act as gatekeepers (MaQuail & Windahl, 1993).

The development of Internet has intensified such criticism about the gatekeeping mechanisms. Against the backdrop of classic gatekeeping theory, everyone on information sharing websites is connected together, and everyone plays the role of gatekeeper. By deciding to share the information or not, individuals act as the gatekeeper for their followers on ISWs. By collectively sharing the information and making the information appear as trending topics on

information aggregators, collective gatekeepers transform gatekeeping into a dynamic and collective process.

Collective gatekeeping has the potential to spread the information further. If the problem of interpersonal effects on information diffusion is the locked-in phenomenon, collective gatekeeping help information bypass the nodes with higher network thresholds, and make the information spread further. With the perspective of threshold models, non-interpersonal effects can be captured by zero-threshold and zero-threshold ratio. The results shows more than 70% of are the news diffusion on Digg can be attributed to collective gatekeeping. As a result, we see the information propagate by jumping in social networks.

Conceptually, the strength of collective gatekeeping can be attributed to the general principle of least effort (Zipf, 1949). The principle of least effort states that human behaviors follow the principle of least effort. Thus, information seekers tends to minimize the effort of information diffusion, especially in choosing information channels (Allen, 1977; Rosenberg, 1967). As we have mentioned before, people tend to jump onto the bandwagon of collective gatekeepers according to bandwagon heuristics hypothesis (Fu, 2012; Fu & Sim, 2011; Sundar, et al., 2008), since bandwagon heuristics supply mental shortcuts for individuals' decision-making. Currently, there is an increasing information overload with the development of social media, especially the advent of social media. It has become even harder to efficiently select good information to share. Therefore, the public increasingly tend to refer to the information recommended by collective gatekeepers.

The relative strength of collective gatekeeping contributes to our understandings on the theory of information cascade (Chamley, 2004; Easley & Kleinberg, 2010). Information cascade theory is framed in the context of making decisions with both private information and external

information. A cascade develops when people "abandon their own information in favor of inferences based on earlier people's actions" (Easley & Kleinberg, 2010). There are some basic assumptions for information cascade theory: First, agents make decisions sequentially; second, agents make decisions rationally based on the information they have; third, agents do not have access to the private information of others; and fourth, a limited action space exists (e.g. a sharing/ignoring decision).

However, with the perspective of network analysis, threshold models argue that people have some notion of the private beliefs of those in their local network. Therefore, information from interpersonal network constructs an important part of agents' private information, which relaxes the assumption of information cascades that agents have no information about the private information of others. According to the theory of information cascade, the "crowd" in the context of Digg is the aggregated decisions made by collective gatekeepers. This finding is consistent with prior research. E.g., Bala and Goyal (1998) show that if a group of individuals is observable by all agents, they may initiate an informational cascade. Such group is named as "royal family". Thus, the finding about collective gatekeeping overriding interpersonal effects demonstrates that people are not only act on observable behaviors taken by concrete individual, instead, they are more likely to follow the aggregated decisions made by collective gatekeepers.

7.2.4 Temporality Hypothesis of Diffusion Driving Forces

Bursts of public attention underline the temporality of public attention, which refines our knowledge about the dynamic evolutions and the underlying mechanisms of public attention. As the "species" living in the ecological environment, media products are given birth by the antecedents of public attention, and they live on public attention. However, since the principle of limited capacity for public attention (Kahneman, 1973; Norman & Bobrow, 1975), media

products have to fiercely compete with each other to survive. One natural result of the principle of limited capacity for public attention is the temporality of media products. Newborn media products suddenly leap into prominence, remain there for a short time, and then fade away from the center of public attention (although still largely unnoticed). Burst of public attention describes the sudden and violent growth and decline of collective attention. The most remarkable feature of information diffusion in the temporal dimension is burst. Bursts offer deep insights into the rhythm of how we collectively allocate attentions, and reflect the temporality of driving forces.

The temporality hypothesis conjectures that the temporality of different driving forces for information diffusion is the determinant factor governing the rise and fall of public attention allocated to specific media product. Based on the influences of the allocation of public attention (e.g., exogenous impacts) and the underlying principles (e.g., the limited capacity for public attention), one speculation is the limited duration of driving forces for information diffusion generates burst in public attention. Measuring burst with peak fraction for the diffusion of YouTube videos, the results show that burst is negatively related to the popularity, search, and social influence of online videos, but positively related to recommendation. Of various categories of videos, news videos experience the strongest bursts, whereas copyright-protected videos the weakest bursts. Thus, this present dissertation contributes to our knowledge by articulating the temporality hypothesis which builds a temporal link between different influential factors of information diffusion and burst of public attention.

Theoretically, the findings of this study render support for the theorization of temporality hypothesis of public attention. Conceptually, there are two aspects of temporality in public attention: the temporality of "latent" antecedents and the temporality of "manifest" temporal

patterns (e.g., bursts). First, the "latent" antecedents are featured by their properties of temporality. E.g., exogenous shocks are very instantaneous. Second, the "manifest" temporality concerns bursts, precursors before bursts, and recoveries after bursts. The temporality hypothesis claims that the temporality underlying different antecedents of public attention leads to the bursts in public attention. In light of the temporality hypothesis, this present study steps further to detect how the temporality of latent antecedents shapes the manifest burst in public attention. Employing the digital traces of the diffusion of YouTube videos, and positioning it in communication research on public attention, this study serves as a starting point to empirically study the origin of bursts in public attention.

As a theoretical generalization for the origin of bursts in public attention, the temporality hypothesis exhibits itself in many concrete aspects, including popularity of YouTube videos, the diffusion channels, and the content of the videos. First, bursts are negatively associated with views and lifetime. Those videos of higher popularity, their bursts are relatively small. Put it in another way, the popularity of viral videos persists over time. Viral videos tend to seek continuous, incremental improvements, not great ups and downs. Thus, going viral means long-term growth rather than a temporary eruption. If the fragmentation of public attention implies "winner takes all", the temporality of public attention implies that the winner grows steadily over time. Second, the temporality of public attention informs the categories of YouTube videos and their survival strategies. According to the origin of bursts and the extent to which social networks have been activated, Crane and Sornette (2008) practically classify YouTube videos into three categories: viral videos, quality videos, and junk videos. Viral videos ($0\% \leq F \leq 20\%$) are videos fueled by word-of-mouth, which primarily spread through social networks. Quality videos ($20\% < F < 80\%$) are similar to viral videos, but experience a sudden burst, and subsequently trigger a

cascade through social networks. Junk videos ($80\% \leq F \leq 100\ \%$) experience a burst of activity for exogenous reason but fail to spread through social networks. Further, considering both the peak fraction and peak time of bursts, YouTube videos can be categorized into four kinds: large & early bursts, small & early bursts, small & late bursts, and large & late bursts.

7.2.5 Theoretical and Practical Implications

In addition to the major theoretical contributions, the present dissertation also demonstrates some other implications in both theoretical and methodological aspects. First, the present dissertation illustrates the communication power of networked individuals in the age of social media. According to the idea of Wellman (2002), individuals had been separated and trapped within the densely-connected "little box". However, with the evolution of social networks and the emergence of Internet and mobile, social lives have been transformed "from densely knit family, neighborhood, and group relationship towards more far-flung, less tight, more diverse personal networks" (Rainie & Wellman, 2012). Person-to-person network suggests how community has transcended group boundaries. The social change from small groups to broader personal networks has been greatly advanced by the Internet, especially social media. On the participatory medium of ISWs, networked individuals are "creative audiences" with communication powers (Castells, 2009, p. 316).

Second, digital traces of human communication behaviors have been extensively employed in this dissertation. Based on the behavioral, relational, longitudinal, and big data generated from digital traces, the measurements could be more solid and convincing. Instead of relying too much on the latent psychological mediations (e.g., attitude, intention, and willingness), digital traces grant diffusion scholars an opportunity to unobtrusively "observe" the dynamic process of online information diffusion.

Third, public attention has become the "currency" of digital world. Both information and attention had been the most important concepts in communication research. However, just as the logic of attention economy argues, in the global digital age, the audiences are overloaded with the excessive supply of information. However, the capacity of public attention is relatively limited. Further, the information diffusion can also be viewed as the flow of attention. Note the directions of information diffusion and attention flow are reversed. In our everyday use of the Internet, individuals selectively allocate their attentions to different websites and concrete information. In practice, public attention can be naturally measured, such as the number of views, retweets, comments, favorites. Thus, in this way, we are trying to establishing a link between diffusion research and attention research. For example, diffusion channels can help attention scholars understand the bursts of public attention, and the limited capacity of attention can help diffusion researchers interpret the temporality of information diffusion.

7.3 Limitations and Directions for Future Research

First, the major limitation of the present dissertation lies in its deficiency of analytic models. As a well-established stream of research, diffusion phenomena have been long studied. In addition to statistical analysis, analytical models have also been widely employed in diffusion research, such as aforementioned Bass diffusion models (Bass, 1969, 2004) and epidemic models (e.g., SIR, SIS) (Keeling & Rohani, 2011; Kermack & McKendarick, 1927; Ross, 1911). However, the analytical models are based on top-down deductions which have the risk of deviating from social realities. Statistical models of information diffusion help us to gain insights and to learn about the hidden principles underlying diffusion patterns. The findings of this present dissertation help build a foundation for future investigations of these and other important issues about information diffusion. For example, previous analytical models primarily focus on

the person-to-person cascades which are usually driven by interpersonal influence, such as persuasion and imitation. However, exogenous impacts (e.g., system recommendations) also play an important role in information diffusion. Thus, it is necessary to incorporate exogenous impacts into analytical models.

Second, it is necessary to note the differences between interpersonal effects in information diffusion and social influences in behavioral contagion. Interpersonal effects in information diffusion primarily emphasize on individuals' exposures, awareness, and attitudes, while social influences in social contagion mainly concern the other behavioral changes of higher costs, such as adoption of new products, participation in social mobilizations. The line of research on social influences in social contagion aims in separating correlation from causation by distinguishing homophily (Aral, et al., 2009), simultaneity (Godes & Mayzlin, 2004), unobserved heterogeneity (Van den Bulte & Lilien, 2001), and other contextual and correlated effects (Manski, 1993). In the context of information diffusion, it is possible that, at the same time of interpersonal exposures to information, networked individuals have also been influenced by exogenous factors (e.g. media effect). Thus, interpersonal effects measured by network thresholds may overestimate social influence. Although to accurately estimate social influences in the process of online information diffusion is beyond the scope of the present dissertation, it may become the concern of future research on information diffusion.

Third, the scope of the present dissertation primarily focuses on the information diffusion on one information sharing website. It is also necessary to trace the whole information diffusion process on the internet. For example, researchers should look into the information sharing on news media, blogs, microblogs, online social network sites, which extends the diffusion research to scope of the inter-media diffusion. However, in such cases, it is usually hard to trace the

accurate diffusion networks and raises great challenges for the large-scale research (Kim, Newth, & Christen, 2013).

Fourth, although this present dissertation is a network analysis of information diffusion, the focus of analysis is the ego network. Therefore, the global structural properties of social networks are not sufficiently considered. Similar problems exist for the analysis of the diffusion networks. Another concern is related to the interplay between social networks and diffusion networks. While these works are not the central focus of the present dissertation, in the light of specific research purposes, it is still necessary to incorporate such analysis into the future research.

References

Allen, T. J. (1977). *Managing the Flow of Information*. Cambridge, MA: MIT Press.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences, 106*(51), 21544-21549.

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science, 57*(9), 1623-1639.

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2011). Spatio-temporal analysis of topic popularity in Twitter. *arXiv preprint arXiv:1111.2904*.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of the fourth ACM international conference on Web search and data mining, USA, 11*, 65-74.

Bakshy, E., Karrer, B., & Adamic, L. A. (2009). Social influence and the diffusion of user-created content. *Proceedings of the 10th ACM conference on electronic commerce, USA. 10*, 325-334.

Bala, V., & Goyal, S. (1998). Learning from neighbours. *The Review of Economic Studies, 65*(3), 595-621.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Baran, S. J., & Davis, D. K. (2011). *Mass communication theory: Foundations, ferment, and future* (6th ed.). Boston, MA: Wadsworth Publishing Company.

Bass, F. (1969). A New Product Growth Model for Consumer Durables. *Management science, 15*(5), 215-227.

Bass, F. (2004). Comments on “A new product growth for model consumer durables": The Bass Model. *Management science, 50*(12), 1833-1840.

Becker, N. G. (1992). Infectious diseases of humans: Dynamics and control. *Australian Journal of Public Health, 16*(2), 208-209.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205.

Bernoulli, D. (1760). Essai d’une nouvelle analyse de la mortalite causee par la petite verole. *Memoires de Mathematiques et de Physique* (pp. 1-45). Paris: Acadademie Royale Sciences.

Bogun á, M., Pastor-Satorras, R., & Vespignani, A. (2003). Absence of epidemic threshold in scale-free networks with degree correlations. *Physical Review Letters, 90*(2), 28701. doi: 10.1103/PhysRevLett.90.028701

Bordogna, C. M., & Albano, E. V. (2007). Statistical methods applied to the study of opinion formation models: A brief overview and results of a numerical study of a model based on the social impact theory. *Journal of Physics-Condensed Matter, 19*(6), 06514401-06514416.

Bowman, S., & Willis, C. (2003). We Media: How Audiences are Shaping the Future of News and Information. *The Media Center at The American Press Institute. Reston, VA*.

Braddock, R. (1958). An extension of the “Lasswell Formula”. *Journal of Communication, 8*(2), 88-93.

Budd, R. W., MacLean Jr, M. S., & Barnes, A. M. (1966). Regularities in the diffusion of two major news events. *Journalism Quarterly, 43*(22), 221-230.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology, 92*(6), 1287-1335.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of modern physics, 81*(2), 591-646.

Castells, M. (1996). *The Rise of the Network Society*. Massachusetts: Blackwell Publishing.

Castells, M. (2007). Communication, power and counter-power in the network society. *International Journal of Communication, 1*(1), 238-266.

Castells, M. (2009). *Communication power*. New York: Oxford University Press.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194-1197.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology, 113*(3), 702-734.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 1-14. doi: 10.1145/1298306.1298309

Cha, M., Pérez, J., & Haddadi, H. (2009). *Flash floods and ripples: The spread of media content through the blogosphere*. Paper presented at the The Third International Conference on Weblogs and Social Media, San Jose, CA.

Chaffee, S. (1975). The diffusion of political information. In S. Chaffee (Ed.), *Political communication: issues and strategies for research* (Vol. 4, pp. 85-128). Beverly Hills, CA: Sage.

Chamley, C. (2004). *Rational herds: Economic models of social learning*. Cambridge: Cambridge University Press.

Chowdhury, S., & Landoni, M. (2006). News aggregator services: User expectations and experience. *Online Information Review, 30*(2), 100-115.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661-703

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician, 35*(1), 54-54.

Coleman, J. S., Katz, E., & Menzel, H. (1966). *Medical innovation: A diffusion study*. New York, NY: Bobbs-Merrill Company

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *Political polarization on Twitter.* Paper presented at the International Conference on Weblogs and Social Media, Barcelona, Spain.

Cooper, S. D. (2006). *Watching the watchdog: Bloggers as the fifth estate*. Spokane, Washington: Marquette Books.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences, 105*(41), 15649.

Davenport, T. H., & Beck, J. C. (2001). *The attention economy: Understanding the new currency of business*. Boston: Harvard Business School Press.

DeFleur, M. L. (1987). The growth and decline of research on the diffusion of the news, 1945-1985. *Communication Research, 14*(1), 109-130.

DeFleur, M. L., & Cronin, M. M. (1991). Completeness and accuracy of recall in the diffusion of the news from a newspaper vs. a television source. *Sociological Inquiry, 61*(2), 148-166.

Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J. B., & Vujnovic, M. (2008). Participatory journalism practices in the media and beyond. *Journalism practice, 2*(3), 326-342.

Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution, 22*(6), 469-493.

Dutton, W. H. (2009). The fifth estate emerging through the network of networks. *Prometheus, 27*(1), 1-15.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge: Cambridge University Press.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, California: Stanford University Press.

Figueiredo, F., Benevenuto, F., & Almeida, J. M. (2011). The tube over time: Characterizing popularity growth of Youtube videos. *Proceedings of the fourth ACM international conference on Web search and data mining*, 745-754. doi: 10.1145/1935826.1935925

Finamore, A., Mellia, M., Munafò, M. M., Torres, R., & Rao, S. G. (2011). Youtube everywhere: Impact of device and infrastructure synergies on user experience. *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 345-360. doi: 10.1145/2068816.2068849

Frey, D. (1986). Recent research on selective exposure to information. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 41-80). Orlando, Florida: Academic Press.

Fu, W. W. (2012). Selecting online videos from graphics, text, and view counts: The moderation of popularity bandwagons. *Journal of Computer-Mediated Communication, 18*(1), 46-61.

Fu, W. W., & Sim, C. C. (2011). Aggregate bandwagon effect on online videos' viewership: Value uncertainty, popularity cues, and heuristics. *Journal of the American Society for Information Science and Technology, 62*(12), 2382-2395.

Funkhouser, G. R., & McCombs, M. E. (1971). The rise and fall of news diffusion. *The Public Opinion Quarterly, 35*(1), 107-113.

Gans, H. J. (1979). Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time. Chicago: Northwestern University Press.

Gantz, W., & Bradley, S. D. (2005). Review and criticism—Bradley S. Greenberg: Advancing communication research, making a difference. *Journal of Broadcasting & Electronic Media, 49*(1), 135-146.

Glasgow, K., & Fink, C. (2013). Hashtag lifespan and social networks during the London riots. In W. G. Kennedy & N. D. Bos (Eds.), *Social computing, behavioral-cultural modeling and prediction* (pp. 311-320). Verlag Berlin Heidelberg: Springer.

Glynn, C. J., & Park, E. (1997). Reference groups, opinion intensity, and public opinion expression. *International Journal of Public Opinion Research, 9*(3), 213-232.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science, 23*(4), 545-560.

Goldhaber, M. H. (1997). The attention economy and the net. *First Monday, 2*(4), Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/519/440.

Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of dunbar's number. *PloS one, 6*(8), e22656.

González-Bailón, S., Borge-Holthoefer, J., Baños, R., & Moreno, Y. (2013). Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks, In press*.

González-Bailón, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific reports, 1*(197), 1-7.

Goode, L. (2009). Social news, citizen journalism and democracy. *New Media & Society, 11*(8), 1287-1305.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420-1443.

Granovetter, M., & Soong, R. (1983). Threshold models of diffusion and collective behavior. *The Journal of Mathematical Sociology, 9*(3), 165-179.

Granovetter, M., & Soong, R. (1986). Threshold models of interpersonal effects in consumer demand. *Journal of Economic Behavior and Organization, 7*(1), 83-99.

Granovetter, M., & Soong, R. (1988). Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence. *Sociological Methodology, 18*(6), 69-104.

Greenberg, B. S. (1964a). Diffusion of news of the Kennedy assassination. *Public Opinion Quarterly, 28*(2), 225-232.

Greenberg, B. S. (1964b). Person to person communication in the diffusion of a news event. *Journalism Quarterly, 41*(3), 489-494.

Headen, R. S., Klompmaker, J. E., & Rust, R. T. (1979). The duplication of viewing law and television media schedule evaluation. *Journal of Marketing Research, 16*(3), 333-340.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology, 21*(1), 107-112.

Hethcote, H. W. (2009). The basic epidemiology models: Models, expressions for R_0, parameter estimation, and applications. In S. Ma & Y. Xia (Eds.), *Mathematical understanding of infectious disease dynamics*. Singapore: World Scientific Publishing Company.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635-650.

Huberman, B. A. (2008). Crowdsourcing and attention. *Computer, 41*(11), 103-105.

Huberman, B. A., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday, 14*(1), Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2317/2063.

Im, Y. H., Kim, E., Kim, K., & Kim, Y. (2011). The emerging mediascape, same old theories? A case study of online news diffusion in Korea. *New Media & Society, 13*(4), 605.

Iribarren, J. L., & Moro, E. (2009). Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters, 103*(3), 387021-387024.

Ivković, Z., & Weisbenner, S. (2007). Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. *Review of Financial Studies, 20*(4), 1327-1357.

Kadushin, C. (2006). Personal influence: A radical theory of action. *The ANNALS of the American Academy of Political and Social Science, 608*(1), 270.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Karsai, M., Kivela, M., Pan, R. K., Kaski, K., Kertesz, J., Barabasi, A. L., & Saramaki, J. (2011). Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E, 83*(2), 0251021-0251024.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly, 21*(1), 61-78. doi: 10.1086/266687

Katz, E., & Lazarsfeld, P. F. (1955). Personal influence: The part played by people in the flow of mass communications. New York: Free Press.

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals*. Princeton & Oxford: Princeton University Press.

Kermack, W. O., & McKendarick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, 115*, 700-721.

Kim, M., Newth, D., & Christen, P. (2013). Modeling dynamics of diffusion across heterogeneous social networks: News diffusion in social media. *Entropy, 15*(10), 4215-4242.

Kincaid, D. L. (1979). The convergence model of communication. Honolulu: East-West Communication Institute.

Klamer, A., & Van Dalen, H. P. (2002). Attention and the art of scientific publishing. *Journal of economic methodology, 9*(3), 289-315.

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science, 311*(5757), 88-90.

Krassa, M. A. (1988). Social groups, selective perception, and behavioral contagion in public opinion. *Social Networks, 10*(2), 109-136.

Lang, A. (1995). Defining audio/video redundancy from a limited-capacity information processing perspective. *Communication Research, 22*(1), 86-115.

Lang, A. (2006). Using the limited capacity model of motivated mediated message processing to design effective cancer communication messages. *Journal of communication*, *56*(s1), S57-S80. doi: 10.1111/j.1460-2466.2006.00283.x

Lang, A., Geiger, S., Strickwerda, M., & Sumner, J. (1993). The effects of related and unrelated cuts on television viewers' attention, processing capacity, and memory. *Communication Research, 20*(1), 4-29. doi: 10.1177/009365093020001001

Lang, A., Newhagen, J., & Reeves, B. (1996). Negative video as structure: Emotion, attention, capacity, and memory. *Journal of Broadcasting & Electronic Media, 40*(4), 460-477.

Larsen, O. N., & Hill, R. J. (1954). Mass media and interpersonal communication in the diffusion of a news event. *American Sociological Review*, *19*(4), 426-433.

Lasica, J. D. (2003). What is participatory journalism. *Online Journalism Review, 7*(08), 2003.

Lasswell, H. D. (1948). The structure and function of communication in society. In Bryson (Ed.), *The communicaiton of ideas*. New York: Harper and Brothers.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). The people's choice: How the voter makes up his mind in a presidential election. *New York: Duell, Sloan and Pearce*.

Lazarsfeld, P. F., & Franzen, R. H. (1945). Prediction of political behavior in America. *American Sociological Review, 9*(1), 261-273.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L.,Brewer, D., Christakis, N., & Contractor, N., Fowler, J.,Gutmann, M.,Jebara, T.,King, G.,Macy, M.,Roy, D.,Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721-723.

Lehmann, J., Gon çalves, B., Ramasco, J. J., & Cattuto, C. (2012). *Dynamical classes of collective attention in Twitter.* Paper presented at the Proceedings of the 21st international conference on World Wide Web.

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In W. W. Cohen & S. Gosling (Eds.), *Proceedings of 4th International Conference on Weblogs and Social Media*. Washington, DC, USA: The AAAI Press.

Leskovec, J., Singh, A., & Kleinberg, J. (2006). Patterns of influence in a recommendation network. *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining, Heidelberg, 10,* 380-389.

Lewin, K. (1947). Frontiers in group dynamics II: Channels of group life; Social planning and action research. *Human Relations, 1*, 143-153.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social Selection and Peer Influence in an Online Social Network. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 68-72.

Lewis, T. G. (2011). *Network science: Theory and applications*. Hoboken, New Jersey: Wiley.

Lippe, P. v. d. (1994). The dynamics of informational cascades: The Monday demonstrations in Leipzig, 1989–1991. *World Politics, 47*, 42-101.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020.

Lou, T., & Tang, J. (2013). *Mining structural hole spanners through information diffusion in social networks*. Paper presented at the International World Wide Web Conference Committee, Rio de Janeiro, Brazil.

Lyytinen, K., & Damsgaard, J. (2001). What's wrong with the diffusion of innovation theory. In M. A. Ardis & B. L. Marcolin (Eds.), *Diffusing Software Products and Process Innovations* (pp. 173-190). Norwell, Massachusetts: Springer.

Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In W. Jackson (Ed.), *Communication Theory* (pp. 503-512). New York: Academic Press.

Manski, C. F. (1993). Identification problems in the social sciences. *Sociological Methodology, 23*, 1-56.

MaQuail, D., & Windahl, S. (1993). *Communication models: For the study of mass communication* (2 ed.). New York: Addison Wesley Longman.

Markus, M. L. (1987). Toward a "critical mass" theory of interactive media: Universal access, interdependence and diffusion. *Communication Research, 14*(5), 491-511.

McCombs, M. (1981). The agenda-setting approach. *Handbook of political communication*, 121-140.

McCombs, M., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176-187.

McCombs, M., & Zhu, J. H. (1995a). Capacity, diversity, and volatility of the public agenda. *Public Opinion Quarterly, 59*(4), 495-525.

McCombs, M., & Zhu, J. H. (1995b). Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. *Public Opinion Quarterly, 59*(4), 495-525. doi: 10.1086/269491

Microblogging. (2013). In Wikipedia. Retrieved March 27, 2013, from http://en.wikipedia.org/wiki/Microblogging.

Miller, D. C. (1945). A research note on mass communication: How our community heard about the death of president Roosevelt. *American Sociological Review, 10*(5), 691-694.

Neuman, W. R. (1990). The threshold of public attention. *Public Opinion Quarterly, 54*(2), 159-176.

Newcomb, T. M. (1953). An approach to the study of communicative acts. *Psychological review, 60*(6), 393-404.

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication, 24*(2), 43-51.

Noelle-Neumann, E. (1993). *The spiral of silence: Public opinion--our social skin* (2nd ed.). Chicago: University of Chicago Press.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive psychology, 7*(1), 44-64. doi: 10.1016/0010-0285(75)90004-3

Nyquist, H. (1928). Thermal agitation of electric charge in conductors. *Physical review, 32*(1), 110-113. doi: 10.1103/PhysRev.32.110

Oliveira, J. G., & Barabási, A. L. (2005). Human dynamics: Darwin and Einstein correspondence patterns. *Nature, 437*(7063), 1251-1251.

Onnela, J.-P., & Reed-Tsochas, F. (2010). Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences, 107*(43), 18375-18380. doi: 10.1073/pnas.0914572107

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D. K., Kimmo, Kertész, J., & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences, 104*(18), 7332-7336.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters, 86*(14), 3200-3203.

Pollmann, T. (1998). On forgetting the historical past. *Memory & cognition, 26*(2), 320-329.

Price, V. (1992). *Communication concepts 4: Public opinion*. Newbury Park, CA: Sage.

Prior, M. (2005). News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout. *American Journal of Political Science, 49*(3), 577-592.

Rainie, L., & Wellman, B. (2012). *Networked: The new social operating system*. Cambridge, Massachusetts: MIT Press.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., & Vespignani, A. (2010). Characterizing and modeling the dynamics of online popularity. *Physical Review Letters, 105*(15), 158701.

Robins, G., Pattison, P., & Woolcock, J. (2005). Small and other worlds: Global network structures from local processes. *American Journal of Sociology, 110*(4), 894-936.

Roehner, B., Sornette, D., & Andersen, J. (2004). Response functions to critical shocks in social sciences: An empirical and numerical study. *International Journal of Modern Physics C, 15*(6), 809-834. doi: 10.1142/S0129183104006236

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). New York: The free press.

Rogers, E. M. (2000). Reflections on news event diffusion research. *Journalism & Mass Communication Quarterly, 77*(3), 561-576.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: The free press.

Rogers, E. M., & Bhowmik, D. K. (1970). Homophily-heterophily: Relational concepts for communication research. *Public Opinion Quarterly, 34*(4), 523-538.

Rogers, E. M., & Kincaid, D. L. (1981). *Communication networks: Toward a new paradigm for research*. New York: Free Press.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political Hashtags, and complex contagion on Twitter. *Proceedings of the 20th international conference on World wide web, India, 20*, 695-704.

Rosenberg, V. (1967). Factors affecting the preferences of industrial personnel for information gathering methods. *Information Storage and Retrieval, 3*(3), 119-127.

Ross, R. (1911). *The prevention of Malaria* (2 ed.). London: Murray.

Ryan, B., & Gross, N. C. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural sociology, 8*(1), 15-24.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5762), 854-856.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology, 1*(2), 143-186.

Schramm, W. (1954). How communication works. In W. Schramm (Ed.), *The process and effects of mass communication*. Urbana: University of Illinois Press.

Scott, J. (1988). Social network analysis. *Sociology, 22*(1), 109-127.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research, 40*(2), 211-239.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois press.

Shao, J., Havlin, S., & Stanley, H. E. (2009). Dynamic opinion model and invasion percolation. *Physical review letters, 103*(1), 18701.

Shaw, D. L., & Hamm, B. J. (1997). Agendas for a public union or for private communities? How individuals are using media to reshape American society. In M. E. McCombs, D. L. Shaw & D. H. Weaver (Eds.), *Communication and democracy: Exploring the intellectual frontiers in agenda-setting theory* (pp. 209-230). Mahwah, NJ: Lawrence Erlbaum Associates.

Shoemaker, P. (1991). *Gatekeeping*. Newbury Park, CA: Sage.

Shoemaker, P., Johnson, P., Seo, H., & Wang, X. (2010). Readers as gatekeepers of online news: Russia, China and the United States. In E. Vartanova (Ed.), *Content, channels and audiences in the new millennium: Interaction and interrelations* (pp. 73-103). Moscow, Russia: Lomonosov Moscow State University.

Shoemaker, P., & Vos, T. (2009). *Gatekeeping theory*. New York: Routledge.

Shoemaker, P., Vos, T. P., & Reese, S. D. (2008). Journalists as gatekeepers. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *Handbook of Journalism Studies* (pp. 73-87). New York: Routledge.

Shumate, M. (2010). Communication power. *Journal of Communication, 60*(2), E1-E5.

Sornette, D., Desch âres, F., Gilbert, T., & Ageon, Y. (2004). Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Physical Review Letters, 93*(22), 228701. doi: 10.1103/PhysRevLett.93.228701

Sornette, D., & Helmstetter, A. (2003). Endogenous versus exogenous shocks in systems with memory. *Physica A: Statistical Mechanics and its Applications, 318*(3), 577-591.

Steeg, G. V., Ghosh, R., & Lerman, K. (2011). What stops social epidemics? *Proceedings of the 5th International Conference on Weblogs and Social Media, Spain, 5*, 377-384.

Sun, E., Rosenn, I., Marlow, C. A., & Lento, T. M. (2009). Gesundheit! Modeling contagion through Facebook news feed. *Proceedings of the Third International Conference on Weblogs and Social Media, San Jose, California, 3*, 146-153.

Sundar, S. S., Oeldorf-Hirsch, A., & Xu, Q. (2008). The bandwagon effect of collaborative filtering technology. *The proceedings of CHI '08 Extended Abstracts on Human Factors in Computing Systems*, 3453-3458.

Tewksbury, D. (2005). The seeds of audience fragmentation: Specialization in the use of online news sites. *Journal of Broadcasting & Electronic Media, 49*(3), 332-348.

Tonkin, E., Pfeiffer, H. D., & Tourte, G. (2012). Twitter, information sharing and the London riots? *Bulletin of the American Society for Information Science and Technology, 38*(2), 49-57.

Treisman, A., & Fearnley, S. (1969). The Stroop test: Selective attention to colours and words. *Nature, 222*, 437-439.

Treisman, A. M. (1964). Selective attention in man. *British Medical Bulletin, 20*(1), 12-16.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402-418.

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences, 109*(16), 5962-5966.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K. I., Kondor, I., & Barabási, A. L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E, 73*(3), 036127.

Valente, T. W. (1993). Diffusion of innovations and policy decision-making. *Journal of Communication, 43*(1), 30-45. doi: 10.1111/j.1460-2466.1993.tb01247.x

Valente, T. W. (1995). *Network models of the diffusion of innovations*. Cresskill, New Jersey: Hampton press.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks, 18*(1), 69-89.

Van den Bulte, C., & Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology, 106*(5), 1409-1435.

VanderWeele, T. J. (2011). Sensitivity analysis for contagion effects in social networks. *Sociological Methods & Research, 40*(2), 240-255.

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How nonprofit organizations are using Facebook. *Public Relations Review, 35*(2), 102-106.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences, 99*(9), 5766-5771.

Watts, D. J. (2004). The "new" science of networks. *Annual review of sociology*, *30*(3), 243-270.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research, 34*(4), 441-458.

Weber, J. (1956). Fluctuation dissipation theorem. *Physical review*, *101*(6), 1620-1626. doi: 10.1103/PhysRev.101.1620

Webster, J. G. (1985). Program audience duplication: A study of television inheritance effects. *Journal of Broadcasting & Electronic Media*, *29*(2), 121-133.

Webster, J. G. (2005). Beneath the veneer of fragmentation: Television audience polarization in a multichannel world. *Journal of Communication, 55*(2), 366-382.

Webster, J. G. (2011). The duality of media: A structurational theory of public attention. *Communication Theory, 21*(1), 43-66.

Webster, J. G., & Ksiazek, T. B. (2012). The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication, 62*(1), 39-56. doi: 10.1111/j.1460-2466.2011.01616.x

Weiss, A. (2005). The power of collective intelligence. *Networker, 9*(3), 16-23.

Wellman, B. (2001). Physical place and cyberplace: The rise of personalized networking. *International journal of urban and regional research, 25*(2), 227-252.

Wellman, B. (2002). Little boxes, glocalization, and networked individualism. In M. Tanabe, P. v. d. Besselaar & T. Ishida (Eds.), *Digital cities II: Computational and sociological approaches* (pp. 10-25). Berlin Heidelberg: Springer.

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports, 3*, 2522.

Westley, B. H., & MacLean, M. S. (1957). A conceptual model for communications research. *Journalism & Mass Communication Quarterly, 34*(1), 31-38.

White, D. M. (1950). The "gate keeper": A case study in the selection of news. *Journalism Quarterly, 27*, 383-391.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686-688.

Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences, 104*(45), 17599-17601.

Wu, F., Huberman, B. A., Adamic, L. A., & Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical Mechanics and its Applications, 337*(1), 327-335.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on Twitter. *Proceedings of the 20th international conference on World Wide Web, India, 20*, 705-714.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In I. King, W. Nejdl & H. Li (Eds.), *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 177-186). Hong Kong, China: ACM.

Yu, L., Asur, S., & Huberman, B. A. (2011). *What trends in Chinese social media*. Paper presented at the 5th SNA-KDD Workshop San Diego, CA USA.

Zhang, J., Liu, B., Tang, J., Chen, T., & Li, J. (2013, August 3-9). *Social influence locality for modeling Retweeting behaviors.* Paper presented at the 23rd International Joint Conference on Artificial Intelligence, Beijing, China.

Zhu, J. H. (1992). Issue competition and attention distraction: A zero-sum theory of agenda-setting. *Journalism Quarterly, 69*(4), 825-836.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Addison-Wesley.

Table 1. Comparison of Diffusion Theories

	Two-step flow	News diffusion	Diffusion of innovations
Discipline	Communication	Communication	Communication
Driving forces	Interpersonal effects vs. Media effects	Interpersonal effects vs. Media effects	Personal attributes, system features, innovation characteristic, and diffusion processes.
Patterns/features	Two-step flow	J-curve	S-shaped curve
Implications	Diffusion network	Salience of content	Categories of innovators
Limitation	simplistic, linear, source-dominated	Fails to consider utility or value of information for receivers	linear, source-dominated, underestimates power of media

Table 2. Comparison of Diffusion Models

	Bass diffusion model	Epidemic model	Threshold model
Discipline	Marketing	Epidemiology	Sociology
Driving forces	Innovativeness vs. imitation	Physical contacts	Interpersonal effects
Patterns/features	Differential equations of Bass diffusion	Differential equations of SI, SIS, and SIR, etc.	Differential equations of utility-cost assumption
Implications	Mathematical modeling	Mathematical modeling	Mathematical modeling, and Threshold
Limitation	Presumes an accurate knowledge of the market size; driven by the initial set of assumptions.	Single source; driven by the initial set of assumptions.	It's difficult to measure threshold.

Table 3. Coding Scheme of Categories of Sina Weibo Tweets

Categories	**Example**
Fun	一女生跟男朋友分手了，旁边她同学安慰她：“那男的有什么好，土木工程的，一听就知道又“土”又“木”的！”旁边软件工程的男同学一听，心都凉了半截。
Life encyclopedia	【脸部各部位长痘痘的原因及祛痘方法】各位姐妹们，还在为脸上长痘苦恼吗？为你们奉上祛痘方法，让我们一起做清新小美人吧
Star	#艾玛沃特森##Emma Watson#3 月 30 日 Emma 和朋友在美国洛杉矶好莱坞.http://t.cn/zOSr3DY
Movie	怡！！！簡直笑爆肚啊！[哈哈] [哈哈] http://t.cn/zOKU4cS 哩部系超级好睇！[good]'
News	【记者暗访：皮革废料制药囊熬制时铬大量残留】央视昨天报道了浙江省新昌县一些胶囊厂非法使用工业明胶生产药用胶囊，这种工业明胶被检出重金属铬超标，所用原料是一种所谓“蓝矾皮”的工业皮革废料。http://t.cn/zO0PVkS
Mood	是的我爱你，但是我也爱烤鸡翅，所以你以为你是谁?
Hobby	#游戏推荐#【小小帝国 Little Empire】开发过 Shoot U 的 Camel Games 公司非常低调但是出品的游戏却非常有实力比如这款《小小帝国》是世界首款应用了 3D 和真实地理位置技术结合的策略类网络游戏. http://t.cn/St43LA
Product	校园用什么 3G 手机促销？普通低价的普通 3G 手机？以中兴 V880 为首的智能机？一经比较，优劣自现，一经传播，口碑即来。#相信沃没错的##极速互联随我行#
Opinion	建议全国上访户，向上海大火案灾民先学习，春节期间将上访改为向各界领导拜年。

Table 4**.** Summary of Linear Regression

	Model 1 Diffusion size >= 100			**Model 2** Diffusion size < 100			**Model 3** All diffusion sizes		
	Beta	SD		Beta	SD		Beta	SD	
Individual attributes									
Verified identity	-0.09	0.03	***	0	0.01		0.08	0.03	**
Number of followers	-0.02	0.01		0.18	0	***	-0.03	0.01	**
Number of tweets (Log)	0.14	0.01	***	0.07	0.01	***	0.11	0.01	**
Diffusion channels									
Mean threshold (Log)	0.06	1.1	***	-0.26	0.04	***	-0.16	0.65	**
Zero-threshold ratio	0.07	0.09	***	0.28	0.05	***	0.07	0.1	**
Diffusion features									
Comments (Log)	0.53	0.01	***	0.18	0.01	***	0.12	0.01	**
lifetime	0.39	0.02	***	0.32	0.01	***	0.39	0.02	**
Diffusion depth	0.14	0.01	***	0.25	0.01	***	0.13	0.01	**
Diffusion depth * Mean threshold (Log)	--	--		--	--		-0.01	0.31	**
Information categories									
Fun	0.03	0.06		-0.01	0.02		0.32	0.05	**
Star	0.05	0.07	***	0.1	0.03	***	0.05	0.06	**
Movie	0.02	0.07		0.01	0.03		0.1	0.07	**
Mood	0.03	0.07	.	-0.04	0.02	.	0.04	0.06	
Hobby	0.02	0.07		0.01	0.03		0	0.07	**
Product	0.04	0.08	**	-0.02	0.03		0.03	0.08	**
News	0.03	0.07	.	0.03	0.02		0.04	0.06	**
Life encyclopedia	0.06	0.06	**	-0.01	0.02		0.03	0.06	**
Information features									
picture	0	0.04		0.02	0.01		0.06	0.04	*
video	-0.03	0.05	*	0.02	0.03		0.01	0.05	
url	0.03	0.04	*	0.01	0.02		-0.01	0.04	
conversation	-0.01	0.03		0	0.01		0.01	0.03	
emotion	-0.01	0.03		0.01	0.01		-0.01	0.03	.
Adjusted R-square	80.3%			78.6%			91.8%		
N	2500			1000			3500		

Notes: the control group of the categories of information is the category of opinion expression.

Dependent variable is the size of diffusion.

Table 5. Collective Gatekeeping and Interpersonal Effects

	Collective Gatekeeping	**Interpersonal Effects**
Process	• Collaborative filtering • Selection of trending news • News aggregation	• Building ego network • News learning from online friends • Decision making of news sharing
Characteristics	• Empowered authority • Collective intelligence • Bottom up & top down • Anonymity • Efficiency • Temporary	• Personalized preference • Local wisdom • Bottom up • Virtual friends • Slow • Long-term

Table 6 OLS regression of interpersonal effect, collective gatekeeping, and news category on the size of news diffusion

	Size of News Diffusion (Log)		
	Beta	SE	
Number of views	0.47	0.00	***
Collective gatekeeping			
Zero-threshold ratio	0.86	0.08	***
Promotion time	-0.03	0.00	***
Interpersonal effect			
Average threshold	0.41	0.54	***
Submitter's fans	0.02	0.00	**
News Category			
Gaming News	0.00	0.03	Ns
Lifestyle news	-0.02	0.02	**
Offbeat news	-0.01	0.02	Ns
Science news	-0.01	0.02	Ns
Sports news	-0.02	0.02	**
Technology news	0.03	0.02	**
World & Business news	0.07	0.02	***
(Intercept)	-----	0.07	***
R Square (%)	84.8***		

Note: Sample size N = 3553. The baseline group of news category is entertainment. The dependent variable is the logarithm form of the size of news diffusion (i.e. how widely the news diffuses). * P < 0.05. ** P < 0.01. *** P< 0.001.

Table 7. Temporal Property of Different Impacts

Levels of impacts	Property of temporality	Different influential factors
Exogenous shocks	Strong instantaneous	1. External shocks (e.g., news events of earthquakes) 2. Recommendations
Endogenous perturbations	Cumulative	Social influence
Audiences interest	Stable	1. Searching 2. Ratings, comments, and favorites
Novelty of media products	Decay	1. Popularity of videos (views & lifetime) 2. Categories of videos (news videos vs. copyright-protected videos)

Table 8. Prediction of Burst of Public Attention Using OLS Regression

	Bursts of public attention		
	Beta	b	SE
Popularity of videos			
Duration of lifetime (log)	-.60***	-.33	.004
Number of views (log)	-.20***	-.07	.005
Audience interest			
Comments (log)	.02	.01	.004
Favorites (log)	-.36***	-.12	.004
Rating(log)	.25***	.08	.005
Diffusion channels of videos			
System recommendation (log)	.14***	.04	.003
Search (log)	-.08***	-.02	.001
Social influence (log)	-.02*	-.01	.002
Mobile (log)	.04***	.01	.002
Category of videos **(baseline = News)**			
Copyright-protected	-.17***	-.35	.024
Hobby	-.16***	-.31	.023
Broadcast	-.13***	-.34	.026
Education	-.07***	-.30	.030
Entertainment	-.06***	-.27	.031
How-to	-.08***	-.30	.029
Ngo	-.03***	-.12	.030
Tech	-.08***	-.29	.029
Constant		.53	.036
R Square	56.6%		

Notes: * $P < .05$. ** $P < .01$. *** $P < .001$. N= 14978. Bursts of public attention is measured by the peak fraction (*F*). The variance inflation factors (VIF) which are much smaller than 10 indicate that there is no presence of multicollinearity.

Table 9. Data Sets Used in This Dissertation

Data	Weibo 2012	Digg 2009	YouTube 2011
Category	Microblog	Social news website	Video sharing website
Diffusion mechanism	Interpersonal effects	1. Collective gatekeeping 2. Interpersonal effects	1. Recommendation 2. Searching 3. Interpersonal effects
Popularity	Log-normal and Power law (max size =422500 retweets)	Log-normal (max size =47260 views; 24100 diggs)	Log-normal (max size = 175700000 views)
Burst	Strong (M = 0.8)	Strong (M = 0.7)	Modest (M = 0.3)

Figure 1. Structure of the Dissertation

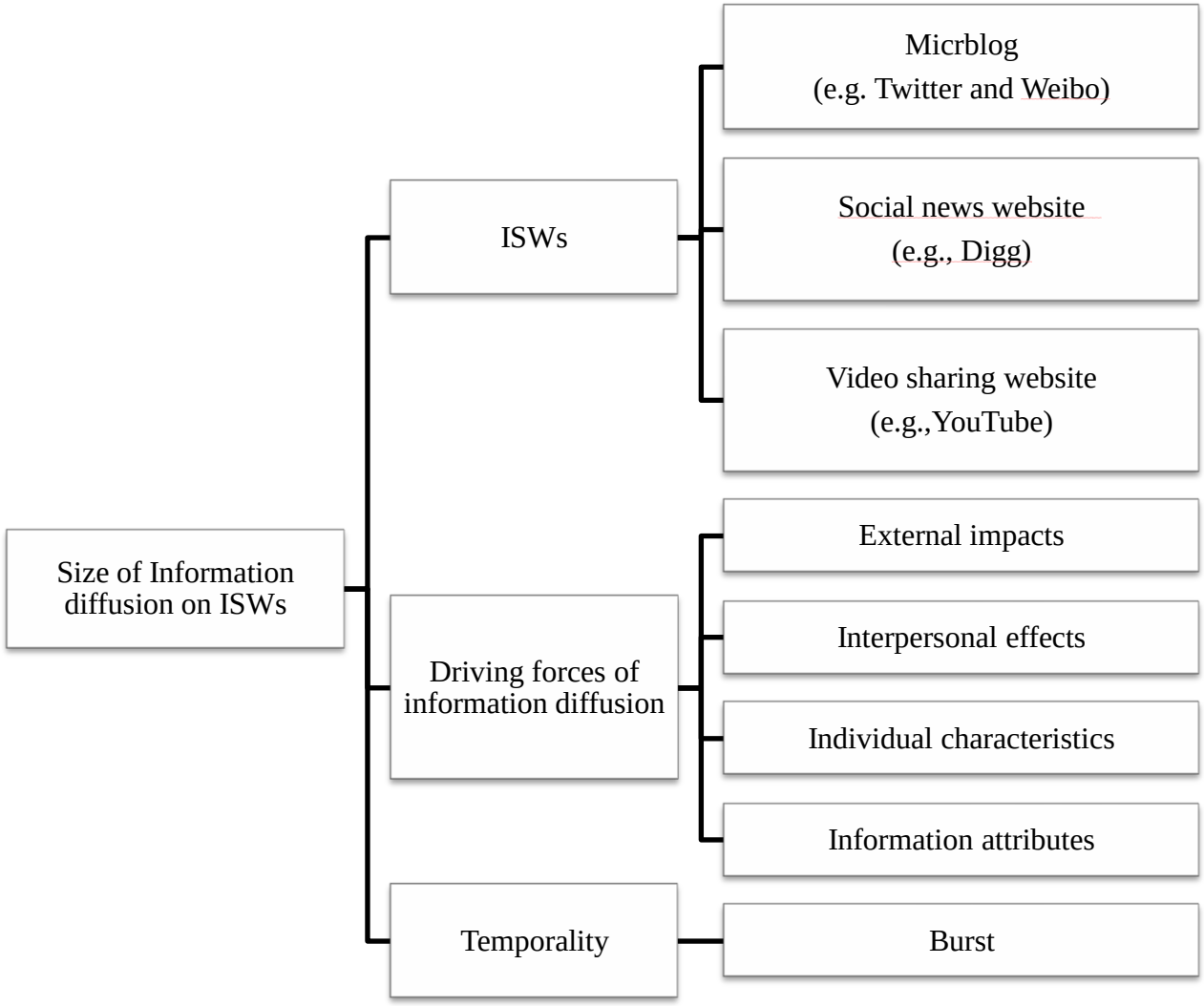
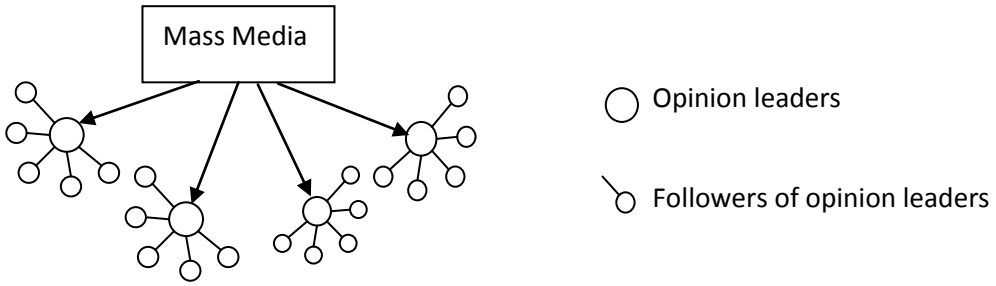
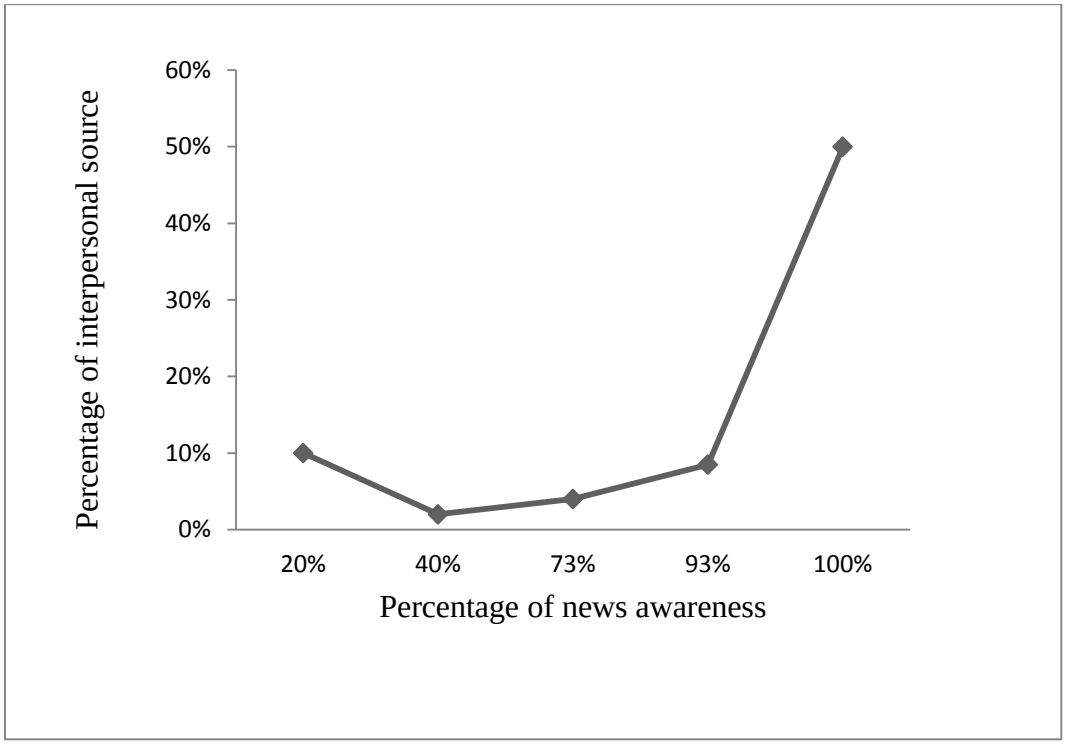


Figure 2. Two-step Flow Model



Source: (Katz, 1957)

Figure 3. J-curve Model of News Diffusion



Source: (Greenberg, 1964b)

Figure 4. Growth, Decline, and Revival of News Diffusion Research

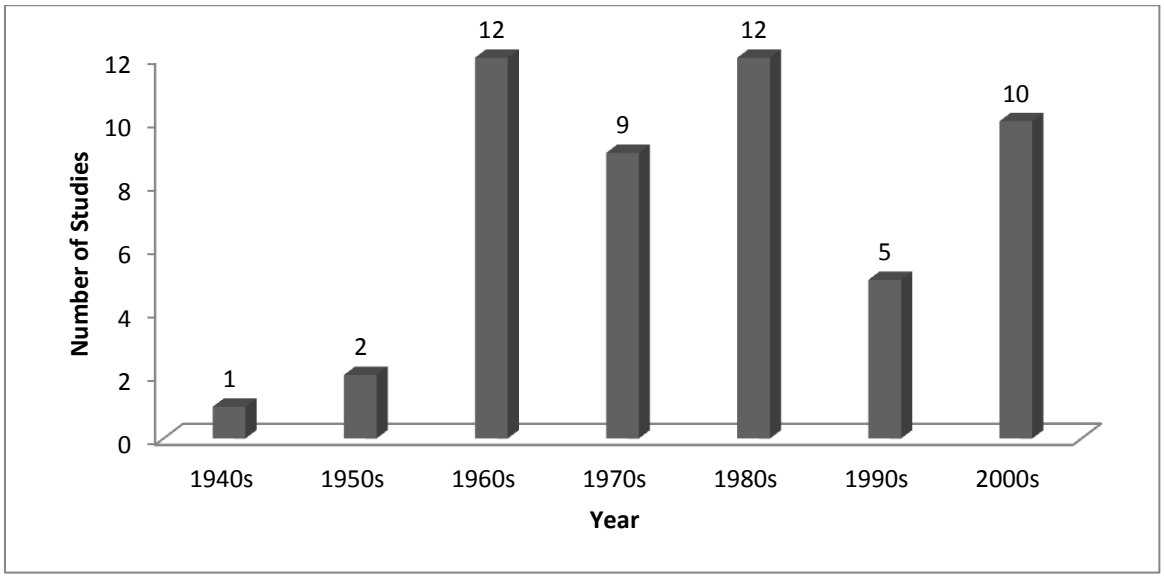
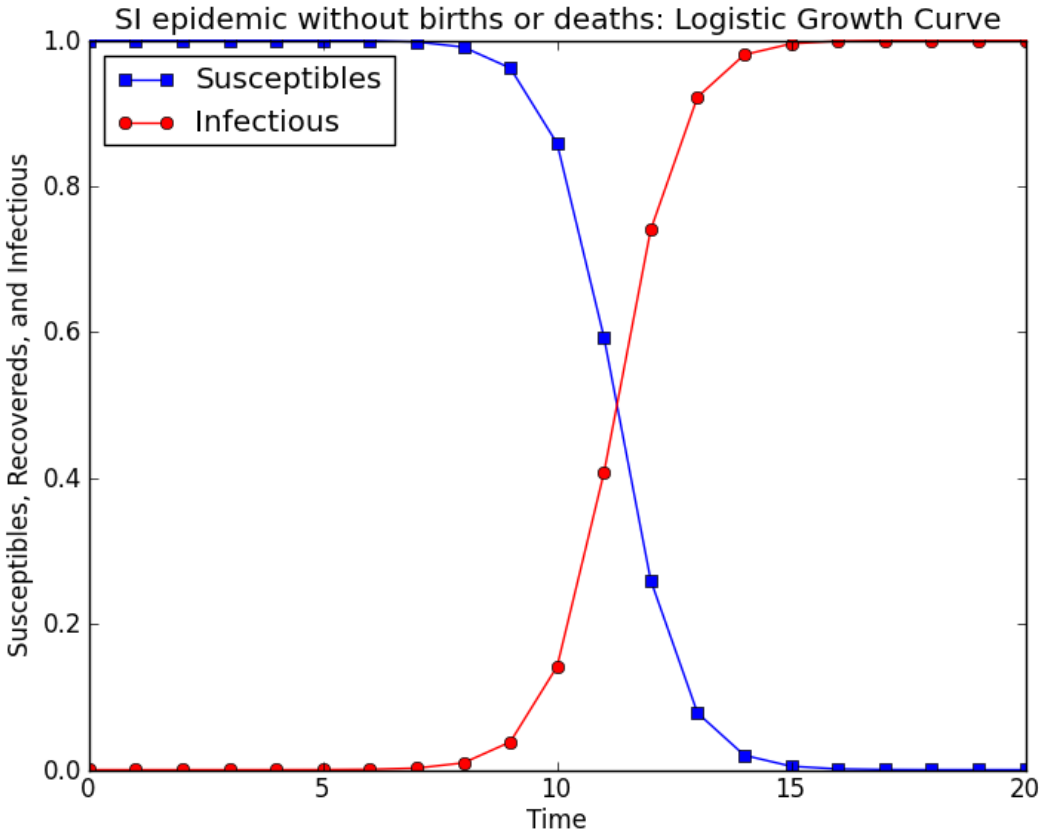
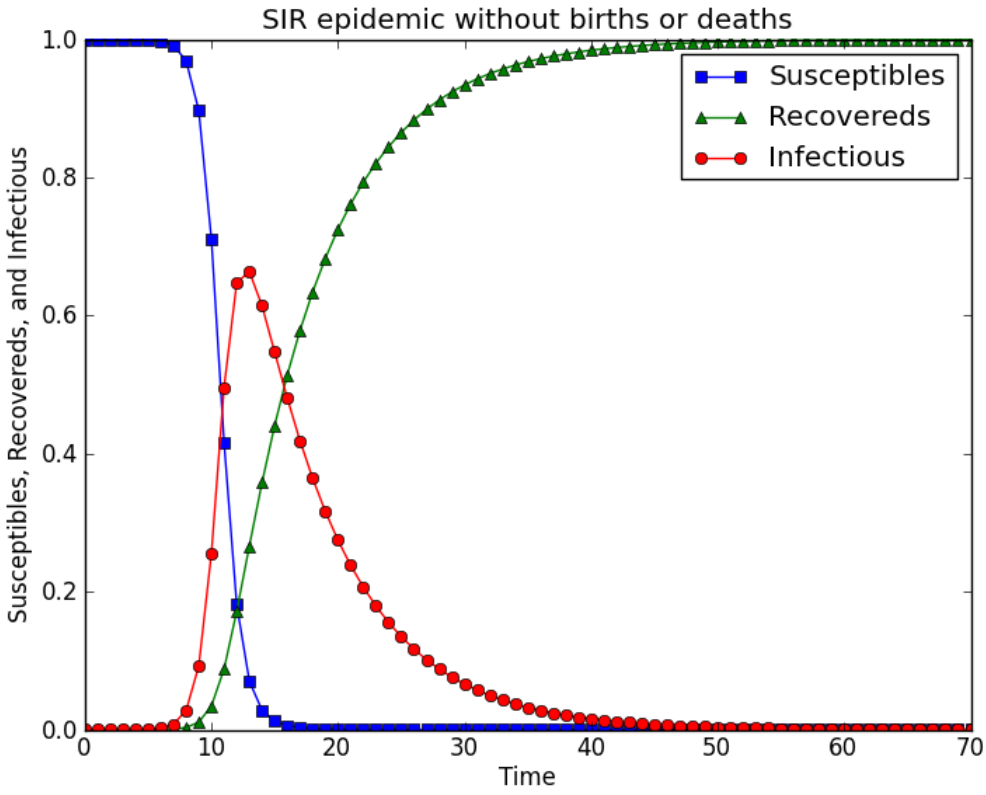


Figure 5. A SI Epidemic Model: The Logistic Growth Curve



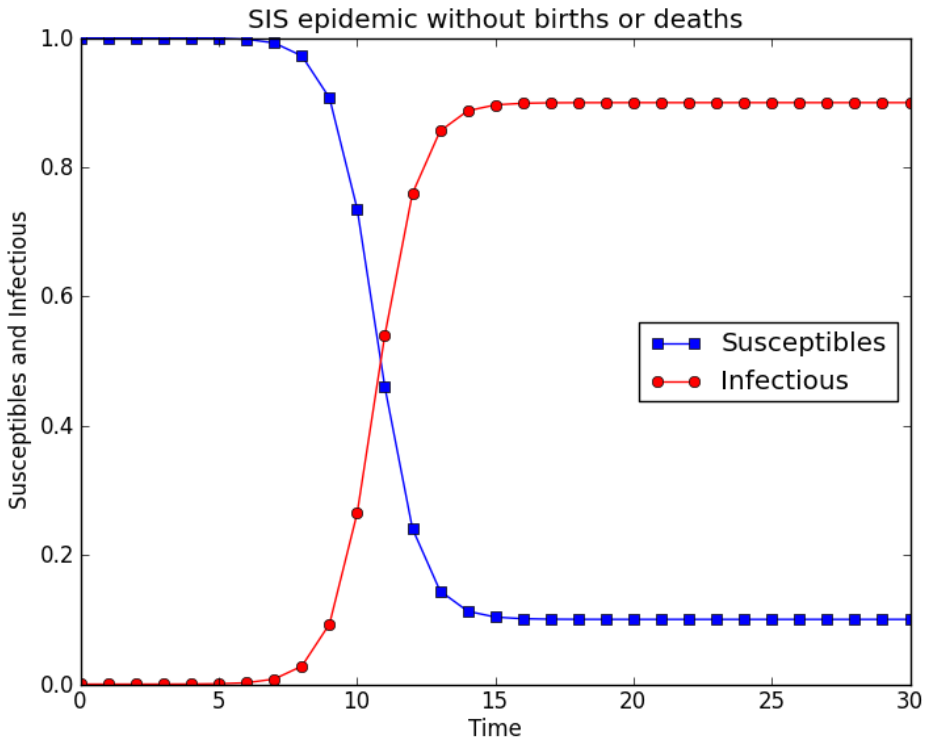
Note: the two curves in the figure show the fraction of the population in the states of susceptible, infected as a function of time. The parameters are $\beta = 1.4247$, S_0=1-1e-7, x_0=1e-7.

Figure 6. A SIR Epidemic Model



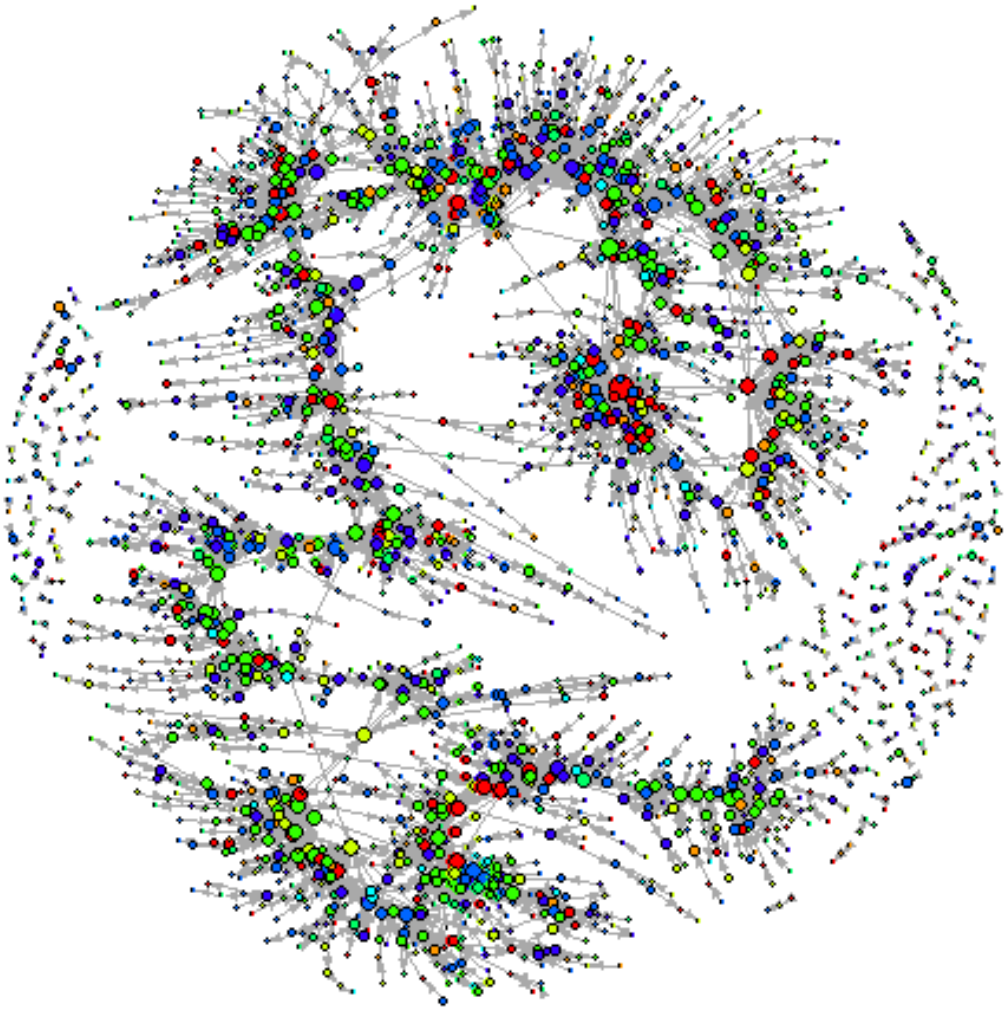
Notes: the three curves show the fraction of the population in the states of susceptible, infected, and recovered over time. The parameters are $\beta = 1.4247$, $\gamma = 0.14286$, I_0=1-1e-6, x_0=1e-6.

Figure 7. A SIS Epidemic Model



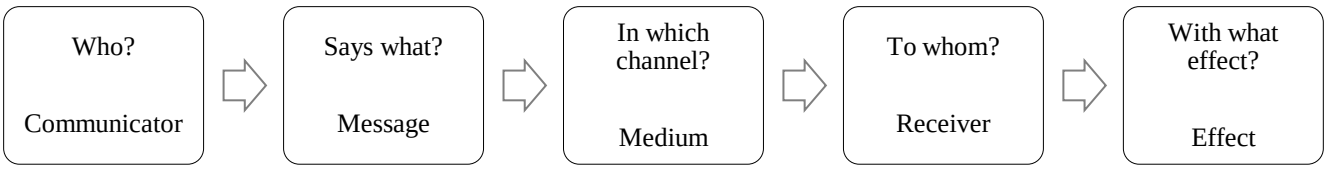
Note: the two curves in the figure show the fraction of the population in the states of susceptible, infected as a function of time. The parameters are $\beta = 1.4247$, $\gamma = 0.14286$, I_0=1-1e-7, x_0=1e-7.

Figure 8. Clickstream Network of 3553 Digg News Stories in June 2009



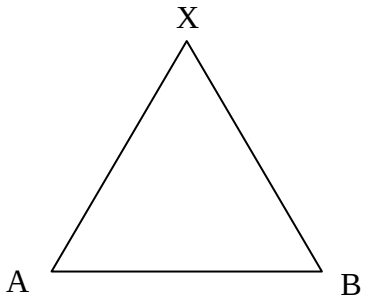
Notes: the nodes denote news stories, and the links denote the flow of public attention between the nodes of news stories. The size of a node denotes the popularity of the news story. The linear skeleton of the clickstream network indicates that public attention flows in parallel with the time.

Figure 9. The Lasswell formula



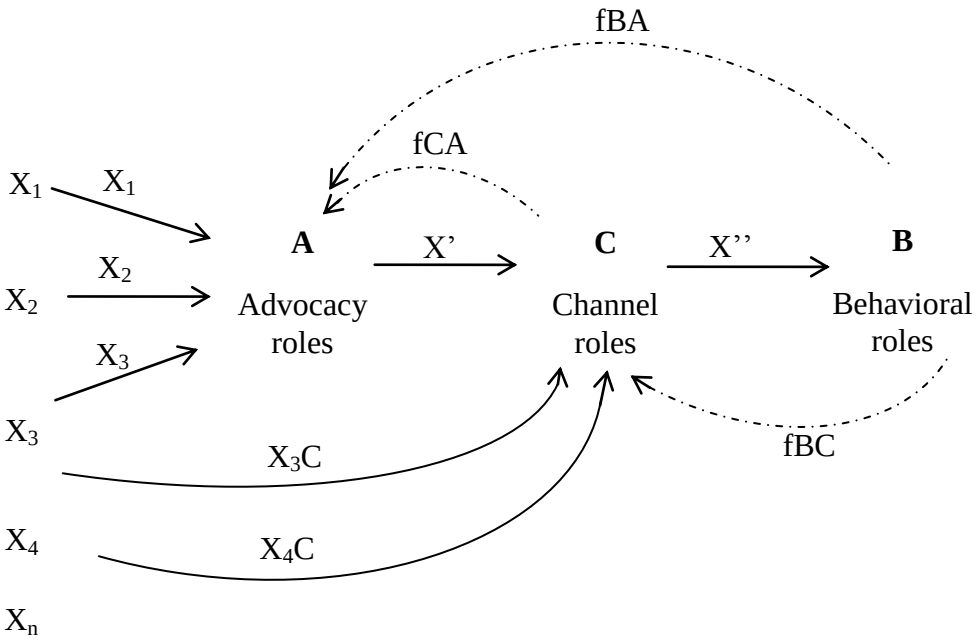
Source: (Lasswell, 1948)

Figure 10. The Newcomb ABX model



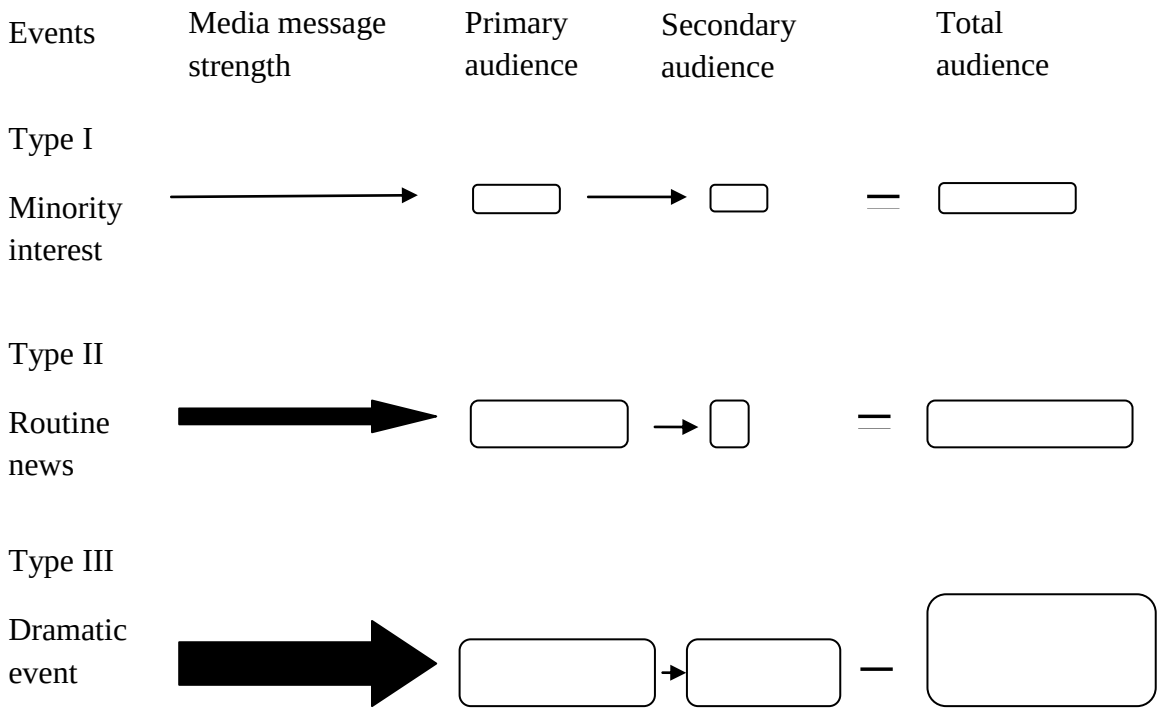
Source: (Newcomb, 1953)

Figure 11. Westley and Maclean's Conceptual Model of Mass Communication



Source: (Westley & MacLean, 1957)

Figure 12. Diffusion Model Underlying the J-Curve



Source: (MaQuail & Windahl, 1993)

Figure 13. Webpage of Information Aggregator on YouTube

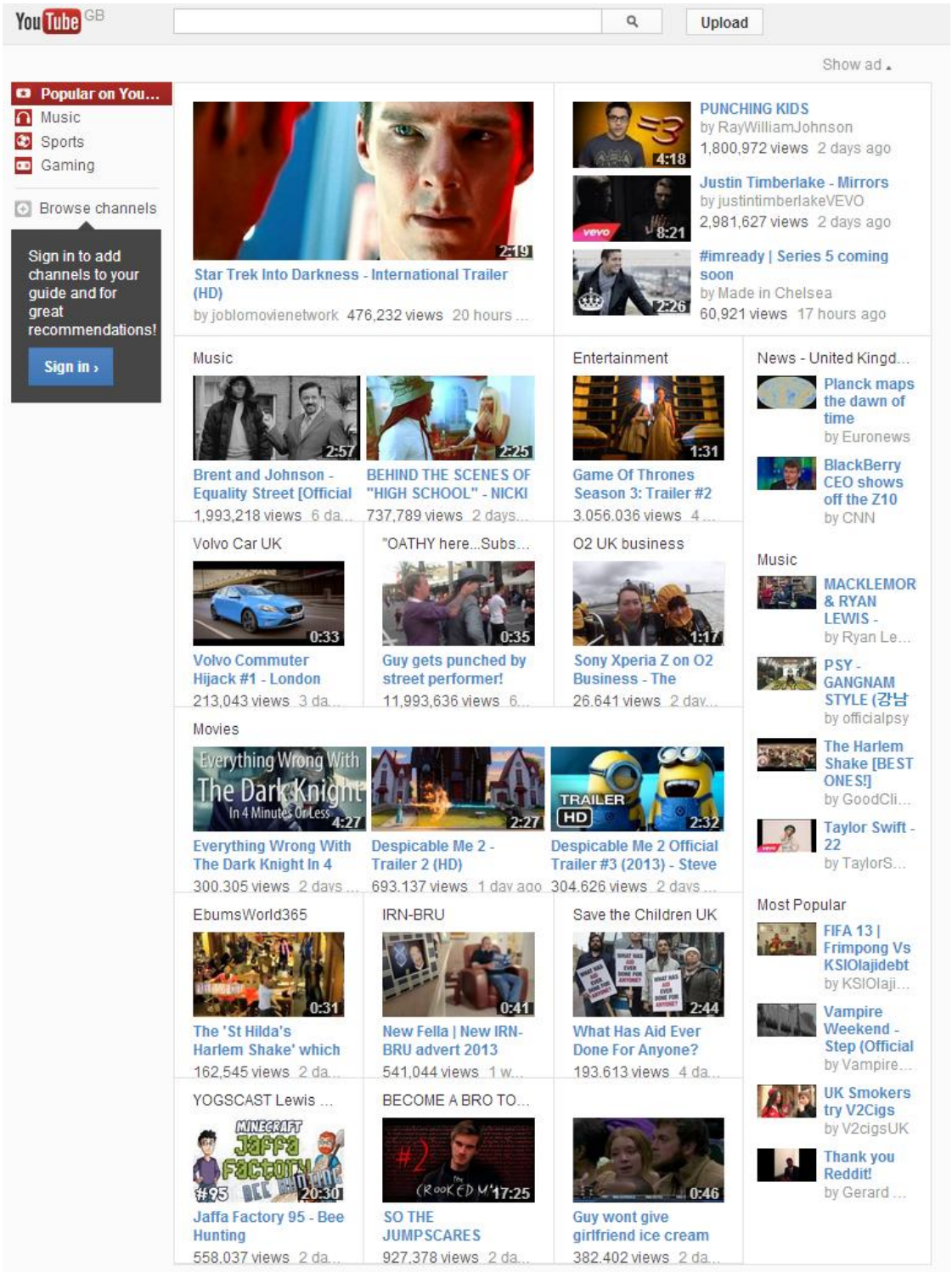
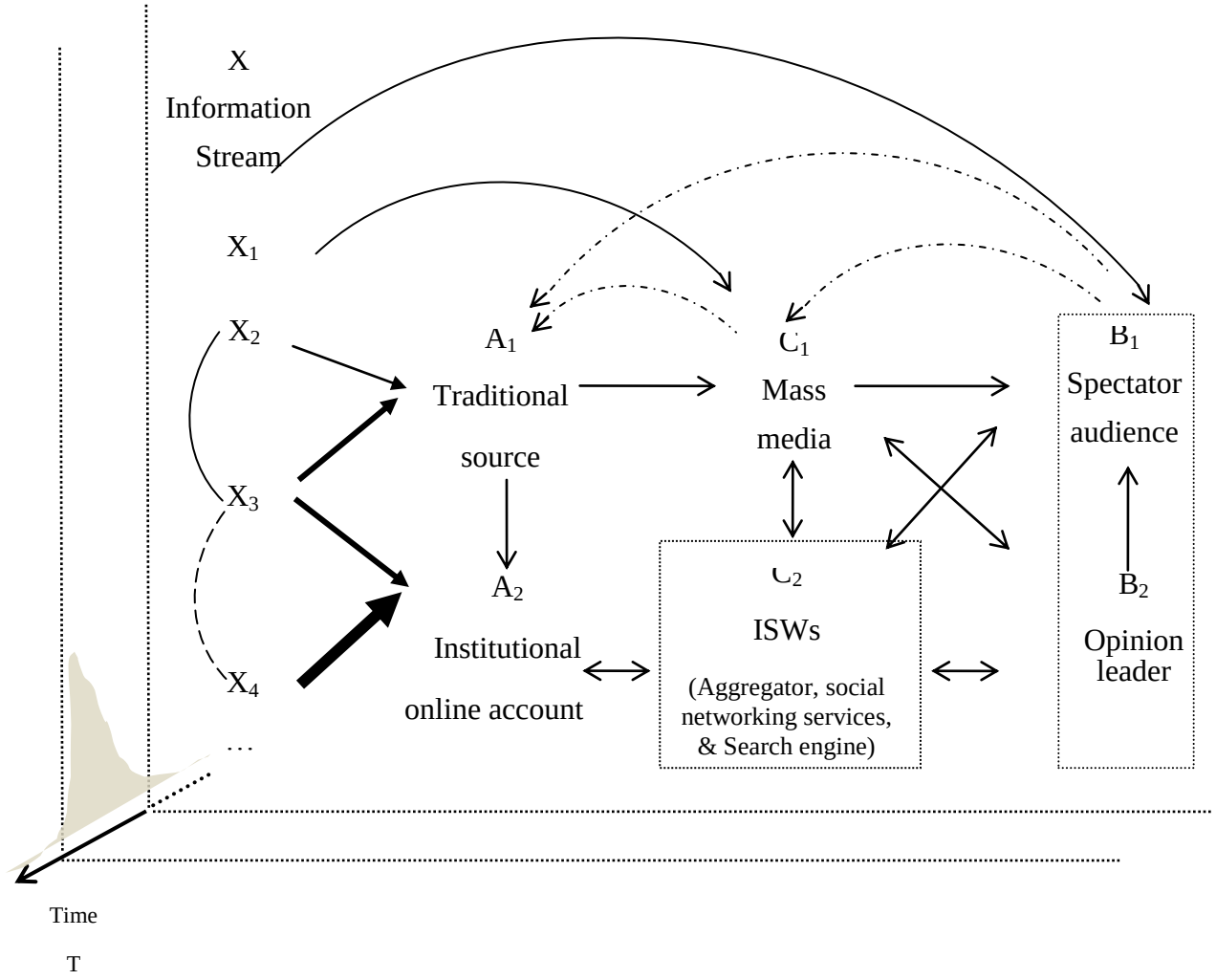
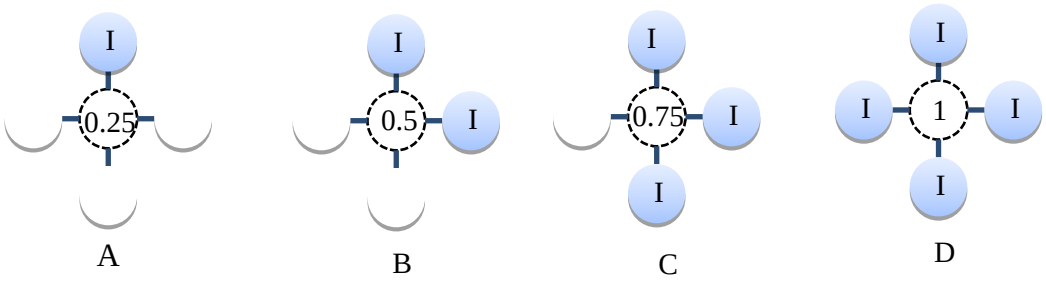


Figure 14. The ABXCT Model of Information Diffusion



Notes: In addition to information (I), advocacy roles (A), communication channels (C), and behavioral roles (B) (e.g., audiences), this model includes the time (T). The solid line between X_2 and X_3 and X_4 denotes the competition relationship, and the dash line between X3 and X4 denotes the mutually beneficial relationship.

Figure 15. The Duality of Network Threshold in Information Diffusion



Notes: the central node with the dashed outline is the ego at the critical point (e.g. when the ego decides to spread the information). Threshold is featured by its duality. A larger threshold suggests a larger local pressure as well as a larger obstacle or cost. In case A, one friend out of four has already engaged in diffusing the information (denoted with "I"), thus its threshold is 0.25. Similarly, the threshold for situation B, C, and D is 0.5, 0.75, and 1, respectively. Therefore, network threshold measures the extent to which individual actors' personal network (e.g. local network) have been activated.

Figure 16. Distribution of the Size of Information Diffusion on Sina Weibo (N= 600’000)

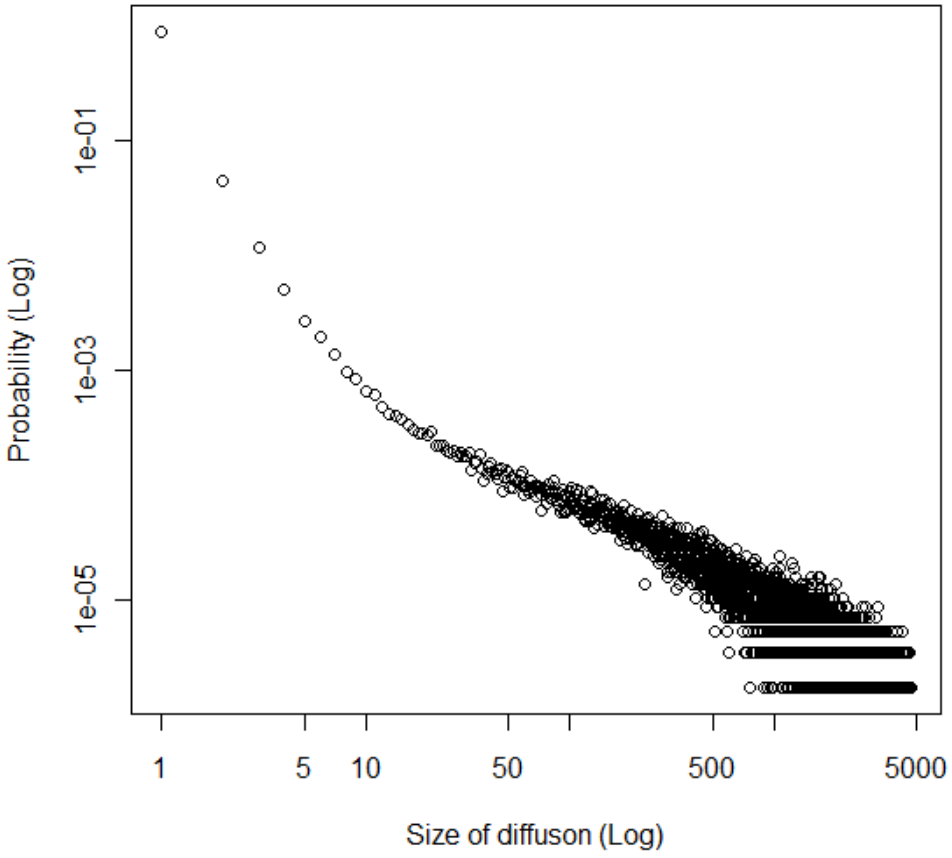
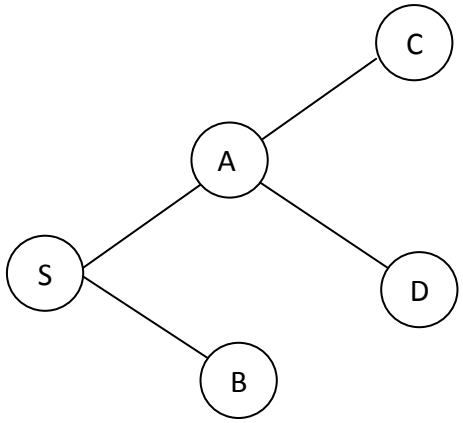


Figure 17. Depth of Diffusion Network



Notes: S denotes information submitter; A, B, C, and D are four diffusers, respectively. Diffusion depth is the largest distance from diffusers to the information submitter. The distance from S to A, B, C, and D is 1, 1, 2, and 2, respectively. The overall depth of diffusion is the maximum value of the diffusion steps for all of the nodes, and in this case, it is 2.

Figure 18. Categories of Information on Sina Weibo

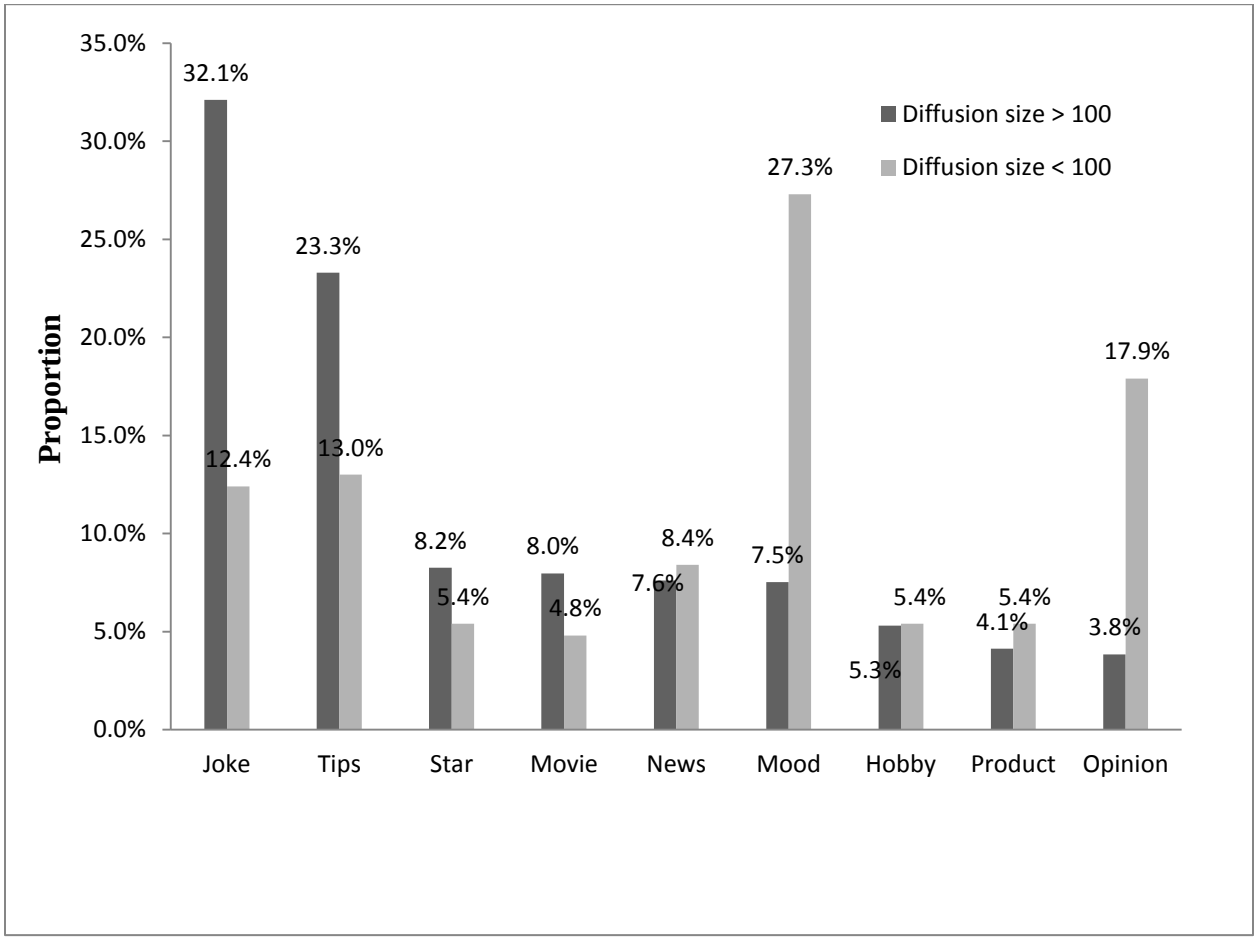


Figure 19. The Nonlinear Relationship between Threshold and Diffusion Size

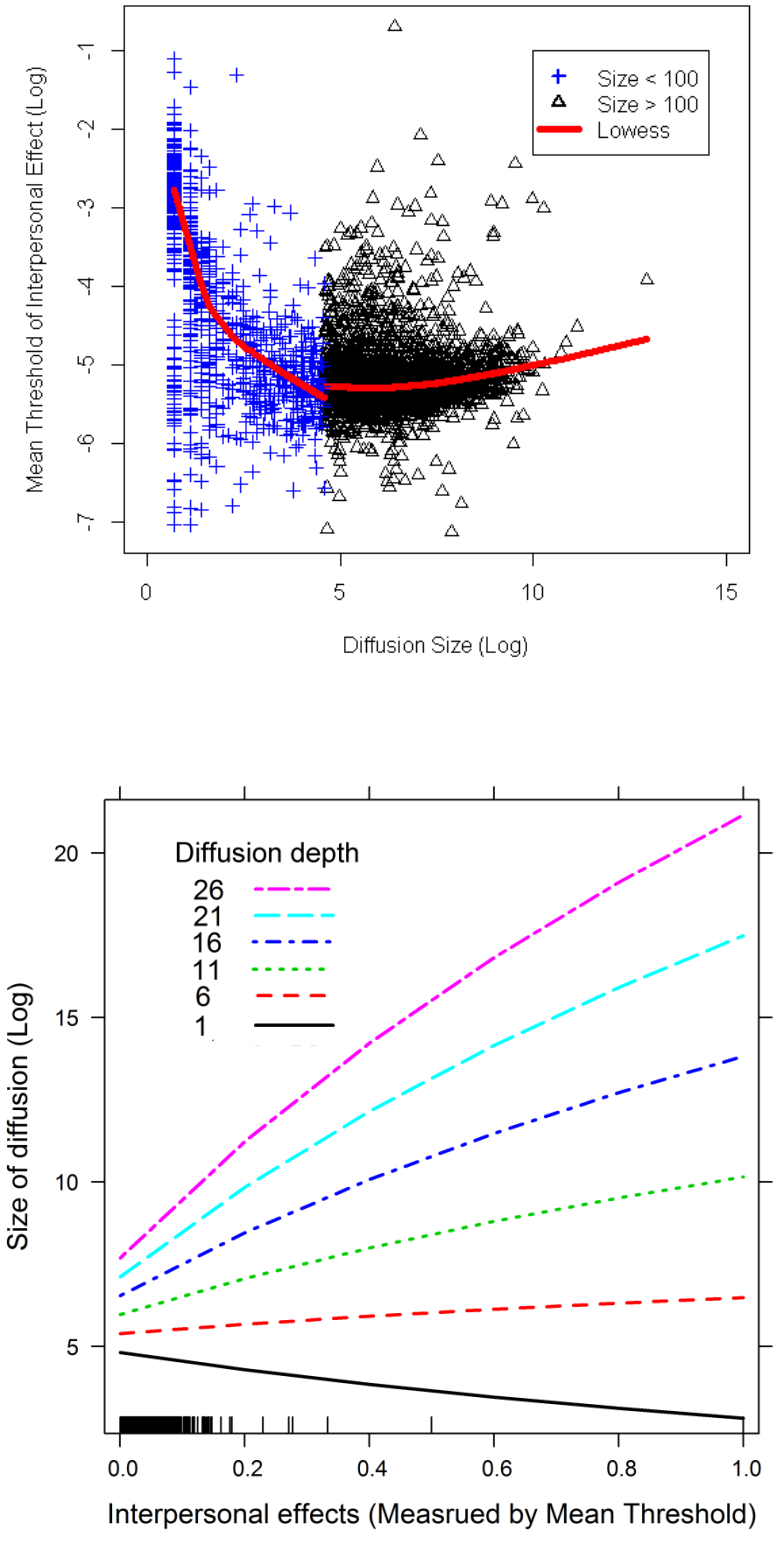


Figure 20. The Scaling Relationship between Comments and Diffusion Size

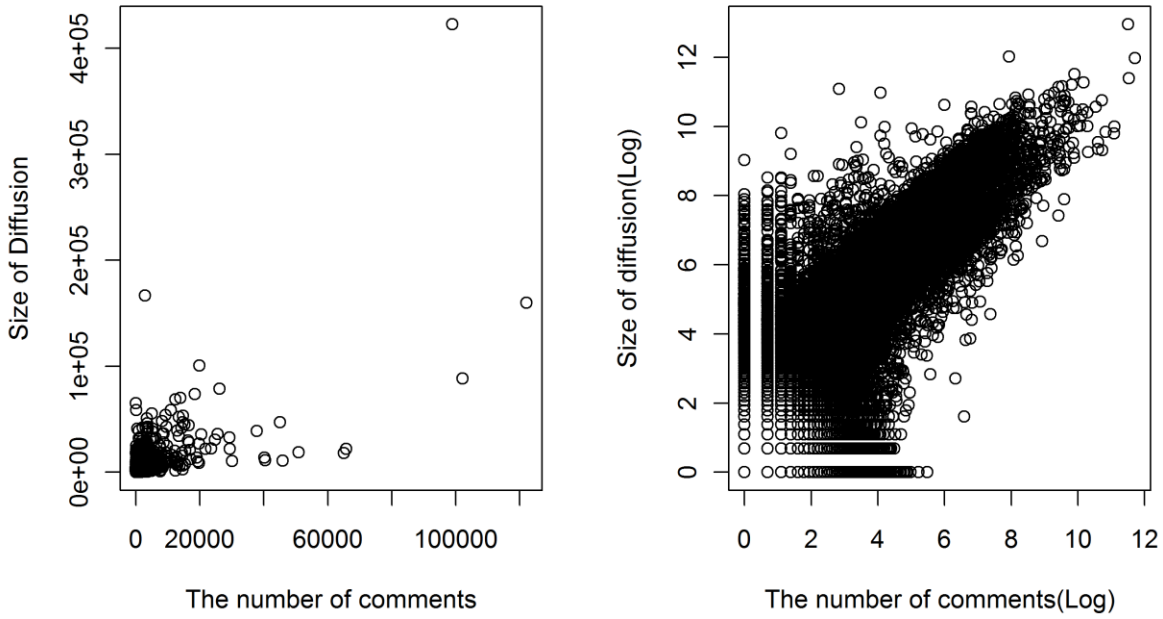


Figure 21. Diffusion Depth and Diffusion Size

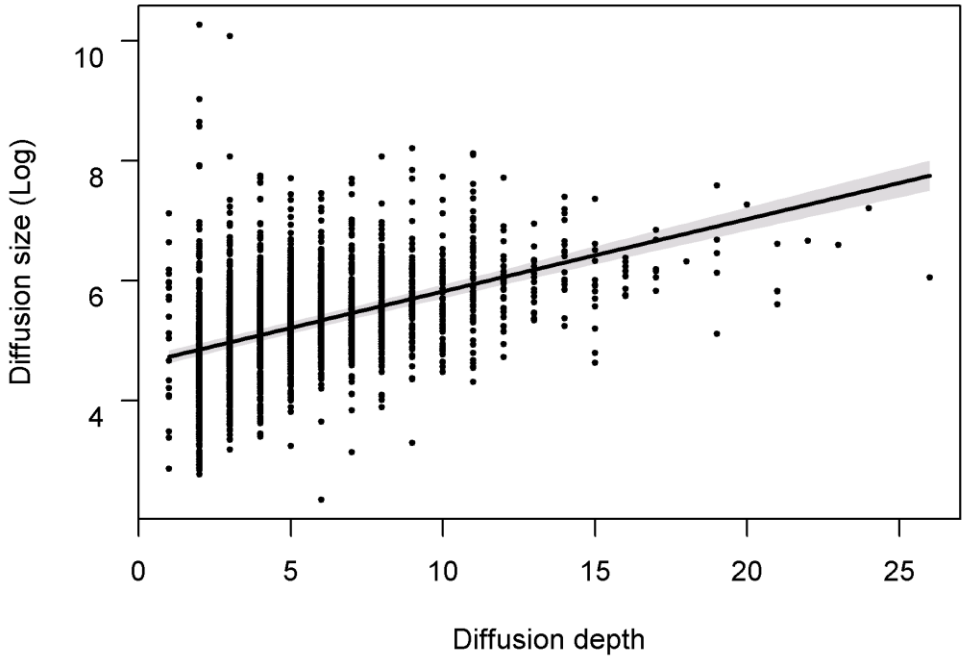


Figure 22. The Evolution of Information Diffusion on Sina Weibo over Time

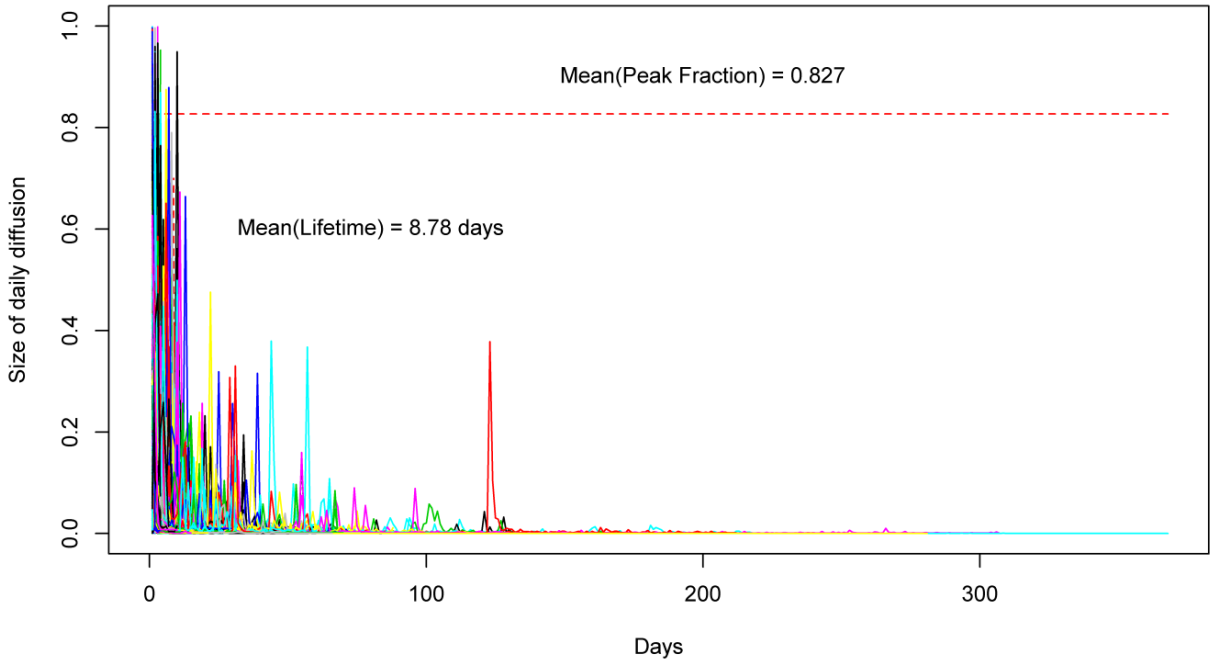


Figure 23. The Rise and Fall of Information Diffusion on Sina Weibo during September 2010-May 2011 (N= 100)

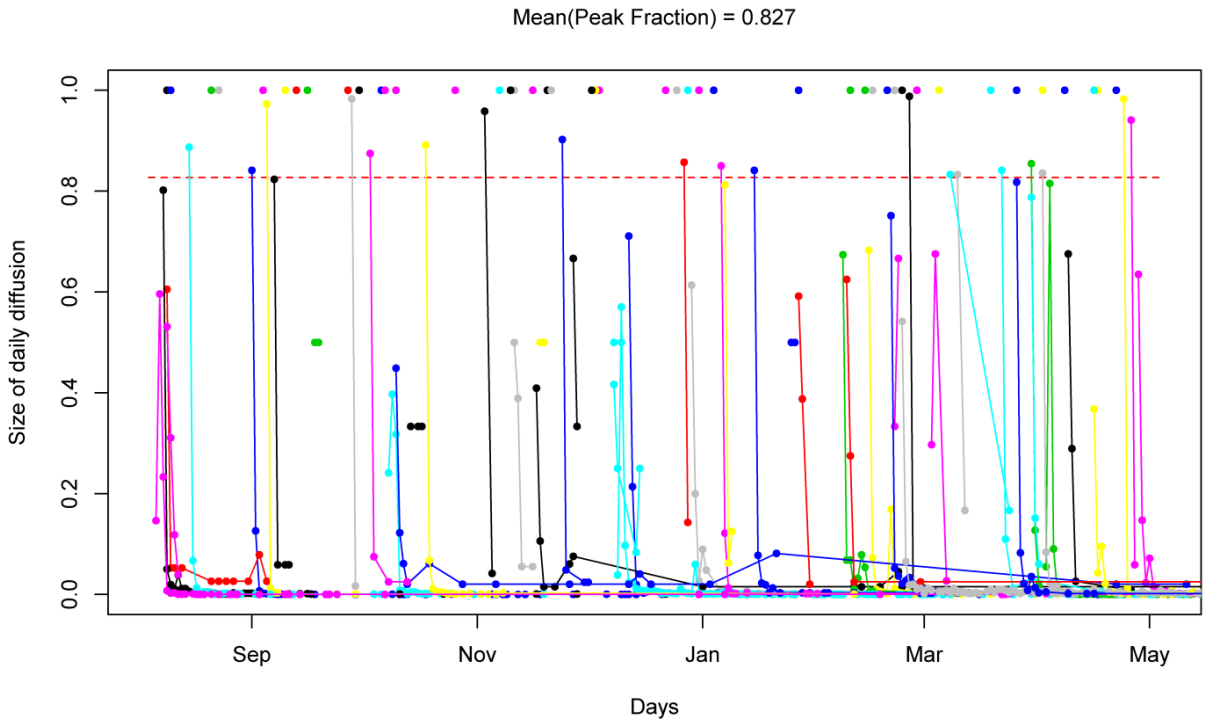
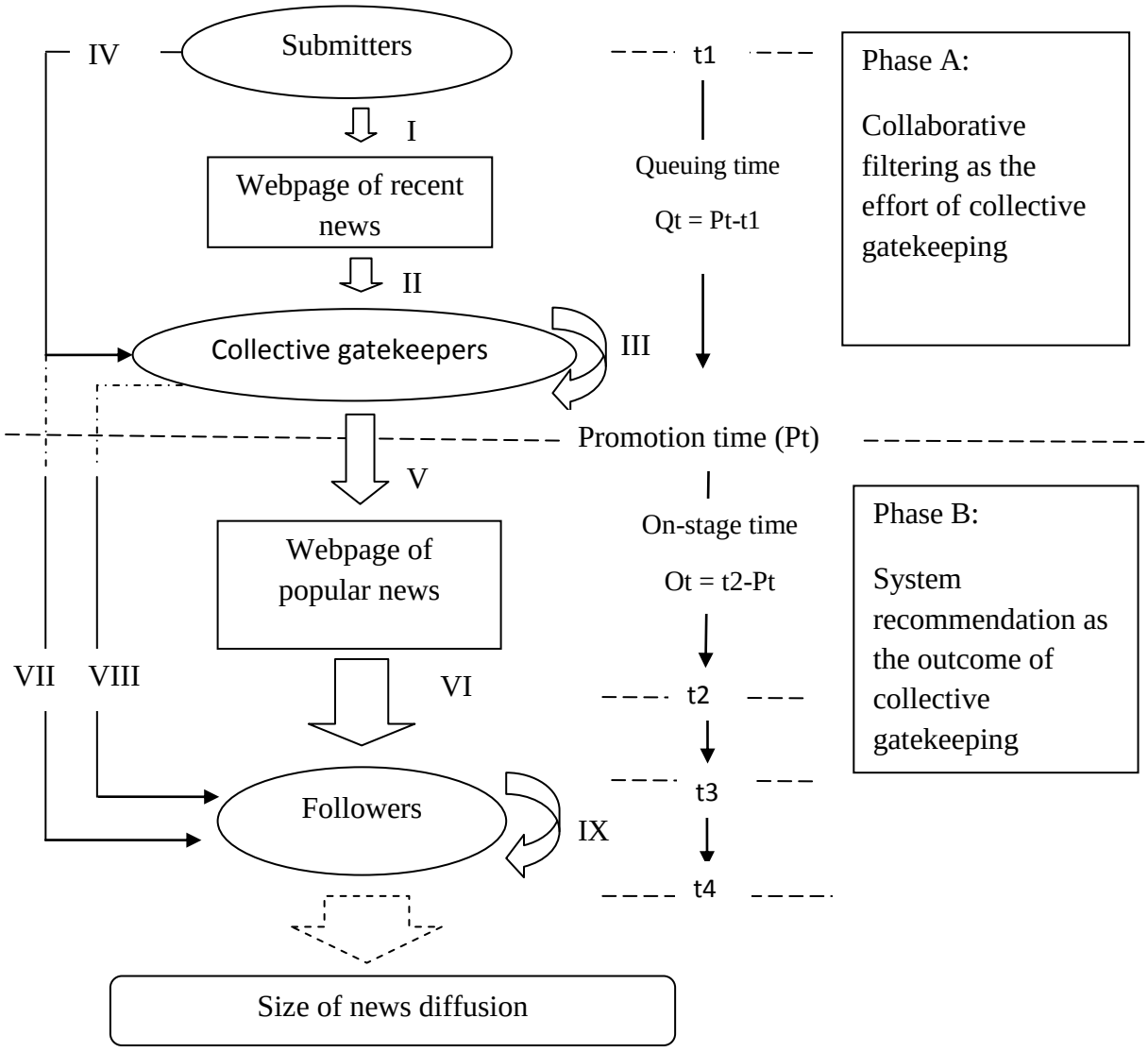
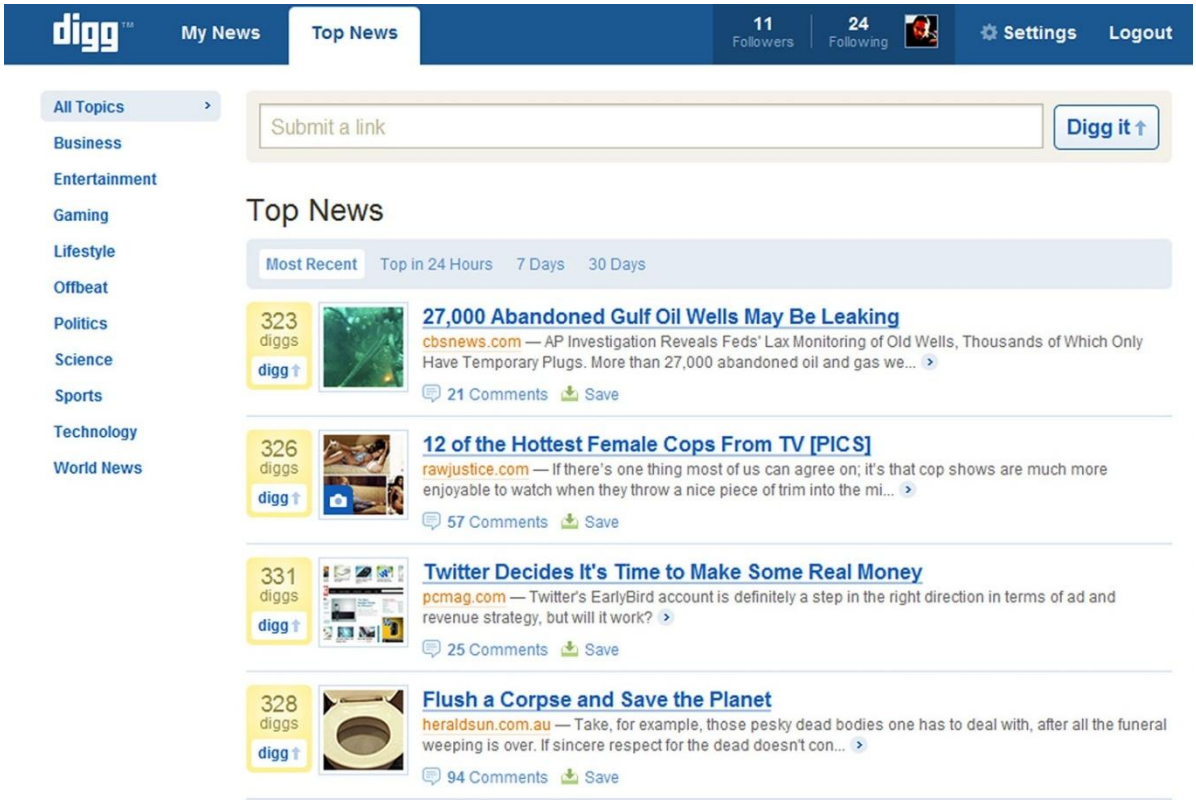


Figure 24. Flow Chart of the Process of Online News Diffusion on SNWs



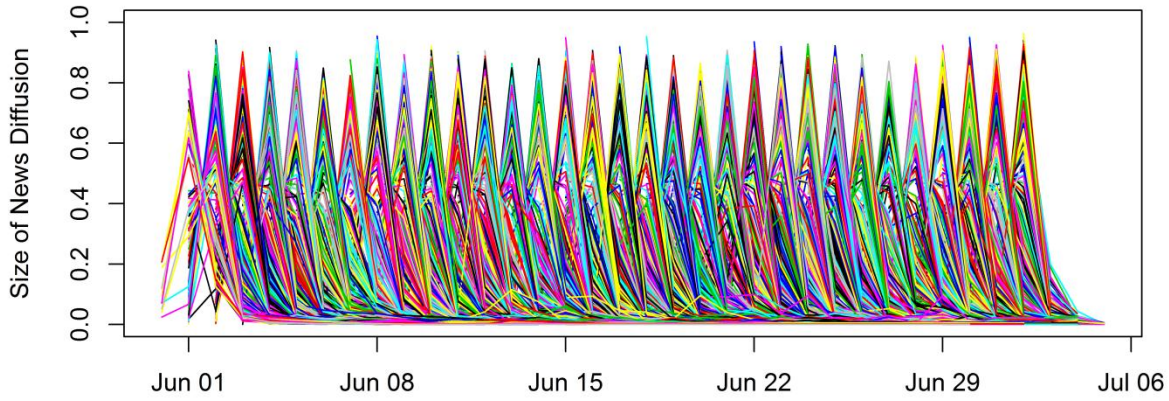
Note: phases VI, VII, and VIII denote the process of interpersonal effects, while phases II,III, IV, V denote the process of collective gatekeeping. VII, VIII, IX

Figure 25 The Preface of Digg



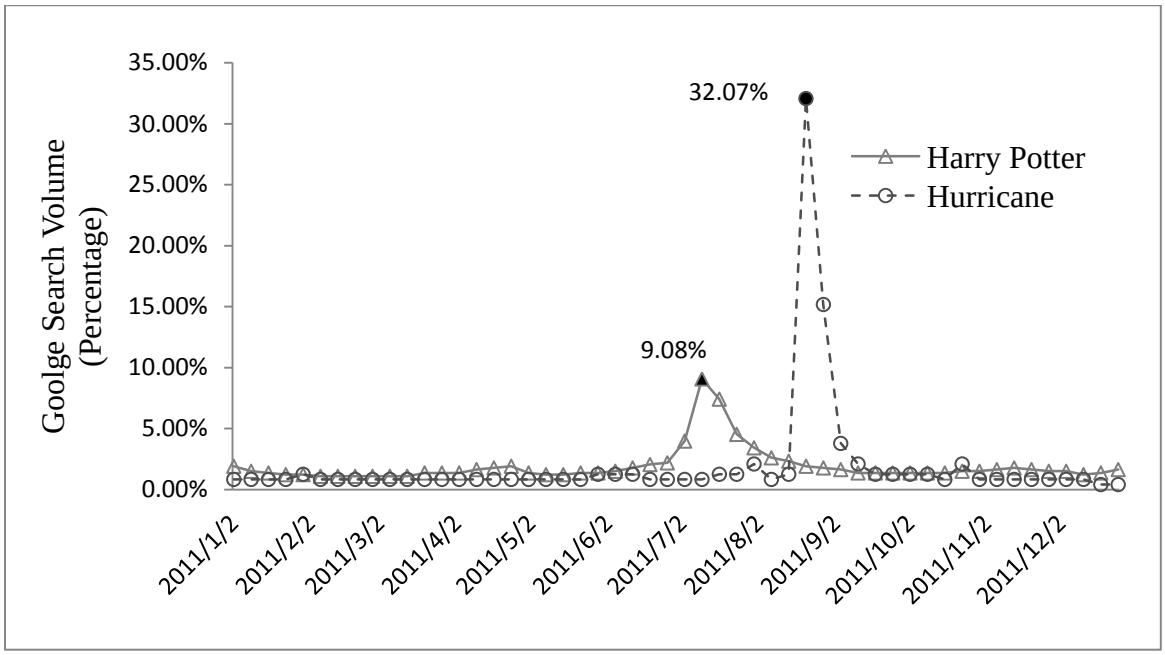
Note: If you visit the page of Digg while logged into your Digg account, “My News” documents the stories voted by your friends. Most recent news presents the most recent news, and top news in 24 hours, 7 days, and 30 days.

Figure 26. Size of News Diffusion over Time on Digg



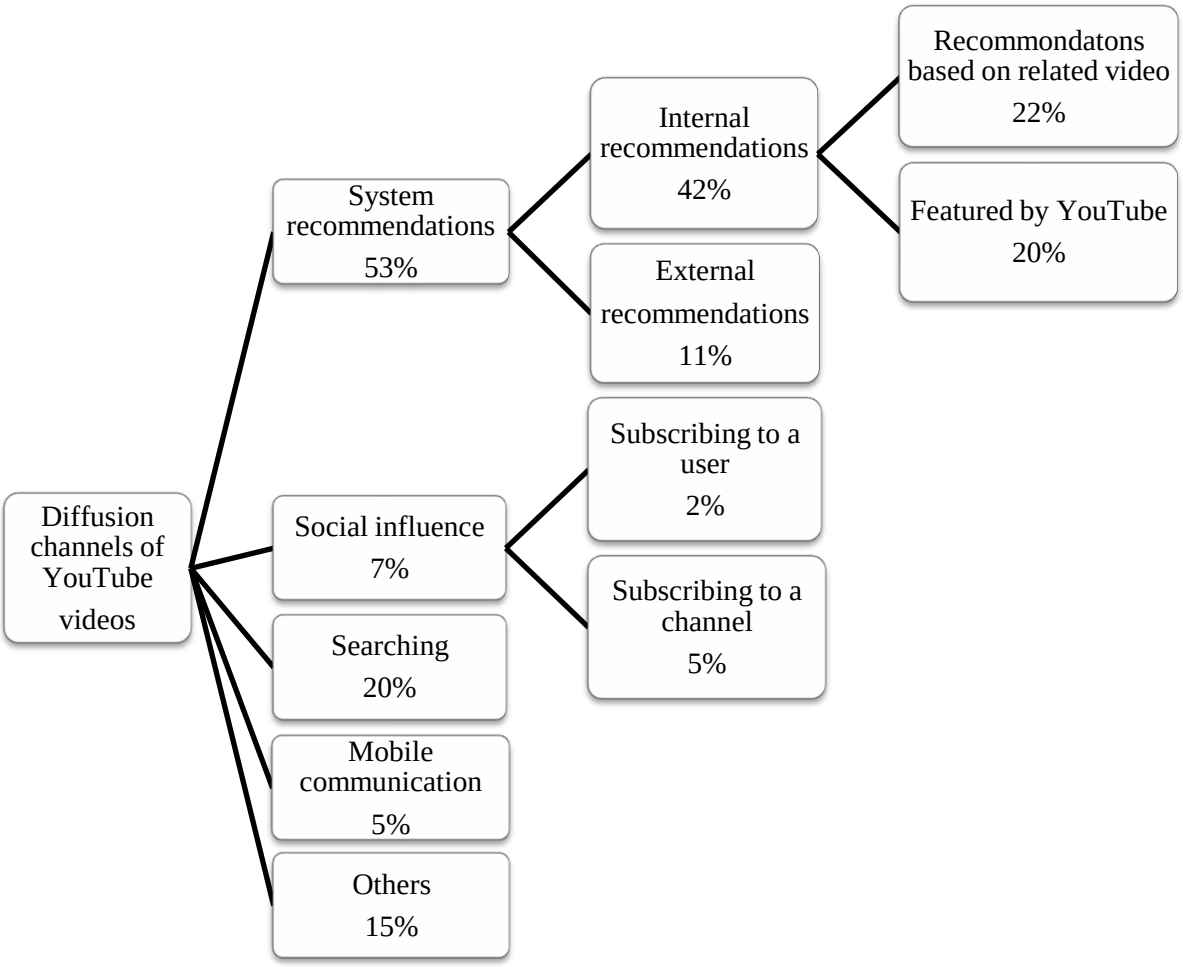
Note: This figure visualizes 3553 news diffusions over one month. The daily size of news diffusion is normalized to the range of 0 and 1. The colors are randomly selected to distinguish nearby diffusion curves, thus the same colors of different diffusion curves doesn't mean that they are the same news story.

Figure 27. Google Search Queries of “Harry Potter” and “Hurricane”



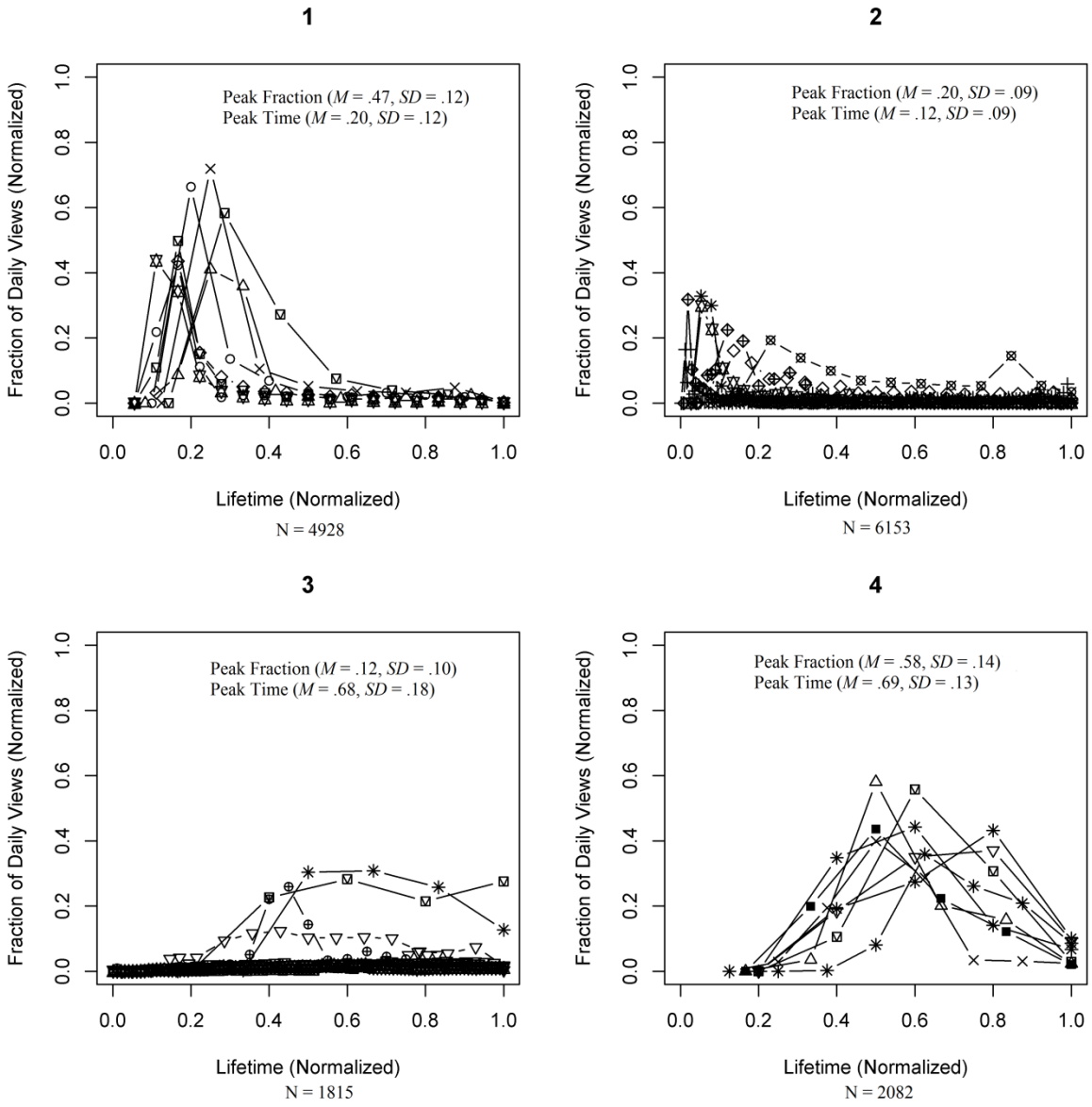
Notes: The data of this figure is collected from the Google Trends. The peak for the search query of “Harry Potter” is relevant to the release of the movie *Harry Potter and the Deathly Hallows*: Part 2 (2011), and the peak for the search query of “Hurricane” is caused by the hurricane Irene. The peak fraction of “Harry Potter” (32.07%) is much higher than that of “Hurricane” (9.08%); the relaxation process for the search query of “Harry Potter” is one month longer than that of “Hurricane”. Source: (Crane & Sornette, 2008)

Figure 28. Diffusion Channels of YouTube



Note: the number in each box denotes the percentage of views of the total views for14978 videos. For each video, YouTube supplies the significant discovery events—the noteworthy referral sources of viewers to a video, based on which, the percentage of diffusion channels of YouTube videos can be inferred. For example, the percentage of mobile communication is 5%, which means that 5% of total views come from mobile devices.

Figure 29. Bursts in the Diffusion Curves of YouTube Videos



Note: in each graph, only 8 randomly selected diffusion curves have been demonstrated to visually illustrate the features of bursts.

Appendix A. The Diffusion of URLs on Twitter

This appendix is one part of the diffusion of information on microblogs. However, since it is about the diffusion of URLs on Twitter, which is not quite coherent with Chapter 4. Thus, I put it here as an appendix. The overall purpose of this appendix is to understand the diffusion of URLs on Twitter. Users can submit tweets with URLs embed in them, and retweet the information. As a retweet, there is a label of "RT @" within the content. Thus, by mining the symbol of "RT @", the retweets of URLs can be detected. Following this logic, the proportion of interpersonal effects can be identified and compared with exogenous influence.

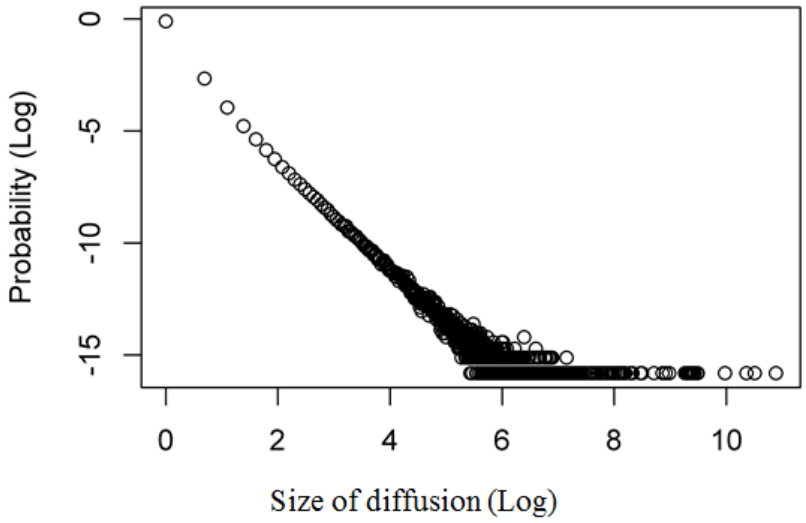
I employ the Tweets 2009 data collected by Yang & Leskovec (2011) [8] to describe the diffusion of URLs on Twitter. 476 million Twitter posts from 20 million users covering 7 months ranging from June 1 to December 31 in 2009. They estimate that this is about 20-30% of all public tweets published on Twitter during the particular time frame According to the latest research from eMarketer, the number of active users will be at around 18 million by the end of 2009 (For a summary of this data, see Table 1).

Table 1 Summary of Tweet 2009 Data

	Number	Percentage
Number of users	17,069,982	
Number of tweets	476,553,560	20-30%
Number of URLs	181,611,080	38.11%
Number of Hashtags	49,293,684	10.34%
Number of re-tweets	71,835,017	15.07%

[8] http://snap.stanford.edu/data/twitter7.html

Figure 1 The distribution of diffusion size for URLs onTwitter



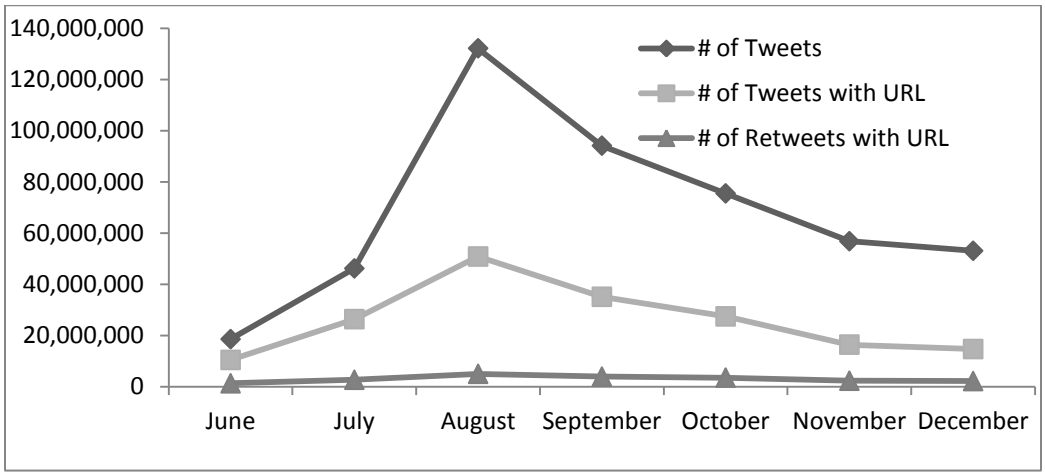
The distribution of the frequency of URLs is highly skewed (see Figure 1). In the tweets of June, 61.3% of the URLs have only been shared once; most URLs (82.2%) have only been shared for no more than 10 times; 90% of the URLs have been shared less than 100 times. To study the diffusion of tweets with URLs, I focus on those tweets which have been shared for more than 100 times in June 2009. Thus, I get 1994 URLs which have been disseminated about one million times in total.

Table 2 Summary of tweets by month

Month	# of Tweets	# of Tweets with URL	# of Retweets with URL	# of Retweets without URL	# of conversations without URL
June	18572084	10438309 (56.20%)	1307456 (7.04%)	1184294 (6.38%)	2984078 (16.06%)

July	46203172	26377469	2697781	2873958	7776501
		(57.09%)	(5.84%)	(6.22%)	(16.83%)
August	132210436	50840619	5014384	6541130	33138985
		(38.45%)	(3.79%)	(4.95%)	(25.07%)
September	94176126	35114523	3933239	5616726	24556872
		(37.29%)	(4.23%)	(5.96%)	(26.08%)
October	75520065	27449513	3528854	5425413	20674078
		(36.35%)	(4.67%)	(7.18%)	(27.38%)
November	56838024	16425281	2315358	4025743	17643612
		（28.90%）	(4.07%)	(7.08%)	(31.04%)
December	53088653	14698726	2178227	4264393	16700623
		(27.72%)	(4.11%)	(8.04%)	(31.49%)
Sum	476608560	181344440	20975299	29931657	123474749
		(38.04%)	(4.40%)	(6.28%)	(25.91%)

Figure 2 The evolution of Tweets with URLs



To summarize: I find that 38.04% of tweets contain URLs, and more than sixty percent (61.96%) of tweets do not include URLs. 15.97% of these tweets containing URLs have been retweeted by Twitter users, that is to say, 84.03% of the tweets containing URLs are driven by exogenous factors. This ratio of exogenous influence is even higher for the tweets with popular URLs (94.35%). Thus, interpersonal effects are not so powerful in disseminating the tweets of URLs.