ATTRACTABILITY AND VIRALITY:

THE ROLE OF MESSAGE FEATURES AND SOCIAL INFLUENCE IN HEALTH NEWS DIFFUSION

Hyun Suk Kim

A DISSERTATION

in

Communication

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Joseph N. Cappella, Gerald R. Miller Professor of Communication

Graduate Group Chairperson

_____

Joseph Turow, Robert Lewis Shayon Professor of Communication

Dissertation Committee

Michael X. Delli Carpini, Professor of Communication and Walter H. Annenberg Dean

Robert C. Hornik, Wilbur Schramm Professor of Communication

ATTRACTABILITY AND VIRALITY:

THE ROLE OF MESSAGE FEATURES AND SOCIAL INFLUENCE IN

HEALTH NEWS DIFFUSION

ACKNOWLEDGMENTS

I have been extremely fortunate to work on this project while enjoying the aid and support of many wonderful people to whom much credit and thanks are due. First and foremost, I wish to express my deepest gratitude to my advisor, Dr. Joseph Cappella, whose unfailing support and encouragement made this dissertation possible. He has always provided me with invaluable guidance and inspiration throughout this project and my entire graduate career, for which I cannot thank him enough. He is a model scholar, teacher, and mentor who I can only hope to emulate. I also offer my sincere appreciation to the other members of my dissertation committee: Drs. Michael Delli Carpini and Robert Hornik. Their insightful comments and suggestions have significantly strengthened this dissertation, and working with them has made me realize what it means to have the guidance of truly distinguished scholars. I could not have asked for a better committee.

In addition to my committee, I would like to thank Dr. Elihu Katz for his thoughtful feedback and encouragement when I was first developing the research questions behind this dissertation. I am also grateful to Dr. Paul Allison for his statistical advice on the analysis of pooled time-series cross-sectional data. With regard to the data collection and management for this dissertation, I owe a debt of gratitude to Tejash Patel, Vamsee Yarlagadda, Chandrakanth Maru, Radu Chebeleu, Dayeon Kim, Jung Min You, Sun-Ha Hong, and Rosie (EunGyuhl) Bae. I am also especially thankful to Michelle Jeong for her careful reading and thoughtful comments on the early drafts.

I would like to acknowledge the funding support for this work provided by the National Cancer Institute at the National Institutes of Health (R01CA160226 and 5U01CA154254) and the Annenberg (Dissertation Research Fellowship) and Wharton (Russell Ackoff Doctoral Student Fellowship) Schools at the University of Pennsylvania. I am also grateful to my great colleagues at the CECCR (and TCORS) Message Core team, including Dina Shapiro, Rui Shi, Holli Seitz, Minji Kim, Christine Skubisz, Laura Gibson, Emily Brennan, Heather Forquer, Erin Maloney, and Sungkyoung Lee.

I would also like to extend my heartfelt thanks to my dear friends at Penn. I am especially grateful to Susan Mello and Elizabeth Roodhouse, the best officemates one could ask for. Their support and friendship has kept me going, and my journey at Annenberg would not have been such a joy without the good times with them. Thanks also to Cabral Bigman-Galimore, Sarah Parvanta, Andy Tan, Susanna Dilliplane, Ashley Sanders-Jackson, Angela Lee, Chul-joo Lee, Young Min Baek, Kyung Lee, Minseop Kim, Jin Woo Kim, Minchul Shin, Eunsun Lee, Stella (Juhyun) Lee, Kwon Sang Lee, and Eunji Kim for all the conversations and laughter during my time at Penn.

Lastly, but most importantly, I am forever grateful to my family – my parents and brother – for their unwavering love and support, and for always believing in me. Words cannot express my most sincere appreciation for all that they have given me as parents and brother. I can never say thank you enough. I dedicate this dissertation to you.

ABSTRACT

ATTRACTABILITY AND VIRALITY:

THE ROLE OF MESSAGE FEATURES AND SOCIAL INFLUENCE IN

HEALTH NEWS DIFFUSION

Hyun Suk Kim

Joseph N. Cappella

What makes health news articles attractable and viral? Why do some articles diffuse widely by prompting audience selections (attractability) and subsequent social retransmissions (virality), while others do not? Identifying what drives social epidemics of health news coverage is crucial to our understanding of its impact on the public, especially in the emerging media environment where news consumption has become increasingly selective and social. This dissertation examines how message features and social influence affect the volume and persistence of attractability and virality within the context of the online diffusion of New York Times (NYT) health news articles. The dissertation analyzes (1) behavioral data of audience selections and retransmissions of the NYT articles and (2) associated article content and context data that are collected using computational social science approaches (automated data mining; computer-assisted content analysis) along with more traditional methods (manual content analysis; message evaluation survey). Analyses of message effects on the total volume of attractability and virality show that articles with high informational utility and positive sentiment invite more frequent selections and retransmissions, and that articles are also more attractable when presenting controversial, emotionally evocative, and familiar content. Furthermore, these analyses reveal that informational utility and novelty have stronger positive

associations with email-specific virality, while emotion-related message features, content familiarity, and exemplification play a larger role in triggering social media-based retransmissions. Temporal dynamics analyses demonstrate social influence-driven cumulative advantage effects, such that articles which stay on popular-news lists longer invite more frequent subsequent selections and retransmissions. These analyses further show that the social influence effects are stronger for articles containing message features found to enhance the total volume of attractability and virality. This suggests that those synergistic interactions might underlie the observed message effects on total selections and retransmissions. Exploratory analyses reveal that the effects of social influence and message features tend to be similar for both (1) the volume of audience news selections and retransmissions and (2) the persistence of those behaviors. However, some message features, such as expressed emotionality, are relatively unique predictors of persistence outcomes. Results are discussed in light of their implications for communication research and practice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Internet plays a central role in today's news media environment. About 50% of U.S. adults consume news online on an average day (Pew Research Center, 2012) and get most national and international news from the Internet (Pew Research Center, 2013a). The Internet and digital communication technologies have also affected the way news is consumed (Napoli, 2011; Rainie & Wellman, 2012; Shirky, 2008; Tewksbury & Rittenberg, 2012; Williams & Delli Carpini, 2011). The emerging media landscape has turned news consumption into an increasingly more *selective* and *social* communication behavior (Pew Research Center, 2010; K. C. Smith, Niederdeppe, Blake, & Cappella, 2013; Southwell, 2013).

People exercise greater selectivity in their news choice than ever before. Selective exposure is everywhere in today's news ecosystem where news sources and channels proliferate and individuals have a high level of control over what to choose (Bennett & Iyengar, 2008; Sunstein, 2007). News consumption is also a "socially-engaging and socially-driven" communication behavior in the new information environment (Pew Research Center, 2010, p. 4). News websites provide news-sharing tools to make it easier for their users to retransmit articles via email or social media such as Facebook and Twitter. A recent survey indicates that about 53% of U.S. adults get news forwarded to them via email or social media, and about 36% pass along news to others through those communication channels (Pew Research Center, 2010). Moreover, social influence cues are pervasive on news websites (Thurman & Schifferes, 2012).

Public signals about popular news stories (e.g., "most-viewed" and "most-emailed" articles), based on an automated aggregation of news consumption data, are presented saliently to online news consumers (Thurman & Schifferes, 2012).

Then, what makes news stories more "attractable" and "viral" in this complex information environment? In other words, why do certain news articles diffuse widely by triggering audience selections (attractability) and subsequent social retransmissions (virality), while others do not? Identifying factors that drive social epidemics of news coverage is essential to our understanding of its impact on audience cognitions, emotions, and behaviors in the new public communication environment (Bennett & Iyengar, 2008, 2010; Cappella, 2002; Holbert, Garrett, & Gleason, 2010; Hornik, 2002; Hornik & Yanovitzky, 2003; Slater, 2007; K. C. Smith et al., 2013; Southwell & Yzer, 2007).

Admittedly, the question is not new. The idea that media exposure is selective and media messages flow through social networks has received scholarly attention from early on in the communication literature (Katz, 1957, 2006; Katz & Lazarsfeld, 2006; Klapper, 1960; Lazarsfeld, Berelson, & Gaudet, 1968; Rogers, 2003). Decades of research have shed light on social and psychological factors that drive selective exposure to and social flow of media content. Selective exposure research has identified psychological factors that underlie audience message selection, such as congeniality bias (Hart et al., 2009; Iyengar & Hahn, 2009; Lazarsfeld et al., 1968; Stroud, 2011) and mood management or adjustment (Knobloch, 2003; Zillmann, 1988, 2000). Diffusion research has highlighted the role of social influence (or social contagion) in the spread of media messages through social networks (Bakshy, Rosenn, Marlow, & Adamic, 2012;

Cha, Benevenuto, Haddadi, & Gummadi, 2012; Katz, 1957; Katz & Lazarsfeld, 2006; Myers, Zhu, & Leskovec, 2012; Rogers, 2003; Tarde, 1903).

Yet, despite all of the research outlined above, there still remain questions that warrant theoretical and empirical attention. First, relatively little attention has been paid to how content characteristics relate to the attractability and virality of media messages (Katz, 1968, 1999; Rogers, 2003). Only recently has research begun to expand in this direction (e.g., Berger, 2013; Hastall & Knobloch-Westerwick, 2013; Knobloch-Westerwick & Sarge, 2013). Second, and more importantly, there is virtually no research that investigates how content features and social influence jointly impact what media messages people choose and share with their social networks. Third, while audience message selections and retransmissions are sequentially connected communication behaviors, they have rarely been examined together in the previous literature (Kim, Lee, Cappella, Vera, & Emery, 2013). Fourth, very little research has been conducted to investigate how message propagation channels (e.g., email vs. social media) affect what kind of media content people share with their social networks (Barasch & Berger, 2014). Finally, most existing research has focused on the volume of attractability and virality as an outcome, while leaving their persistence relatively understudied (Cappella, 2002).

This dissertation aims to fill the gaps in the literature by providing a more comprehensive framework for understanding drivers of audience message selections and retransmissions. Within the context of the online diffusion of New York Times (NYT) health news articles, the dissertation examines how message features, social influence, and their interactions affect news attractability and virality, both in terms of volume and persistence. The dissertation also investigates how digital news-sharing channels (email

vs. social media) shape what news goes viral. Employing computational social science approaches (Lazer et al., 2009; Parks, 2014) coupled with traditional research methods, this dissertation collects and analyzes (1) behavioral data on audience selections and retransmissions of the NYT articles, and (2) associated content and context data. Results of the dissertation shed new light on the role played by message features, social influence, and communication channels in driving online health news diffusion.

Overview of the Dissertation

This dissertation is organized as follows. Chapter 2 first conceptualizes the notions of attractability and virality under the framework of an epidemiological approach to message effects. The chapter then reviews theoretical and empirical literature on factors driving audience message selections and retransmissions, focusing on the role of message features, social influence, and online news-sharing channels such as email and social media. The notion of the persistence (or sustainability) of attractability and virality is also discussed.

Chapter 3 provides details of the methodology used to collect time-series behavioral data of audience selections and retransmissions of NYT health news articles, and associated content and context data on the articles. Three methodological approaches are detailed: (1) machine-based data mining, (2) content analysis, and (3) message evaluation survey.

Chapters 4 to 6 develop specific hypotheses and research questions, and present analysis methods and results of empirical tests. Chapter 4 investigates how message features relate to the total volume of news selections and retransmissions. It also tests

how the relationships between content characteristics and the volume of virality differ by online news-sharing channels: email and social media (Facebook and Twitter). Chapter 5 examines how public signals about news popularity (i.e., social influence cues) and their interactions with message features drive the temporal dynamics of news attractability and virality over the full course of news diffusion. It further evaluates how article content characteristics impact early news popularity in terms of selections and email-based retransmissions. Chapter 6 focuses on the persistence of news attractability and virality. It first shows how the volume and persistence measures are associated with each other (for both attractability and virality), and then explores the role of message features and social influence in shaping the persistence of news attractability and virality.

The final chapter, Chapter 7, summarizes the major findings of this dissertation in relation to existing research literature, and discusses their theoretical and practical implications. The chapter also points to limitations of the dissertation and suggests directions for future research.

CHAPTER 2

THEORETICAL AND EMPIRICAL BACKGROUND

Message effects research has mostly centered on persuasion or information processing outcomes. However, messages can also exert *diffusive effects*, such that certain messages are more likely than others to achieve enormous popularity by attracting audience attention and going viral (Berger & Milkman, 2012; Cappella, 2002; Hartmann, 2009; Jenkins, Ford, & Green, 2013; Kim et al., 2013). That is, media messages can be viewed not only in terms of their persuasiveness, but also in terms of their diffusiveness (Berger, 2013; Cappella, 2002; Gleick, 2011; Heath & Heath, 2007; Jenkins et al., 2013).

In line with this view, this dissertation employs an *epidemiological approach* to message effects with the outcome being message diffusion. In this approach, audiences as well as media are conceptualized as propagators of certain messages (Cappella, 2002). The approach assumes that there are certain features of messages that make them attract much attention and get widely shared, which have biological and/or sociocultural roots (Blackmore, 2000; Dawkins, 2006; Schaller & Crandall, 2004; Schudson, 1989; Shoemaker, 1996; Sperber, 1996). The approach also posits that social influence, as a contextual feature surrounding messages and audiences, has a vital role in message diffusion (Bass, 1969; Bikhchandani, Hirshleifer, & Welch, 1992, 1998; Christakis & Fowler, 2009; Granovetter, 1978; Muchnik, Aral, & Taylor, 2013; Salganik & Watts, 2009a; Schelling, 2006; Tarde, 1903; Watts, 2007). Specifically, this dissertation applies the epidemiological approach to the case of online health news diffusion.

In this chapter, I first explicate online news diffusion in the new information environment. I suggest that online diffusion of news stories involves two audience communication behaviors: selections and retransmissions. News attractability and virality are proposed as their respective corresponding message-level outcomes. Second, building upon the selective exposure and diffusion literature, I postulate how message features and social influence drive selective consumption and social sharing of health news. I also focus on the role of digital news-sharing channels (email vs. social media) in shaping the relationships between content characteristics and virality. Finally, I discuss the notion of persistence of news selections and retransmissions. I propose an exploratory proof-of-concept test of how message properties, social influence, and news retransmission channels impact the sustainability of attractability and virality.

Online News Diffusion: Attractability and Virality

In a broad sense, diffusion can be defined as "the spread of (1) an item, idea, or practice, (2) over time, and (3) to adopting units (individuals, groups, corporate units), embedded in (4) channels of communication, (5) social structures (networks, community, class), and (6) social values, or culture" (Katz, 1999, p. 147; see also Katz, Levin, & Hamilton, 1963; Rogers, 2003). In the context of online health news diffusion, each news article represents a "diffusing item," and an article is said to be "adopted" if it is read by a news consumer (i.e., a potential adopter). In other words, the total number of "adoptions" of an online news article is indicated by the total number of "exposures" that the article receives from news consumers.

As shown in Figure 2-1, health news articles diffuse on the Internet via two primary routes: *broadcast* and *viral* paths (Goel, Watts, & Goldstein, 2012; Katz & Lazarsfeld, 2006; Myers et al., 2012; Van den Bulte & Lilien, 2001). The broadcast diffusion path accounts for the portion of total adoptions of news articles that is made through news consumers' direct exposures to the articles. In other words, it represents the news exposure context in which people visit news websites and select certain articles for consumption. The viral diffusion path constitutes the portion of total exposures to news articles which results from social sharing (i.e., retransmission) of the articles. Getting access to news by following recommendations from people in one's social network (e.g., friends, family members, etc.) is a representative viral route to news diffusion (Goel et al., 2012; Hermida, Fletcher, Korell, & Logan, 2012; Myers et al., 2012; Pew Research Center, 2010).

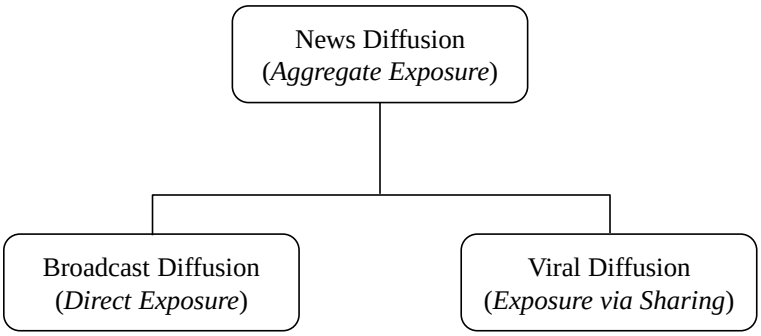

Figure 2-1. Two Primary Routes to News Diffusion

As is the case when diffusing other items, individuals play a dual role – potential adopters and propagators – in news diffusion, and they exercise selectivity in deciding both what to choose and what to share (Kim et al., 2013; K. C. Smith et al., 2013). For example, as potential adopters, people decide whether to read a certain article by picking it out of multiple available articles on news websites such as the New York Times website and Yahoo News. Once exposed to the article, as potential propagators, they

further decide whether to forward the article to their social networks via email or social media such as Facebook and Twitter, which in turn might lead the recipients to consume it (see Figure 2-2).

This dissertation focuses on these audience communication behaviors – selections and retransmissions – that underlie and determine health news diffusion. The dissertation proposes the notions of news *attractability* and *virality* to describe variability among news stories in triggering audience selections and retransmissions, respectively. News attractability is defined as the extent to which a news article invites selections from the audience, and news virality refers to the extent to which an article gets shared by people who consume it.

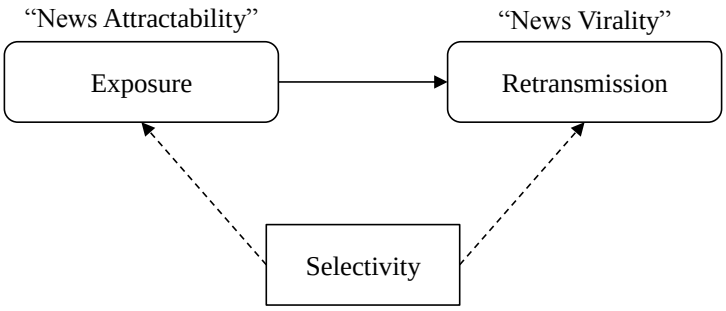

Figure 2-2. Selectivity in Audience News Selections and Retransmissions

In sum, online news stories diffuse more widely when they are both attractable and viral. Accordingly, this dissertation focuses on factors likely to shape both news selection and retransmission behaviors.

Before discussing drivers of news attractability and virality, it is important to note that selective exposure to online news articles and subsequent social sharing of the articles differ in their behavioral characteristics. From a message-effect standpoint, the two behaviors may take place in response to different message components of news articles. When selecting an article on a news website, people typically base their choice

only on the article's title and/or summary (or teaser). On the other hand, their news retransmission behaviors tend to take place after reading the article's full text.

This dissertation proposes that selection and sharing behaviors involve different motivations. News selection is a personal behavior, and tends to be driven by self-oriented motivations. Past research has identified motivations that underlie selective exposure behaviors, such as confirmation-seeking (Hart et al., 2009; Iyengar & Hahn, 2009; Jamieson & Cappella, 2008; Lazarsfeld et al., 1968; Stroud, 2011), mood management or adjustment (Knobloch, 2003; Knobloch & Zillmann, 2002; Strizhakova & Krcmar, 2007; Zillmann, 1988, 2000; Zillmann & Bryant, 1985), and informational-utility-seeking (Atkin, 1973, 1985; Freedman & Sears, 1965; Hastall, 2009; Katz, 1968; Knobloch-Westerwick, 2008; Knobloch-Westerwick, Carpentier, & Blumhoff, 2005). On the other hand, news retransmission is a social behavior, and might thus involve additional considerations compared to news selection. While news propagation is also driven in part by relatively self-focused motivations, people might also consider factors related to their target audience when engaging in this behavior such as audience characteristics (e.g., background and preference) and the nature (or strength) of their relationship with the audience (see also Falk, Morelli, Welborn, Dambacher, & Lieberman, 2013; Falk, O'Donnell, & Lieberman, 2012). Research has documented motivational and relational factors that trigger message retransmission behaviors, including altruistic or socializing motivations, status-seeking or self-enhancement motivations, and social connection/tie strength (De Angelis, Bonezzi, Peluso, Rucker, & Costabile, 2012; Harvey, Stewart, & Ewing, 2011; Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004; Ho & Dempsey, 2010; Huang, Lin, & Lin, 2009; C. S. Lee & Ma, 2012;

C. S. Lee, Ma, & Goh, 2011; Phelps, Lewis, Mobilio, Perry, & Raman, 2004; Sundaram, Mitra, & Webster, 1998). Taken together, the differences between news selections and retransmissions will be considered when discussing factors likely to affect these communication behaviors in sections below.

Determinants of Attractability and Virality

Message Features

This dissertation focuses on message features that previous research has suggested affect attractability and virality: informational utility, content valence, emotional evocativeness, novelty, and exemplification.

Informational Utility

Scholars have identified informational utility as a key driver of audience message selections (Hastall, 2009; Knobloch-Westerwick, 2008) and retransmissions (Berger, 2013; Berger & Milkman, 2012). The findings of a recent meta-analysis supported the notion that informational utility drives selective exposure (Hart et al., 2009). The meta-analysis revealed that while there is an overall tendency for individuals to prefer congenial over uncongenial messages (i.e., confirmation-seeking), the opposite is true when uncongenial messages have higher informational utility (see also Knobloch-Westerwick & Kleinman, 2012). Similarly, Knobloch-Westerick and colleagues have also highlighted the significant role of informational utility in fostering message exposure. According to them, when encountering external stimuli accompanied by potential threats or opportunities, people tend to seek out media messages with greater intensity in the following four dimensions: (1) perceived magnitude of challenges or gratifications, (2)

perceived likelihood of their realization, (3) perceived proximity in time or immediacy, and (4) perceived efficacy to influence the external stimuli (Hastall, 2009; Knobloch-Westerwick, 2008; Knobloch, Carpentier, & Zillmann, 2003).

Previous research has also suggested that messages with higher informational utility are more widely shared or circulated by individuals (Berger, 2013; Bordia & DiFonzo, 2005; DiFonzo & Bordia, 2007; Shibutani, 1966). Studies on message-sharing motivations have shown that people engage in this behavior to help or encourage their recipients by sharing useful information (Hennig-Thurau et al., 2004; C. S. Lee et al., 2011; Phelps et al., 2004; Sundaram et al., 1998). The idea that messages with high informational utility enjoy a retransmission advantage is supported by recent research on news virality. Thorson (2008) revealed that news articles offering practical advices about life issues (e.g., medical problems, finance, personal relationships, and jobs) stay longer on the New York Times website's "most e-mailed" list. Similarly, Berger and Milkman (2012) also found that articles conveying practically useful information are more likely to appear on the list.

In light of the theoretical and empirical literature reviewed above, this dissertation proposes that health news stories are more attractable and viral when they provide *efficacy information* (Bandura, 2004, 2009) which addresses effective means to achieve health-related goals such as promoting health and overcoming (or reducing) health threats, because such information has high practical value (Berger, 2013; Hastall & Knobloch-Westerwick, 2013; Kim et al., 2013; Knobloch-Westerwick & Sarge, 2013). Behavior change theories suggest that perceived self-efficacy is one of the primary determinants of health behaviors (Bandura, 2001; Fishbein & Ajzen, 2010), and a meta-analysis revealed

that high-efficacy messages are effective in promoting healthy cognitions and behaviors (Witte & Allen, 2000), all of which imply the high utility of efficacy information . In addition to the presence of efficacy information in health news stories as an intrinsic message feature, this dissertation also investigates how an overall sense of perceived usefulness impacts news attractability and virality (Berger & Milkman, 2012).

Content Valence

Research has shown that negatively valenced messages are more attractable. Scholars have suggested that individuals are hardwired for negative information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Shoemaker, 1996; Tversky & Kahneman, 1981). This psychological tendency, called *negativity bias*, indicates that "in most situations, negative events are more salient, potent, dominant in combinations, and generally efficacious than positive events" (Rozin & Royzman, 2001, p. 297). The negativity bias effect is also well established in selective exposure research (Donsbach, 1991; Knobloch, Hastall, Zillmann, & Callison, 2003; Meffert, Chung, Joiner, Waks, & Garst, 2006; Zillmann, Knobloch, & Yu, 2001). For example, Knobloch, Hastall, and colleagues (2003) revealed a selective exposure tendency toward Internet news stories with relevant threatening photographs (harm-related images), compared to news stories with relevant but innocuous photographs, or those without photographs (see also Zillmann et al., 2001).

In contrast to the case of audience message selections, research suggests that positivity bias operates in deciding what to share (Alhabash et al., 2013; Berger & Milkman, 2012; Kim et al., 2013). As discussed earlier, the decision to pass along news articles might involve more complex considerations than news selection, since news

retransmission is a more social activity. Previous research has shown that people's information-sharing decision involves considerations such as the characteristics of recipients (e.g., background and preference), anticipated responses from recipients (e.g., feelings), expected perceptions of recipients about them (e.g., peer recognition and reputation), and the nature or strength of their relationship with recipients (Hennig-Thurau et al., 2004; Ho & Dempsey, 2010; Huang et al., 2009; C. S. Lee & Ma, 2012; C. S. Lee et al., 2011; Phelps et al., 2004; Sundaram et al., 1998). This suggests that positive news will be retransmitted more frequently because they may make recipients feel good and help build or maintain the sharers' positive images (Berger, 2013). Recent empirical studies also support the idea that positive content is more viral. Berger and Milkman (2012) showed that positive news articles get shared via email more frequently than negative ones. Kim and colleagues (2013) found that tobacco control messages evoking positive rather than negative emotional responses are more likely to be retransmitted by smokers. An experimental study on viral advertising (Eckler & Bolls, 2011) also found that people are more likely to propagate video advertisements with positive sentiment than those with negative tone (see also Alhabash et al., 2013; Campo et al., 2013; Carter, Donovan, & Jalleh, 2011; Shifman, 2012; van den Hooff, Schouten, & Simonovski, 2012).

In sum, this dissertation predicts negativity bias in news selections, but hypothesizes positivity bias in news retransmissions. Building upon previous empirical works on attractability and virality, this dissertation examines content valence focusing on three specific message features. The dissertation evaluates effects of emotional valence both in terms of (1) *emotional responses* evoked by articles and (2) *expressed*

emotions in articles (Berger & Milkman, 2012; Kim et al., 2013). It also investigates how content *controversiality* (i.e., negative valence) impacts attractability and virality (Z. Chen & Berger, 2013; Zillmann, Chen, Knobloch, & Callison, 2004).

Emotional Evocativeness

Independent of the content valence of messages, emotional evocativeness has also been found to drive selective exposure to and social sharing of messages. Studies have suggested that emotionally arousing content captures audience attention (Heath & Heath, 2007; Zillmann et al., 2004). For example, Zillmann and colleagues (2004) found that people selectively seek out news articles presenting lead sentences with emotionally evocative frames (e.g., agony and conflict) rather than articles using lead sentences framed in a less emotionally intensive way (e.g., factual or economy).

Research has documented that the experience of emotional arousal triggers social sharing of the emotion, thereby making emotionally arousing messages spread through social networks (Christophe & Rimé, 1997; Harber & Cohen, 2005; Peters & Kashima, 2007; Rimé, 2009; Southwell, 2013). Scholars have suggested that individuals engage in social sharing of emotion because it has both intrapersonal and interpersonal benefits, such as the collective sense-making of their emotional experience and establishing (or strengthening) social bonds (Harber & Cohen, 2005; Peters & Kashima, 2007; Rimé, 2009). Empirical evidence for the role of emotional evocativeness in boosting content virality is also robust (Berger & Milkman, 2012; Dang-Xuan, Stieglitz, Wladarsch, & Neuberger, 2013; Heath, 1996; Heath, Bell, & Sternberg, 2001; Peters, Kashima, & Clark, 2009). This dissertation thus hypothesizes that the emotional evocativeness of health news articles is positively associated with both attractability and virality. As with the

case of emotional valence, this study examines emotional evocativeness both in terms of evoked and expressed emotions (Berger & Milkman, 2012).

Novelty

This dissertation posits that health news stories are more frequently selected and shared when their content is characterized by novelty (which is one of the prominent news values in journalism; Harcup & O'Neill, 2001; Shoemaker & Cohen, 2006; Stephens, 2007). People may seek out novel, surprising, unusual, or deviant news because such news tends to interrupt their routine information processing (or break the expectation of existing schema), and thus leads them to "stop and think" or consider it as potentially threatening information (Heath & Heath, 2007; Shoemaker, Chang, & Brendlinger, 1987; Shoemaker & Cohen, 2006). For example, a recent study revealed that individuals are more likely to select news stories containing deviant or unusual content (J. H. Lee, 2008; see also J. H. Lee, 2009).

Novelty may also boost virality because unusual or surprising content has high social currency and makes for good conversation material (Berger, 2013; Heath & Heath, 2007; E. Rosen, 2009). Research has shown that people are more likely to retransmit novel, surprising, or counterintuitive messages – including news articles (Berger & Milkman, 2012; Thorson, 2008), antismoking arguments (Kim et al., 2013), and folktales or jokes (Loewenstein & Heath, 2009; see also Moldovan, Goldenberg, & Chattopadhyay, 2011; Norenzayan & Atran, 2004; Norenzayan, Atran, Faulkner, & Schaller, 2006).

Exemplification

Scholars have suggested that messages crafted in a narrative form are more likely to invite social propagations (Berger, 2013; Heath & Heath, 2007; see also Gottschall,

2013). Stories may have a retransmission advantage because (1) they are a fundamental form of human cognition, knowledge, and communication, and easier to comprehend and recall (Bruner, 1986; Fisher, 1999; Schank & Abelson, 1995), and (2) they convey messages vividly and engagingly, thereby providing entertainment and instruction effectively (Berger, 2013; Heath & Heath, 2007).

In the case of news, exemplification has been identified as an intrinsic message feature that makes news more vivid, engaging, and thereby more story-like (Brosius & Bathelt, 1994; Zillmann, 2006; Zillmann & Brosius, 2000). Exemplars in a news article are "personal descriptions by people who are concerned or interested in an issue" (Brosius, 1999, p. 214) that the article addresses, and they act as a delivery vehicle for the article's central information (Cappella, 2006). While news is a highly structured and conventionalized form of narrative (van Dijk, 1988), research suggests that presenting relevant exemplars further enhances its narrativity (Kim, Bigman, Leader, Lerman, & Cappella, 2012). Taken together, this dissertation predicts that exemplification in health news articles boosts their virality.[1]

Social Influence

In the new information environment, people do not consume news in a vacuum. Public signals about news popularity, such as "most-viewed" and "most-emailed" news lists, are prevalent on Internet news websites (Thurman & Schifferes, 2012). News popularity indicators are automatically generated by continuously collecting and

[1] Exemplification may also affect news attractability (e.g., Hastall & Knobloch-Westerwick, 2013; Knobloch-Westerwick & Sarge, 2013). However, this prediction is not tested in this dissertation because exemplars are present in only a few teasers of the 760 New York Times health news articles (i.e., textual units used when predicting attractability; see Chapters 3 and 4 for details).

aggregating audience engagement with news articles (i.e., "aggregated collaborative filtering" or "aggregate user representations"; Thurman & Schifferes, 2012; Walther & Jang, 2012). That is, message features discussed earlier are not the only factor likely to affect audience news selection and retransmission behaviors in this complex public communication enviornment. When consuming news online, people are also pervasively exposed to public signals about what others read and share, which are cumulatively recorded, aggregated, and presented prominently on news websites.

This dissertation proposes that health news articles appearing on "most popular" lists are more likely to invite further audience selections and retransmissions. Public signals about news popularity may enable intially popular articles (either in terms of getting read or shared) to enjoy a cumuluative advantage (DiPrete & Eirich, 2006; Salganik & Watts, 2009a), which generates an information cascade or a "richer-get-richer" phenomenon (Bikhchandani et al., 1992, 1998; Sunstein, 2007).

Specifically, the dissertation posits that public signals about news popularity work as social influence cues for news consumers. When people do not make their decisions or behave independently from each other, social influence (or decision externalities) arises, meaning that "the likelihood of choosing some particular alternative depends in some manner on the choices of others" (Watts, 2007, p. 252). It is well established that people are signifcantly influenced by others' choices and behaviors (Bond et al., 2012; Cialdini & Goldstein, 2004; Fishbein & Ajzen, 2010; Goldstein & Cialdini, 2007; Muchnik et al., 2013; Pentland, 2014). Studies have proposed a wide array of psychological mechanisms to explain why social influence works, such as conformity bias, imitation of socially desirable behaviors (or opinions), and the use of mental

shortcuts (or heuristics) to avoid complex decision-making processes (Watts, 2007, p. 253; see also Bikhchandani et al., 1992; Chaiken, 1987; Sundar, 2008).

Research suggests that social influence occurs not only when the referent people are those within one's social networks (Bond et al., 2012; Fishbein & Ajzen, 2010; Muchnik et al., 2013), but also when they are anonymous or "impersonal" others (Cai, Chen, & Fang, 2009; Y. Chen, Wang, & Xie, 2011; Cialdini, 2003; Mutz, 1998; Rimal, 2008; Zhang, 2010). In particular, studies have shown that public signals about others' choices or behaviors serve as a cue to information credibility (Sundar, 2008; Sundar & Nass, 2001). Sundar and Nass (2001) found that people evaluated an identical set of news articles more favorably when they were told that the articles were selected by other people, compared to when informed that they were selected by other sources such as expert news editors, computerized news gathering system, or even the participants themselves (i.e., tailored recommendations).

Previous research has shown that public signals about the popularity of media content impact what people select and share. Salganik and colleagues (Salganik, Dodds, & Watts, 2006) revealed that online music consumption behavior is strongly driven by popularity information about songs, such that an initially popular song becomes more popular, whereas an initially unpopular song become more unpopular, demonstrating the social influence-driven cumulative advantage effects (see also Salganik & Watts, 2008). Similarly, Messing and Westwood (2012) found that news articles manipluated as receiving more recommendations or "likes" by Facebook users invited more frequent selections than those indicated as less popular, and the presence of this social influence factor made congeniality bias effects – partisan selective exposure – statistically

insignificant (see also Fu & Sim, 2011; Knobloch-Westerwick, Sharma, Hansen, & Alter, 2005). There is also good empirical evidence that social influence cues drive social retransmissions. Studies have shown that online content spreads on informational networks such as Twitter through a viral diffusion path, generating cumulative advantage effects on content virality (Lerman & Ghosh, 2010; Myers et al., 2012). An online experiment conducted on Facebook (Bakshy et al., 2012) also found that people exposed to social signals about their Facebook friends' message propagation behaviors are more likely to retransmit the message than those who are not exposed to such signals (see also Aral & Walker, 2011).

All in all, this dissertation predicts that health news articles that appear on "most-popular" lists in a given time interval will become more attractable and viral in a later time interval.

The Interplay of Social Influence and Message Features

This dissertation hypothesizes synergetic interaction effects between content characteristics and social influence on audience news selections and retransmissions, such that social influence-driven cumulative advantage effects are stronger for health news articles containing the aforementioned message features (i.e., informational utility, content valence, emotional evocativeness, novelty, and exemplification). In other words, the dissertation postulates that, while self-reinforcing effects of social influence cues on news selections and retransmissions are significant, cumulative advantage effects are more pronounced in news articles with certain message properties that enhance inherent attractability and virality (e.g., the presence of efficacy information).

The prediction of mutually reinforcing interaction effects between message features and social influence is based on the "compatibility" hypothesis developed in the diffusion literature (Katz, 1976, 1999; Rogers, 2003; Tarde, 1903). Scholars have suggested that while social influence plays a significant role in driving the social epidemic of items such as innovations and ideas by prompting imitation behaviors, the diffusion process also hinges upon how well the characteristics of diffusing items match the potential adopters in terms of their cultural, social and psychological background (Katz, 1976, 1999; Rogers, 2003; S. Rosen, 1981; Tarde, 1903). This suggests that message features inherently boosting the attractability and virality of news articles due to their good fit with news consumers should also lead to the articles benefiting more from social influence-driven cumulative advantage effects.

While little research has been conducted to examine the interaction effect between social influence and content characteristics on diffusion, the aforementioned music download study by Salganik and colleagues produced relevant results (Salganik et al., 2006; Salganik & Watts, 2008). While the study found evidence for strong social influence-driven cumulative advantage effects on music download behavior overall, it also revealed that the self-reinforcing impact of social influence tends to be stronger for more "appealing" songs (using download data for each song observed in an experimental group where social influence is absent as a measure of inherent "quality" of songs).

Retransmission Channels and Virality

This dissertation investigates how news-sharing platforms impact what health news goes viral. Specifically, the dissertation explores how effects of message features on virality differ between two types of online news retransmission channels of different

audience size (Berger & Milkman, 2012), focusing on the comparison between *email* (narrowcasting) and *social media* (Facebook and Twitter; broadcasting).

Email- and social media-based news retransmissions tend to assume different types of recipients. Email-based news forwarding usually targets an audience that is relatively small and narrow, and sharers specify particular receivers when they retransmit news via email. On the other hand, recipients of social media-based news sharing tend to be relatively large and diverse. When forwarding news articles via social media, sharers are less likely to target specific audience members from their entire online social networks (e.g., Facebook friends or Twitter followers), although it is also possible to do so on social media. As discussed earlier, sharers' consideration of recipients (e.g., recipients' background and preference, the nature or strength of the sharer-recipient relationship) plays a significant role in deciding what to share (Falk et al., 2013; Falk et al., 2012; Huang et al., 2009; C. S. Lee et al., 2011; Phelps et al., 2004). Thus, it seems reasonable to expect that news-sharing channels – email and social media – varying in their target audience impact what news goes viral by activating different motivations of news propagators (Berger & Milkman, 2012). However, not enough empirical evidence has been assembled to allow specific predictions about the impact of those channels. It is only recently that research has begun in this direction, focusing on the difference in sharers' focuses and motivations between when they retransmit messages to a relatively small and narrow audience and when they share those messages with a larger and broader audience (Barasch & Berger, 2014). Therefore, this dissertation poses a research question concerning the role of news retransmission channels (email vs. social media) in shaping the associations between message characteristics and virality.

Persistence of Attractability and Virality

This dissertation so far has focused on the *volume* of news attractability and virality. Yet, audience news selection and retransmission behaviors can also be examined in terms of their persistence or sustainability (Asur, Huberman, Szabo, & Wang, 2011; Berger & Iyengar, 2013, p. 577; Berger & Schwartz, 2011). Social diffusion of any item, regardless of its adoption volume, has a lifecycle such that it increases, reaches a peak, and declines over time until it stops (Rogers, 2003). This is particularly the case for news articles because by nature, the value of a news article tends to decrease with time while constantly facing competition for audience attention from other "newer" news articles (Asur et al., 2011; Leskovec, Backstrom, & Kleinberg, 2009; Szabo & Huberman, 2010; Wang & Huberman, 2012; F. Wu & Huberman, 2007; Yang & Leskovec, 2011).

Identifying drivers of the persistence of attractability and virality can broaden the basis for our understanding of message diffusion. It is also important because there are certain messages that do not achieve a high volume of attractability and virality, but have significant impacts on audience cognitions, emotions, and behaviors by continuing to get read and shared (i.e., surviving for a long time). Messages conveying misinformation or rumors are revealing in this regard (P. Smith et al., 2011; Sunstein, 2009). While such messages do not necessarily diffuse widely, their effects on audience judgments and decisions are consequential and tend to persist even after the particular misinformation or misbelief is corrected (Garrett, 2011; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Sunstein, 2009; Thorson, 2013).

While there appears to be a clear *conceptual* distinction between (1) the volume of audience selections and retransmissions and (2) their persistence, it is unclear whether

the two notions are *empirically* distinguishable because they are likely to be well-correlated with each other. Relatedly, one may also assume that message features and social influence affect the length of time for which news articles are viewed and shared in a similar way to which they shape the volume of news attractability and virality. However, to my knowledge, there is virtually no theoretical or empirical work on the persistence of news selections and retransmissions in terms of (1) its relationship with the volume of selections and retransmissions, and (2) its predictors. All things considered, this dissertation conducts an exploratory proof-of-concept test to address these questions. As an exploratory approach, the dissertation focuses on the same predictors used for examining the volume of attractability and virality: message features and social influence.

The Current Research

Study Context: Online Diffusion of New York Times Health News Stories

In sum, this dissertation investigates how message features and social influence impact the volume and persistence of health news attractability and virality, and how news retransmission channels (email vs. social media) shape what health news goes viral. The dissertation examines the proposed hypotheses and research questions using (1) behavioral data of audience selections and retransmissions of New York Times (NYT) health news articles and (2) associated article content and context data. Specifically, the dissertation focuses on NYT health news stories published online between July 11, 2012 and February 28, 2013.

NYT articles were chosen because at the time the dissertation data were collected, NYT was the only U.S. news outlet that enabled access to viewing and sharing *count* data

for each article.[2] Admittedly, the NYT website is not representative of all Internet news outlets. However, it is one of the most popular online news websites in the United States, which suggests that one can observe a sufficient amount of online news diffusion by focusing on NYT articles. As of September 2012, NYT was ranked first in digital circulation (= 896,352) among U.S. daily newspapers (Alliance for Audited Media, 2012). During the month of October 2012, it attracted about 48.7 million unique visitors online and was ranked as the second most popular online newspaper worldwide (comScore, 2012). Taken altogether, I concluded that using NYT health news data for online news diffusion research was a reasonable trade-off between the measurement quality of diffusion-related outcomes and the generalizability of study findings.

Control Factors

Effects of message features and social influence on attractability and virality are tested while controlling for the following potentially confounding content and context factors. This study includes the total number of selections as a covariate when predicting the total number of retransmissions. It is important to note that the observed frequency with which a news article has been retransmitted is partly a function of the number of times it has been viewed. Given that greater exposure to an article can lead to an increase in the *frequency* of sharing the article (i.e., simply having more opportunity to be shared), the sheer number of times that the article has been shared is confounded by the number of times that it has been viewed (see Godes et al., 2005). Therefore, in an observational setting like the one employed in this dissertation, it is essential to disentangle the

[2] Other news websites including news aggregators (portals) also provided sharing- and/or viewing-related information (e.g., Google News). However, they offered such information in the form of *popularity-rank* (e.g., Top 10 Most Popular articles), not in the form of actual *count* data.

likelihood of *news retransmission* from the likelihood of *news selection* by statistically controlling for the level of exposure when examining drivers of news virality.

This study also includes as a covariate the total amount of time that health news articles were shown in prominent locations on the main page of the NYT Health section to control for effects of an editorial cue to news values on attractability and virality (Graber, 1988; Sundar & Nass, 2001). To control for message factors potentially related to content credibility (Eastin, 2001; Hu & Sundar, 2010; Knobloch-Westerwick, Johnson, & Westerwick, 2013; Westerwick, Kleinman, & Knobloch-Westerwick, 2013), the following variables are included as covariates: (1) mention of professional sources and (2) factual or evaluative statements by expert sources. The month and day of the week in which articles were published online are also controlled to covary out potential seasonal or periodic variations in news selections and retransmissions. Other message-related control variables include: basic linguistic features (i.e., word count, use of complex words), use of words related to death, health, and social processes, mention of diseases or bad health conditions, topical area, writing style, presence of images, number of hyperlinks, and article column (assigned by the NYT). More details about the control variables are provided in Chapters 3 and 4.

CHAPTER 3

DATA

Overview

This dissertation uses data on health news articles that appeared on the *New York Times*' (NYT) website between July 11, 2012 and February 28, 2013 (about seven and a half months; 33 weeks). Health news articles were defined as those published in the *Health* section of the NYT website. All health news articles published *online* during the 33-week period comprised the news sample for this dissertation, except for the following: (1) articles from news agencies (e.g., AP and Reuters), (2) articles listed in the *Recipes for Health* series, (3) interactive articles (e.g., *Well Quiz* and *Think Like a Doctor* series), (4) obituaries, and (5) multimedia-based articles. This exclusion was made to ensure that articles were comparable in their content-type and format. As a result, the final sample consisted of 760 NYT health news articles.

The unit of analysis throughout this dissertation research is the article (N = 760). Specifically, I collected and analyzed data on two types of textual units for each article: *teaser* (title + abstract; for news attractability analyses) and *full text* (for news virality analyses). With respect to these two types of textual units, it should be noted that the article's abstract is not a part of its full text (e.g., lead sentences) but rather an independent summary of the full text.

Article-related data were collected using three broad categories of methodological tools. First, an automated software application was developed for machine-based data mining of diffusion indicators (i.e., aggregate behavioral measures of news selections and

retransmissions), content metadata (e.g., article URL), and context information (e.g., articles shown on the "most-emailed" list). Second, a content analysis, using both human and computerized coding procedures, was performed to measure objective features of article teasers and full texts (e.g., the presence of efficacy information, the number of positive emotion words). Third, a message evaluation survey was conducted to measure subjective (or perceived) features of article texts (e.g., emotional responses, perceived usefulness). In the following sections, details of each data collection method are described.

Machine-Based Data Mining: News Diffusion Tracker

In order to collect news diffusion-related data in an automated manner, this dissertation employed a "big data" method as used in computational social science (Lazer et al., 2009; Mayer-Schönberger & Cukier, 2013; Parks, 2014). An automated software application, the *News Diffusion Tracker* (NDT), was developed to collect diffusion indicators, content metadata, and context information for each health news article.[3]

Real-Time Data Mining

The NDT was programmed to perform two data-mining tasks simultaneously and on a real-time basis: (1) making and maintaining a connection to the NYT's Most Popular API (application programming interface[4]) to import data from the newspaper's database,

[3] The NDT was developed in collaboration with Tejash M. Patel, Vamsee K. Yarlagadda, and Radu Chebeleu from the Systems and Infrastructure Services team at the Annenberg School for Communication at the University Pennsylvania.

[4] An application programming interface (API) is a software interface which enables other applications to access and communicate with it. In the case of the NYT's Most Popular API, it can be viewed as a web service provided by the NYT for accessing data in its database. By

and (2) scraping information from the main page of the Health section of the NYT's website (http://www.nytimes.com/pages/health/).

The NYT's Most Popular API

The NDT fetched diffusion indicators and content metadata via the Most Popular API every 15 minutes. The data collection through the API was carried out in the following manner. The NDT made a connection to the API and requested parameters (i.e., types of data to be returned by the API endpoint) for every 15-minute interval. Specifically, the NDT set the parameters as follows: (1) *section* = health, (2) *diffusion indicators* = viewing count, sharing-via-email count, sharing-via-Facebook count, sharing-via-Twitter count, (3) *time-period* for the diffusion indicators = 24 hours, and (4) content metadata = title, abstract, URL, article category (column), etc.

With regard to the diffusion indicators, upon request by the NDT at every 15 minutes, the API provided *viewing* (news selection) and *sharing* (news retransmission) data for NYT health news articles that were published online *no earlier than 30 days* prior to the time of the request. As the time-period parameter was set to 24 hours, the API returned information about the number of times that a health news article had been viewed (i.e., page-views) and shared (via email, Facebook, and Twitter, separately for each retransmission channel) by NYTimes.com readers *in the last 24 hours* as of the time of request from the NDT. Specifically, at every 15-minute interval, the NDT obtained the following four lists of NYT health news articles (articles published more than 30 days before the time of observation were excluded from each list):

---

following the procedures and rules set by the NYT, the NDT can access and store the data in the format and structure that the NYT specifies.

1) Selection (Viewing): a list of articles that NYTimes.com readers had *viewed at least once* in the last 24 hours (as of the observation time), along with each article's viewing count during the 24-hour time period.
2) Email-Retransmission: a list of articles that NYTimes.com readers had *retransmitted via email using the NYT website's built-in sharing tool at least once* in the last 24 hours (as of the observation time), along with each article's email-retransmission count during the 24-hour time period.
3) Facebook-Retransmission: a list of articles that NYTimes.com readers had *retransmitted via Facebook using the NYT website's built-in sharing tool at least once* in the last 24 hours (as of the observation time), along with each article's Facebook-retransmission count during the 24-hour time period.
4) Twitter-Retransmission: a list of articles that NYTimes.com readers had *retransmitted via Twitter using the NYT website's built-in sharing tool at least once* in the last 24 hours (as of the observation time), along with each article's Twitter-retransmission count during the 24-hour time period.

Thus, of the NYT health news articles published 30 days or less before each time of measurement, articles not included in the "selection (viewing)" list were those that had never been viewed during the 24-hour time period. The same holds true for the retransmission-related article lists.

Web Crawler

The NDT also collected context information of health news articles using its built-in web crawler that scanned and scraped the main page of the Health section of the NYT website every 15 minutes, concurrently to the data mining of the NYT API. Specifically, the NDT fetched the following "snapshot" information by visiting, extracting, and processing the main page's HTML (Hyper Text Markup Language) source code at every visit: (1) a list of articles displayed in prominent locations (top six positions in the upper-left-hand corner of the page) and (2) articles shown on the "most-viewed" and "most-

emailed" lists (10 articles for each list) located under the area labeled "MOST POPULAR – HEALTH" in the right hand side of the page.[5] The data on news articles shown in (1) prominent locations, (2) the "most-viewed" list, and (3) the "most-emailed" list were then automatically transformed into article-level records that indicated whether an article was shown in each area of the Health section's main page at every observation time point.

Real-Time Data Management

All information gathered from the NYT's Most Popular API and the main page of the NYT website's Health section was *machine-readable*, which made it possible for the NDT to process and store the information in a fully automated manner. The API returned its responses in a machine-readable format (JSON or XML), which enabled the NDT to automatically build and update a news article dataset that included the diffusion indicators and content metadata. Similarly, the HTML source codes for the main page of the NYT Health section that were scraped by the NDT were also machine-readable. Therefore, the article records about prominent locations and news popularity lists were automatically integrated with the data collected via the Most Popular API.

The NDT was written in JavaScript and run on the MS SQL server of the Annenberg School for Communication at the University of Pennsylvania. The NDT was soft-launched on June 20, 2012 for beta tests. After a series of program revisions and fixes, its final version was launched at 12:00 AM on July 11, 2012 to ensure data

[5] The list of 10 most-viewed (or most-emailed) articles obtained at each observation time is, unsurprisingly, based on the selection (or email-retransmission) count data collected via the NYT API at the same time point. The NYT website presents the rank-order information of the viewing (or email-retransmission) count data in the last 24 hours. That is, the NDT collected both (1) the popularity information of health news articles that is visible to readers on the NYT website and (2) the actual diffusion data (i.e., selection or email-retransmission count data on which the popularity information is based) that is invisible but accessible via the NYT API.

collection of articles published online from that date onwards. From this time-point on, the NDT fetched and updated news diffusion indicators, content metadata, and context information for each NYT health news article, all simultaneously, at every 15-minute interval.

Post-Hoc Data Mining

Diffusion Indicators

The online news retransmission frequency data obtained from the NYT's Most Popular API (email, Facebook, and Twitter) are based on the aggregation of individual readers' news-forwarding behaviors conducted via the NYT website's built-in sharing tool available on each article's webpage. In other words, the news-retransmission occasions captured by the NYT API are limited to only those taking place on its website. Therefore, the API-provided news retransmission count for a news article is a *lower bound* on the actual sharing count for the article because the API does not keep track of alternative news propagation activities. Specifically, news-forwarding behaviors conducted *online* using other means than the NYT website's sharing tool can be categorized as follows: (1) *URL-only-based* retransmission where one shares a NYT article by just copying-and-pasting the article's URL into an email message, Facebook "status update," or "tweet"; (2) *Facebook's "share" function-based* retransmission where one "shares" a "status update" about an article (along with a URL link to its webpage) posted either by the NYT's Facebook page (www.facebook.com/nytimes) or one's Facebook friends; (3) *Twitter's "retweet" function-based* retransmission where one "retweets" a "tweet" about an article (along with a URL link to its webpage) sent by the

NYT's Twitter account (www.twitter.com/nytimes) or someone else. All the news propagation methods mentioned above are not included in the NYT API data.

The fact that the NYT API only keeps track of news-sharing behaviors taking place on its website (i.e., one particular method of retransmission) can pose a potential threat to the external validity of the retransmission frequency measures collected via the API. This would especially be the case when the measures are not very representative of all online news-forwarding behaviors including those conducted outside the NYT website.

To address this issue, I conducted a post-hoc data mining of aggregate news retransmission behaviors taking place on Facebook and Twitter, after the entire sample of the 760 NYT health news articles were identified. This post-hoc data collection was made possible by using publicly available social media APIs that allow access to Facebook and Twitter data, although it was, of course, impossible to gather data on the frequency of the URL-only-based retransmissions conducted via email. Specifically, the News Diffusion Tracker (NDT) obtained the *total* number of "shares" (i.e., retransmissions) for each article (identified by its unique URL information) on Facebook by accessing Facebook's API. The *total* number of "tweets" that include each article's URL link was collected using Topsy's API. It should be noted that because both the Facebook API and Topsy API use an article's URL as an identifier, the post-hoc data cover "status updates" and "tweets" for each article across the board for all news retransmissions methods discussed above (including the NYT's built-in sharing tool). In other words, the social media-based retransmission behaviors tracked by the NYT API are part of those tracked by the Facebook API and Topsy API.

Content Metadata

Based on the article URL information provided by the NYT API, I obtained additional content metadata for each article. Specifically, article full texts and online publication timestamps (date, hours, and minutes) were collected by parsing and processing each article's HTML source code.

Content Analysis

Objective message features of 760 health news articles were measured using content analysis. The objective message features refer to message variations that are independent of audience perceptions or responses (O'Keefe, 2003). Specifically, the objective message characteristics of the articles were content-analyzed using (1) human coding and (2) computerized coding methods. The content analysis was conducted separately for article teasers (title + abstract) and full texts.

Human Coding

Article Teaser

Article teasers were coded in terms of the following three objective message features: (1) the presence of efficacy information, (2) the mention of professional sources, and (3) the mention of diseases or bad health conditions. Content-coding of these message properties was done separately for *titles* and *abstracts* (brief summaries). Each of the title- and abstract-coding was performed by two trained research assistants who were blind to the hypotheses and research questions of this dissertation.

Efficacy information was coded to be present if a title (abstract) addressed one or more ways to promote health and wellbeing (or remain healthy) or to overcome (or avoid) a health risk/threat (Cappella, Mittermaier, Weiner, Humphryes, & Falcone, 2007;

Moriarty & Stryker, 2008). For example, the following abstract text was coded as efficacy information being present: "In a study, doing at least two and a half hours a week of either aerobic exercise or weight training substantially lowered the risk of Type 2 diabetes – but doing both may offer the greatest benefit."[6] Two coders also judged the presence of a mention of one or more *professional sources* in health areas, such as a specific expert individual, group, institution, or work(s) by these entities (e.g., doctor, researchers, CDC, FDA, a study, etc.). Finally, the coders identified whether there was any mention of one or more *diseases (or bad health conditions)* such as cancer, Alzheimer's disease, flu, sleep loss, and so on.

For each of the title- and abstract-coding tasks, a total of 90 cases were randomly drawn from the full news sample and used as reliability data for the nominal items described above (Krippendorff, 2013). A random half of the rest of the full sample was assigned to each coder. Inter-coder reliability estimates, measured using Krippendorff's α (Hayes & Krippendorff, 2007; Krippendorff, 2013), are shown in Table 3-1.

Table 3-1. Inter-coder Reliability for Article Titles and Abstracts

	Krippendorff's α	
	Title	Abstract
Presence of Efficacy Information	.78	.77
Mention of Professional Sources	.94	.89
Mention of Diseases / Bad Health Conditions	.82	.79

Note. All content codes are nominal.

Article Full Text

As with teasers, article full texts were coded by two trained research assistants who were unaware of this dissertation's hypotheses and research questions. The two

[6] This is an abstract of a NYT article titled "Weight Training May Lower Diabetes Risk": http://well.blogs.nytimes.com/2012/08/07/weight-training-may-lower-diabetes-risk

coders assessed the following items: efficacy information, exemplification, credibility statements, topical area, and writing style.

As with the case of teasers, *efficacy information* was defined as information that addresses way(s) to promote health and wellbeing (or remain healthy) or to overcome (or avoid) a health risk/threat (Cappella et al., 2007; Moriarty & Stryker, 2008). The coders judged the presence of efficacy information in an article's full text. The coders were also given the following instruction for coding efficacy information in article full texts (adapted from the codebook of Cappella et al., 2007):

> Efficacy information usually gives specific details about what can be done or explicit instructions on about how to remain healthy. This includes any of the possible means/strategies that can (or should have been done) prevent or treat health outcomes (or promote health and wellbeing), among which are medicines, treatments, prevention behaviors (e.g., exercise, diet, nutrition, etc.), screening/testing.

Exemplification was defined as a discussion (or mention) of a narrative (personal case/experience) of a person or family that is related to the subject of a given news article. The coders were asked to record the presence of exemplification in each article's full text. Additional instructions were given as follows (adapted from the codebook of Cappella et al., 2007):

> Any human being, including celebrities and historical figures, can constitute an exemplar as long as the person is not a fictional character. An exemplar must contain some concretizing or identifying information about the person or family described in the article (e.g., name, age, gender, location, or health outcome, etc.).

Credibility statements were assessed using two coding items. First, the coders were asked to find *statement(s)* attributed to (or made by) a specific expert individual, group, or institution (e.g., doctors, researchers, and government organizations such as the CDC and FDA) that provides factual information or offers an evaluative opinion with regard to the subject of a news article. The coders indicated whether a given article's full

text contained (1) no credibility statement, (2) one credibility statement, or (3) two or more credibility statements. Second, the coders judged the presence or absence of one or more credibility statements *opposing or contradicting* to the credibility statement(s) identified by the first coding item (Cappella et al., 2007).

The *topical area* of a news article was coded using categories adapted from a research report on health news coverage (Kaiser Family Foundation & Pew Research Center, 2009). The coders were asked to choose one of the following broad categories for a given news article's topical area. First, a "health policy and health care system" category included articles about issues concerning health policy, law, regulation, health insurance, or other government health programs (e.g., Medicare). Second, a "public health" category included articles that focus on pandemics/epidemics (e.g., bird flu, swine flu, influenza) or environmental health concerns. Third, a "diseases and health conditions" category included articles that discuss the causes, effects, prevention, or treatment of diseases or health conditions (risks). Articles about medical research on the related areas were included here. Fourth, a "global news" category included articles about health issues in countries outside the U.S. Fifth, the coders were instructed to choose a "none of the above" option when a news article was thought to be unrelated to any of the four broad topical areas described above.

Finally, the coders assessed the writing style of a health news article by judging whether the article was *written in a first-person point of view*.

Reliability data for article full texts consisted of 80 cases that were randomly selected from the full news sample. Each coder then assessed a random half of the rest of

the full sample. Table 3-2 presents final inter-coder reliability estimates for the nominal items using Krippendorff's α (Hayes & Krippendorff, 2007; Krippendorff, 2013).

Table 3-2. Inter-coder Reliability for Article Full Texts

	Krippendorff's α
Presence of Efficacy Information	.77
Presence of Exemplification	.92
Credibility Statements	1.00
Presence of Opposing Credibility Statements	1.00
Topical Area	.83
Writing Style (First-Person Point of View)	.84

Note. All content codes are nominal.

Computerized Coding

LIWC 2007

The text analysis software program Linguistic Inquiry and Work Count (LIWC 2007; Pennebaker, Booth, & Francis, 2007) was used for computer-assisted content analysis of article teasers and full texts at the word-level. LIWC counts words that belong to psychologically meaningful categories (e.g., positive/negative emotion words) defined by its own internal dictionary, which is developed based on human judgment of word categories. The LIWC 2007 dictionary classifies approximately 4,500 words and word stems in about 80 categories (for details about the reliability and validity of LIWC 2007, see Bantum & Owen, 2009; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007; Tausczik & Pennebaker, 2010). The LIWC lexicon has also been widely used in previous studies that employed computational social scientific methods (e.g., Golder & Macy, 2011).

Computerized coding was conducted separately for article teasers and full texts. The LIWC 2007 lexicon covered a reasonably broad range of words used in both types of

texts, with high average word-coverage rates. The LIWC 2007 dictionary words covered on average about 80.4% of the words in the 760 article teasers (*SD* = 8.8, *Min* = 50, *Max* = 100). Even when the dictionary coverage rate was assessed with unique words, its average was only slightly reduced (*M* = 76.4, *SD* = 9.0, *Min* = 45.8, *Max* = 96.9). This suggests that the coverage rate for the raw set of words is not a mere artifact of the fact that the dictionary covered frequently-occurring words disproportionately well. For full texts, the LIWC 2007 lexicon covered on average about 80.7% of the words (*SD* = 3.9, *Min* = 66.2, *Max* = 92.3). As with the case of article teasers, the coverage rate decreased only slightly when it was calculated based on unique words (*M* = 74.0, *SD* = 4.2, *Min* = 60.7, *Max* = 86.6).

This dissertation focused on seven word categories of the LIWC 2007 lexicon that tap into basic linguistic features and social-psychological domains. The "word count" and the frequency of "words longer than six letters" (an indicator of writing complexity; Tausczik & Pennebaker, 2010) categories were measured as basic linguistic features. With regard to social-psychological domains, the following word categories were analyzed: "positive emotion" (e.g., *good*, *happy*, and *hope*), "negative emotion" (e.g., *bad*, *fear*, and *sad*), "death" (e.g., *die*, *kill*, and *mortality*), "health" (a category including diseases- and clinic-related words such as *cancer*, *clinic*, *colonoscopy*, *flu*, *mammogram*, *obesity*, and *pill*), and "social processes" (a category covering words pertaining to family, friends, and social interactions; e.g., *daughter*, *friend*, *husband*, *neighbor*, *talk*, and *share*).

Other Computerized Method for Coding Article Full Texts

This dissertation also used a HTML parser. Using URL information for each article, the parser counted the number of hyperlinks embedded in full text by processing

the HTML source code for the article webpage. Hyperlinks for the author(s) of news articles (i.e., byline hyperlinks) were also included.

Message Evaluation Survey

In order to measure perceived (or effect-based) message features of health news articles (O'Keefe, 2003), I conducted a message evaluation survey where respondents read and rated article teasers (title + abstract) and full texts on the Internet. The goal of this survey was to obtain an evaluation score (e.g., usefulness) for each article by aggregating evaluations from multiple respondents who read the same article. This methodological approach is equivalent to crowd-sourcing the evaluation process. Respondents' aggregate assessments are consequential because they are precisely what we want to know about messages: "average" perceptions or reactions regarding the messages.

Before describing the details of the message evaluation survey (MES), it should be emphasized that there is an essential difference between the content analysis (CA; discussed in an earlier section) and the MES in terms of the nature of message features that these methods measure. The CA method assumes that the CA-generated data on an objective (intrinsic) message feature of an article are independent of coders (Krippendorff, 2013). Therefore, the CA method expects substantial agreement between the coders about the message feature (i.e., the coders are interchangeable), and hence there should be a high-level of inter-coder agreement for the CA data to be considered reliable. On the other hand, the MES method focuses on message variations that are assumed to be subjective (i.e., perceived or effect-based message features). For example, the MES

method posits that a person's evaluation about the "usefulness" of an article is in part dependent upon one's own characteristics such as personality and life experience. Consequently, the MES does not necessarily assume a great deal of agreement between respondents' evaluations of the same article about its perceived usefulness. Rather, the MES aggregates (averages) multiple respondents' ratings about the usefulness of an article, thereby canceling out the individual differences in their ratings, and uses this aggregate information as a usefulness score for the article. Therefore, achieving substantial inter-respondent agreement about perceived message features is not necessary for the MES-generated content data to be considered reliable. Instead, the reliability of the MES data hinges on the number of respondents assigned to each article for rating its perceived message features. The more respondents evaluate each article, the more reliable their aggregate evaluations will be in general.[7] Related methodological considerations on the current message evaluation survey are discussed in further detail in Appendix A.

Survey Sample and Design

Survey respondents were recruited through Amazon's Mechanical Turk (MTurk; www.mturk.com), a web-based platform for crowdsourcing human intelligence tasks that include participating in surveys and experiments. Not only has MTurk been widely used recently for survey and experimental studies, but its samples have also been shown to be more diverse and similar to the general population than other traditional convenience

[7] The downside of the "perceived" message features obtained by crowd-sourced aggregate assessments is that, unlike "objective" or "intrinsic" message features, they cannot be manipulated in any obvious way (see O'Keefe, 2003 for further discussion about the difference between the two types of message properties and its implications for message effects research).

samples (e.g., college students). Moreover, using MTurk samples, several studies have replicated the results of well-established social science experiments and those of previous studies that recruited more representative samples (for more details about the validity of survey and experimental studies using MTurk samples, see Behrend, Sharek, Meade, & Wiebe, 2011; Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). Recent computational social science research has also recruited MTurk samples to assess a large number of online messages extracted from web sources (e.g., Bakshy, Hofman, Mason, & Watts, 2011).

A total of 5,092 U.S. adults participated in the message evaluation survey (aged 18 to 80 years; $M = 33$, $SD = 11$). Among the 5,092 respondents, about 51.0% were female, 76.6% were non-Hispanic White, 69.5% were currently employed, and 51.3% completed some college or more education.

The survey was conducted online over about a one-month period. Recruitment advertisements were posted on the MTurk website throughout the period. The advertisements provided a hyperlink to a survey website that was designed for this dissertation research and hosted at the server of the Annenberg School for Communication at the University of Pennsylvania.[8] Interested people who clicked on the hyperlink were redirected to the survey website and presented with an electronic copy of the consent form. Once they agreed to participate, the survey started. Respondents who completed the survey were offered $1 as compensation for their participation (payments were made through MTurk).

[8] The survey website was programmed in collaboration with Tejash M. Patel, Chandrakanth Maru, and Radu Chebeleu from the Systems and Infrastructure Services team at the Annenberg School for Communication at the University Pennsylvania.

During the survey, each participant was asked to read and rate six pieces of article texts (three teasers and three full texts) that were randomly selected from the entire sample of NYT health news articles. The survey was programmed to sample six *different* news articles to ensure that no respondent would evaluate a full text and a teaser of the same article. The survey consisted of three sections; in each section, the respondents evaluated one full text and one teaser text. Of the entire set of 760 news articles, one article was mistakenly excluded from the sampling pool due to an unexpected technical error in the programming script for the survey website. Consequently, a total of 759 articles were evaluated by the respondents in this survey.

As each of the 5,092 respondents rated three article teasers and three full texts, the survey generated 15,276 (=5,092 × 3) message evaluations for each type of article text. As the sampling and assignment of article texts were completely randomized for each respondent, the average number of respondents per article was about 20.1 for teasers (*SD* = 4.5) and full texts (*SD* = 4.4), which is consistent with the expected number derived from the survey design (i.e., 15,276 evaluations ÷ 759 articles). Figure 3-1 presents the frequency distribution of the number of respondents per article.

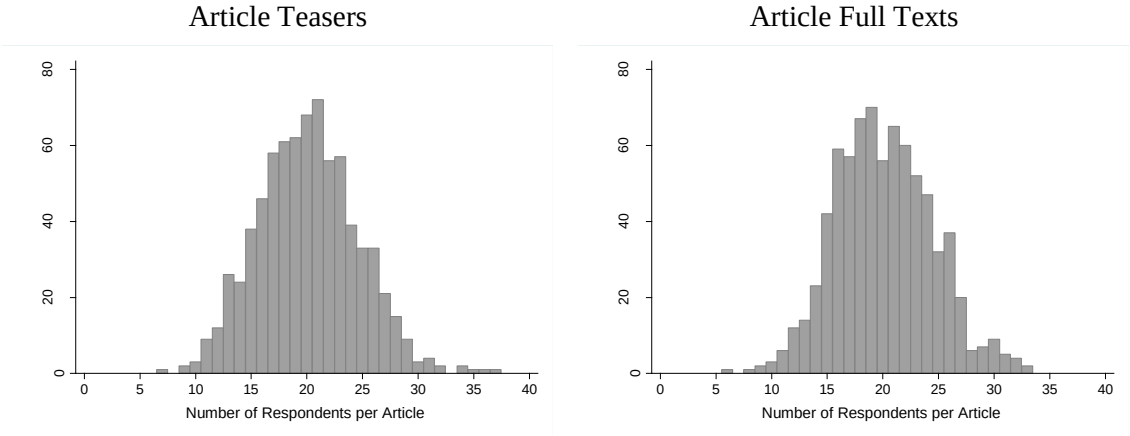

Figure 3-1. Histogram of the Number of Respondents per Article

Measures

Respondents answered a series of questions for each article text. An identical set of questions was asked for both full texts and teasers, with the exception of a minor variation in the question wording that referred to the type of article text (i.e., "article" vs. "article teaser").

For emotional responses-related items, respondents were presented with eight emotion words and asked "How much does each of the following words describe how you felt while reading the article [article teaser]?" with response options ranging from "not at all" (= 1) to "extremely" (= 5). The eight emotion words were as follows: *pride*, *amusement*, *contentment*, *hope*, *anger*, *fear*, *sadness*, *surprise*. The choice of these emotion words was based upon the literature on (1) basic emotions theories (Lazarus, 1991; Shaver, Schwartz, Kirson, & O'Connor, 1987) and (2) the role of discrete emotions in message diffusion and health communication (e.g., Dillard & Nabi, 2006; Heath et al., 2001). Emotional evocativeness (arousal) was measured with a single item. Respondents indicated the extent to which the news article [article teaser] they read made them feel *aroused*, on a 5-point scale ranging from "not at all" (= 1) to "extremely" (= 5).

To measure the perceived novelty of news articles, in addition to the "surprise" item mentioned above, respondents were asked two items that were adapted from previous studies (Kalyanaraman & Sundar, 2006; Turner-McGrievy, Kalyanaraman, & Campbell, 2013). Respondents indicated how strongly they agreed with the statement that the information presented in the article [article teaser] was *new* and *unusual*, on a 5-point scale ranging from "strongly disagree" (= 1) to "strongly agree" (= 5).

Perceived controversiality and usefulness were measured using the same question stems as the "newness" and "unusualness" items. Respondents indicated how much they agreed or disagreed with the statement that the information presented in the article [article teaser] was *controversial* (Z. Chen & Berger, 2013) and *useful* (Berger & Milkman, 2012). Response options ranged from "strongly disagree" (= 1) to "strongly agree" (= 5).

Summary

In this chapter, I described the details of the three broad categories of methodological tools used in this dissertation to collect various sets of data on 760 New York Times health news articles that were published online over 33 weeks: (1) machine-based data mining, (2) content analysis, and (3) message evaluation survey. Table 3-3 summarizes the different article-related data collected via the different methodological tools. Descriptive statistics of these data will be presented in later chapters.

Table 3-3. List of Article-Related Data by Data Collection Methods

Machine-Based Data Mining	Content Analysis	Message Evaluation Survey
Diffusion Indicators - *Real-Time (NYT API)* - Viewing - Sharing (Email, Facebook, Twitter) - *Post-Hoc (Social Media APIs)* - Sharing (Facebook, Twitter) Content Metadata - *Real-time (NYT API)* - Title - Abstract - Article URL - Article Category - Image URL(s) - *Post-Hoc(HTML Parser)* - Article Full Text - Online Publication Timestamp Context Information - *Real-Time (Web Crawler)* - Articles Shown in - Prominent Locations - "Most-Viewed" List - "Most-Emailed" List	Human Coding - Presence of Efficacy Information - Mention of Professional Sources (Teaser Only) - Mention of Disease (Teaser Only) - Credibility Statements (Full Text Only) - Topical Area (Full Text Only) - Writing Style (Full Text Only) Computerized Coding - *LIWC 2007* - Word Count - Words Longer than Six Letters - Positive Emotion Words - Negative Emotion Words - Death-Related Words - Health-Related Words - Social Processes-Related Words - *HTML Parser* - Number of Hyperlinks (Full Text Only)	Survey Items - Emotional Valence - Pride - Amusement - Contentment - Hope - Anger - Fear - Sadness - Emotional Evocativeness (Arousal) - Novelty - Newness - Unusualness - Surprise - Controversiality - Usefulness

CHAPTER 4

MESSAGE EFFECTS ON ATTRACTABILITY AND VIRALITY

Overview

This chapter examines how message features relate to the *total* volume of news attractability and virality using aggregate online behavioral data of news selections and retransmissions observed in a natural setting. The total volume of news attractability and that of virality are defined as the total frequency with which New York Times (NYT) health news articles have been (1) viewed and (2) shared via communication channels (email, Facebook, and Twitter), respectively.

Message effects models are examined to identify content-level ingredients of the total volume of news attractability and virality, with a focus on message features central to this dissertation that are related to (1) informational utility, (2) content valence, (3) emotional evocativeness, (4) novelty, and (5) exemplification (only for news virality). This chapter further investigates how news-sharing platforms (email vs. social media) moderate the impact of message features on virality (Barasch & Berger, 2014; Berger & Milkman, 2012).

The article sample consists of 760 NYT health news articles published online between July 11, 2012 and February 28, 2013. The unit of analysis is the article teaser (i.e., title and abstract) for a message effects model predicting the total volume of news attractability because teasers are article-specific textual information available to the audience in most news-choice environments (e.g., readers visiting the main page of the

NYT's Health section or those receiving email newsletters from the NYT).[9] For the total volume of news virality, the unit of analysis is the article's full text. The assumption here is that readers are more likely to retransmit an article after exposure to its full text rather than making the propagation decision based solely on its teaser without viewing the full text. This assumption appears to hold especially for the news retransmission data analyzed in this dissertation. As detailed in Chapter 3, the NYT's Most Popular API tracked and aggregated audience news-sharing behaviors conducted via the NYT website's built-in news-sharing tool embedded in every article's webpage. This means that the news propagation occasions kept track of by the NYT API were limited to those taking place after readers viewed a given article (or, more technically, after clicking and opening the article webpage).[10]

Focal message characteristics examined as predictors of the total volume of news attractability and virality include the presence of efficacy information, usefulness, emotional valence and evocativeness, controversiality, novelty, and exemplification (for virality only). The effects of emotional valence and evocativeness are assessed using

---

[9] To be sure, it is sometimes only article "titles" that are shown to readers (e.g., articles that appear on the lower part of the main page of the NYT Health section. However, the navigation of the NYT website and the email-newsletter sent by the NYT suggests that it is more frequently the case that readers are exposed to article teasers in their article-choice situations. In any case, it is important to note that this dissertation has no data as to exactly in what situations articles were chosen to be read by NYTimes.com readers. That is, there is uncertainty with respect to whether it was an article's title or teaser that was shown to the readers when they chose to read the article's full text. Given this, I opted to use as much textual information as possible rather than discarding potentially important piece of textual information (i.e., article abstract) for understanding the readers' article selection.

[10] Note, however, that the same does not necessarily hold true for news-sharing behaviors captured by social media API sources described in Chapter 3 (i.e., Facebook API and Topsy API) because, for example, one can click a "share" button for a Facebook post (or "retweet" a tweet message) that contains a link to an article without having to visit the article's webpage.

variables obtained through both a message evaluation survey (i.e., emotional responses) and a computerized coding method (i.e., expressed emotions).

Control variables are as follows: mention of diseases or bad health conditions (only for attractability), mention of professional sources (only for attractability), credibility statements (only for virality), the presence of death-related words, words related to health and social processes, word count, writing complexity, article column (category), article topic (only for virality), the presence of images (only for virality), number of hyperlinks (only for virality), and the article's online publication month and day of the week.

Hypotheses and Research Questions

Building on the theoretical and empirical literature reviewed in Chapter 2, I posit a series of hypotheses and research questions about the effects of message features on the total volume of news attractability and virality.

Message Effects on News Attractability

I offer eight hypotheses about the impact of content characteristics on the total volume of news attractability, focusing on four categories of message features discussed in Chapter 2: informational utility, negativity bias, emotional evocativeness, and novelty.

Informational Utility and Attractability

H1-1: Articles that present efficacy information in their teasers will be more frequently viewed than those without efficacy information.

H1-2: Articles whose teasers provide more useful content will be more frequently viewed.

Negativity Bias and Attractability

H2-1: Articles whose teasers evoke more negative emotional responses will be more frequently viewed.

H2-2: Articles whose teasers contain more negative emotion words will be more frequently viewed.

H2-3: Articles whose teasers provide more controversial content will be more frequently viewed.

Emotional Evocativeness and Attractability

H3-1: Articles whose teasers evoke more emotional arousal will be more frequently viewed.

H3-2: Articles whose teasers contain more emotion words will be more frequently viewed.

Novelty and Attractability

H4: Articles whose teasers provide more novel content will be more frequently viewed.

Message Effects on News Virality

I pose nine hypotheses drawn upon the following five categories of focal message characteristics discussed in Chapter 2: informational utility, positivity bias, emotional evocativeness, novelty, and exemplification.

Informational Utility and Virality

H5-1: Articles that present efficacy information will be more frequently shared than those without efficacy information.

H5-2: Articles that provide more useful content will be more frequently shared.

Positivity Bias and Virality

H6-1: Articles that evoke more positive emotional responses will be more frequently shared.

H6-2: Articles that contain more positive emotion words will be more frequently shared.

H6-3: Articles that provide less controversial content will be more frequently shared.

Emotional Evocativeness and Virality

H7-1: Articles that evoke more emotional arousal will be more frequently shared.

H7-2: Articles that contain more emotion words will be more frequently shared.

Novelty and Virality

H8: Articles that provide more novel content will be more frequently shared.

Exemplification and Virality

H9: Articles that present exemplars will be more frequently shared than those without exemplars.

Retransmission Channels and Virality

Finally, I explore a research question as to how retransmission channels (email vs. social media [Facebook and Twitter]) impact what news goes viral.

RQ1: How do the relationships between focal message features and news virality differ between email-based and social media-based retransmissions?

Method

Measures

Dependent Variables: Total Number of Selections and Retransmissions

Aggregate behavioral data on the total number of selections (attractability) and that of retransmissions (virality) for 760 NYT health news articles were obtained using the data that the News Diffusion Tracker (NDT) collected from the NYT's Most Popular API on a real time basis. As detailed in Chapter 3, the NYT API provides information about the frequency of viewing and sharing that has happened *in the last 24 hours* as of each measurement time (i.e., the automated request from the NDT at every 15 minutes) for health news articles that have been published *no earlier than 30 days* prior to each measurement time. Given this, a measure of the total number of news selections was obtained by aggregating (summing) 30 days of viewing count data (i.e., every 24 hours) for each article: M = 54,135, SD = 92,988. Similarly, a measure of the total number of retransmissions was calculated by aggregating 30 days of retransmission count data for each article by each retransmission channel: email (M = 657, SD = 1,403), Facebook (M = 254, SD = 589), and Twitter (M = 100, SD = 157).

Figure 4-1 shows distributional characteristics of aggregate behavioral measures of total news selections and retransmissions (email, Facebook, and Twitter). Consistent with previous research on online news diffusion (Bandari, Asur, & Huberman, 2012; F. Wu & Huberman, 2007), all diffusion indicators had a long-tailed distribution, or more formally, a lognormal distribution. The Shapiro-Wilk tests for all the four diffusion indicators failed to reject the null hypothesis of lognormality (all p-values > .87).

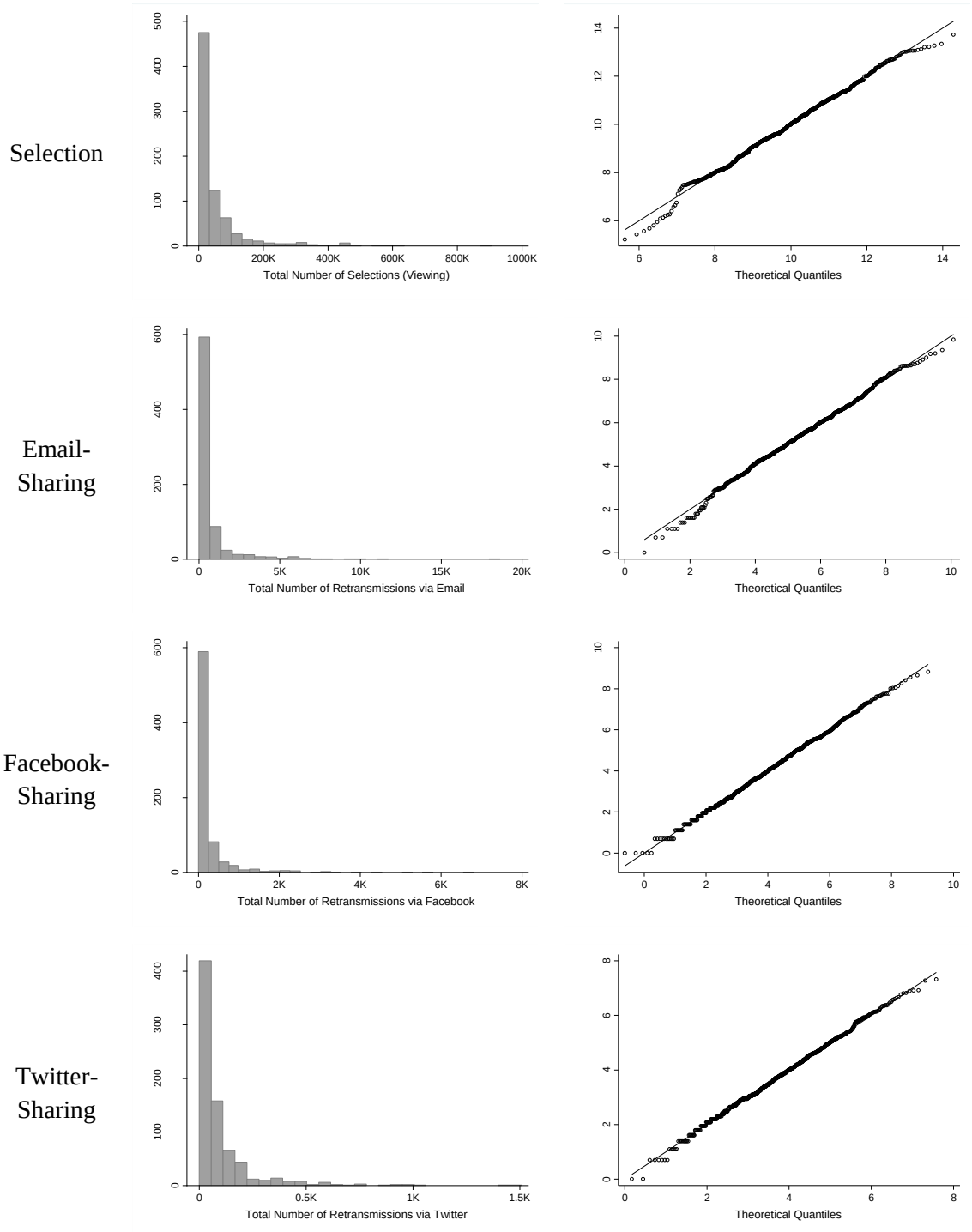

Figure 4-1. Distributional Characteristics of Viewing and Sharing Data

Graphs in the left panel are histograms of diffusion indicators, and those in the right panel are normal Q-Q plots of logged diffusion indicators (where the straight line indicates a normal probability distribution).

In other words, there was substantial inequality among the news articles in terms of their attractability and virality. The Gini coefficient, a measure of inequality (Allison, 1978) which varies between 0 (complete equality) and 1 (complete inequality), was .66 for the total number of selections, .71 for email-retransmission, .73 for Facebook-retransmission, and .61 for Twitter-retransmission.

Taken together, all diffusion indicators were *natural-log-transformed* and used in the analyses reported below.[11] The mean of the logged selection count was 9.95 (*SD* = 1.44). Descriptive analyses of the retransmission data showed a high degree of internal consistency among the three logged retransmission measures (email, Facebook, and Twitter), with a Cronbach's alpha of .95. Therefore, a measure of the total number of retransmissions was obtained by summing the three retransmission measures and taking the logarithm of the summed data (*M* = 5.86, *SD* = 1.48). For the analysis pertaining to RQ1, I used a logged email-retransmission variable (*M* = 5.34, *SD* = 1.57) and a logged summative measure of the Facebook- and Twitter-retransmission data (i.e., news sharing via social media; *M* = 4.82, *SD* = 1.46).[12] As with their sub-measures, both the total retransmission variable and the social-media-retransmission variable followed a lognormal distribution, with the *p*-values from the Shapiro-Wilk tests being greater than .87 (see Figure 4-2). The Gini coefficient was .69 for both variables.

---

[11] Throughout this dissertation, all logarithmic transformations were conducted using *natural* logarithm.

[12] Before taking the natural logarithm of the social-media-retransmission variable, a constant of '1' was added to the original data to make sure that observed zero frequencies are transformed to zeros in the corresponding logged variable. Throughout this dissertation, the same method was used for all log-transformed variables whose original data include zero scores.

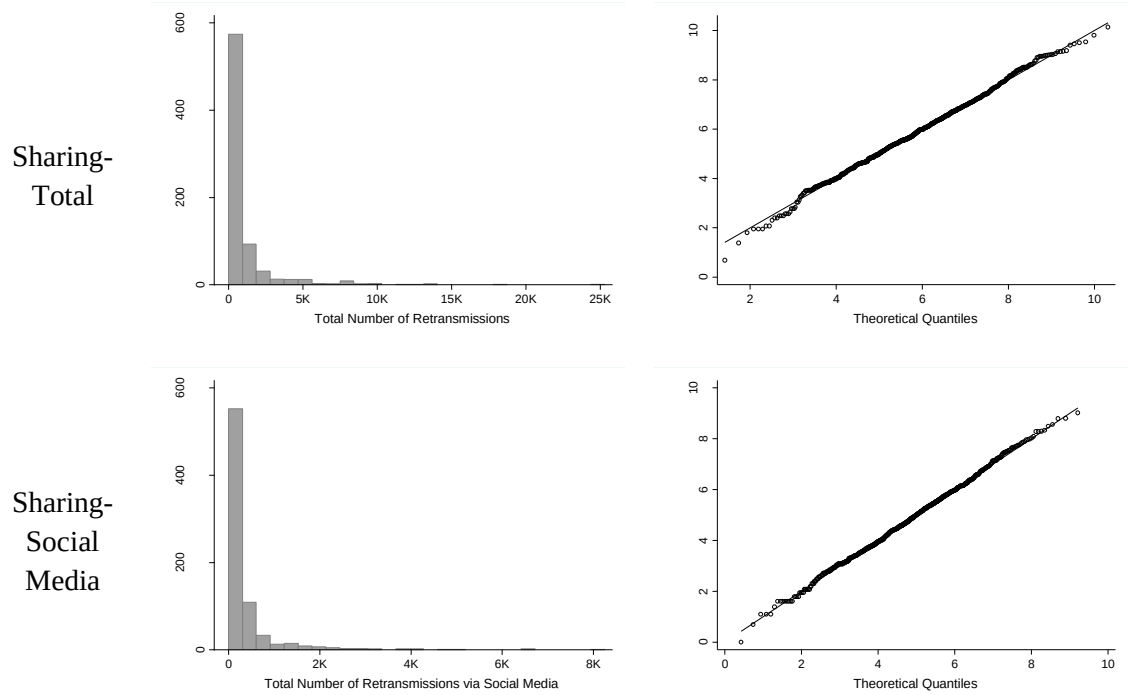

Figure 4-2. Distributional Characteristics of Total- and Social-Media-Sharing Data
Graphs in the left panel are histograms of diffusion indicators, and those in the right panel are normal Q-Q plots of logged diffusion indicators (where the straight line indicates a normal probability distribution).

As detailed in Chapter 3, the retransmission data obtained via the NYT's Most Popular API cover only news-sharing behaviors that take place on the NYT website; therefore, the frequency of retransmissions measured by the NYT API is a lower bound on their actual frequency. This might have resulted in the pattern (found in the NYT API data) of larger numbers of email-sharing than that of Facebook- or Twitter-sharing (reported earlier in this section). Therefore, a concern may arise as to the measurement validity of the social media-based retransmission data collected by the NYT API.

To address this issue, I conducted a post-hoc data collection of Facebook- and Twitter-related retransmission count using publicly available social media APIs (Facebook API and Topsy API, respectively) which keep track of a wider range of news-sharing behaviors than the NYT API (see Chapter 3 for details about this method). As

expected, the number of retransmissions collected through the social media APIs was much larger than its counterpart obtained via the NYT API and it was also larger than the email-based retransmissions: $\text{Facebook}_{\text{Facebook\_API}}$ (M = 726, SD = 1,786), $\text{Twitter}_{\text{Topsy\_API}}$ (M = 569, SD = 1,256). The two measures also followed a lognormal distribution (p-values from the Shapiro-Wilk tests > .87), and showed substantial inequality among the 760 articles (The Gini coefficient = .72 for $\text{Facebook}_{\text{Facebook\_API}}$ and .55 for $\text{Twitter}_{\text{Topsy\_API}}$).

Like when using all retransmission measures obtained from the NYT API, there was a high internal consistency among the logged email-retransmission measure (collected through the NYT API) and the two logged social media-based retransmission measures (collected through the social media APIs): α = .89. The retransmission measures collected via the NYT API were highly correlated with the corresponding measures from the social media APIs: r = .89, p < .001 between the logged $\text{Facebook}_{\text{NYT\_API}}$ and the logged $\text{Facebook}_{\text{Facebook\_API}}$; r = .83, p < .001 between the logged $\text{Twitter}_{\text{NYT\_API}}$ and the logged $\text{Twitter}_{\text{Topsy\_API}}$. Consequently, as shown in Figure 4-3, both the total retransmission measure (i.e., logged sum of the $\text{Email}_{\text{NYT\_API}}$, $\text{Facebook}_{\text{Facebook\_API}}$, and $\text{Twitter}_{\text{Topsy\_API}}$) and the social-media retransmission measure (i.e., logged sum of the $\text{Facebook}_{\text{Facebook\_API}}$ and $\text{Twitter}_{\text{Topsy\_API}}$) were strongly associated with their counterpart measures obtained via the NYT API: r = .95, p < .001 and r = .92, p < .001, respectively. More importantly, the study findings reported below were similar to those from analyses using the Facebook- and Twitter-based retransmissions collected through the social media APIs.

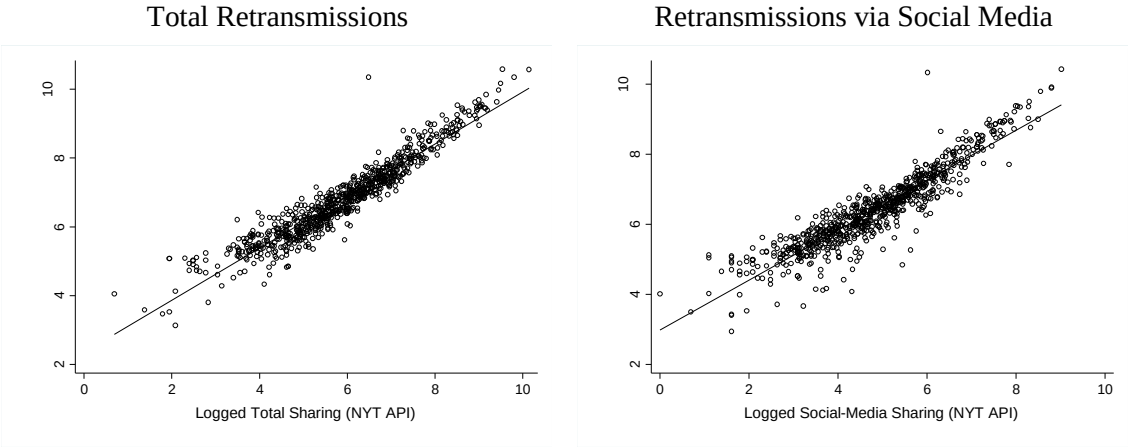

Figure 4-3. Comparison of Retransmission Data: NYT API and Social Media APIs
Graphs are scatterplots with a linear fit (solid straight line). Note that the total retransmission data for the social media APIs include the email-based retransmissions obtained from the NYT API.

Human-Coded Variables

Article teaser. As discussed in Chapter 3, a content analysis was conducted separately for article titles and abstracts, which compose article teasers. Thus, I combined message variations in titles and abstracts to quantify those in teasers. The content analysis revealed that of the 760 NYT health news articles, 150 (19.7%) contained efficacy information, 555 (73.0%) mentioned professional sources, and 432 (56.8%) mentioned diseases or bad health conditions in their teasers.

Article full text. Of the 760 article full texts, 188 (24.7%) presented efficacy information, 207 (27.2%) contained exemplars and 147 (19.3%) were written in the first-person point of view. The content analysis further showed that 59 out of the 760 full texts (7.8%) provided no credibility statements, 222 (29.2%) provided one, 397 (52.2%) provided two or more credibility statements but without opposing statement(s), and 82 (10.8%) provided two or more credibility statements along with opposing statement(s). Articles' topical area was distributed as follows: "health policy / health care system" (n = 107; 14.1%), "public health" (n = 34; 4.5%), "diseases and health conditions" (n = 523;

68.8%), "global news" (*n* = 15; 2.0%), "other (none of the above)" (*n* = 81; 10.7%). Given the small number of cases that were coded as being in the "public health" and "global news" topical areas, these categories were combined with the "other" category; this recoded article-topic variable was used in the analyses reported below.

Computer-Coded Variables

Article teaser. The 760 health news articles had on average about 33.26 words in their teasers (*SD* = 7.42). Given this small word count, LIWC-measured message variations (e.g., positive emotion words) were analyzed in terms of the raw *number* of words rather than the *proportion* (percentage) of words, the latter of which is LIWC's default metric (proportion data tend to be unreliable when the denominators – total word count in this case – are small). Building on the operationalization used in a previous study (Berger & Milkman, 2012), the expressed emotional valence (positivity) of article teasers was measured by the word-count difference in positive and negative emotion words (*M* = −.13, *SD* = 1.66), while the expressed emotionality was quantified as the total number of positive and negative words used in the teasers (*M* = 1.81, *SD* = 1.57). Of the 760 article teasers, 683 (89.9%) had no death-related words, 55 (7.2%) had one, 20 (2.6%) had two, and 2 (0.3%) had three such words. Thus, the original word-count variable was recoded to indicate the presence or absence of death-related words (i.e., present in 77 article teasers [10.1%]). There were, on average, 1.99 health-related words (*SD* = 1.67) and 2.24 social-processes-related words (*SD* = 1.93). The mean number of words with more than six letters in article teasers (i.e., an indicator of writing complexity) was 9.29 (*SD* = 3.34).

Article full text. The average word-count of the 760 article full texts was 796.29 (*SD* = 385.15). Given this substantial word count, LIWC's default percentage metric was used to quantify word-level message variations in the full texts (i.e., % over the total number of words). The average of the expressed emotional valence (positivity) of article full texts, calculated as the percentage difference between positive and negative emotion words (% positive emotion words − % negative emotion words; Berger & Milkman, 2012), was .12 (*SD* = 1.73). The mean of the expressed emotionality of article full texts (% positive emotion words + % negative emotion words; Berger & Milkman, 2012) was 3.88 (*SD* = 1.53). As with the case of article teasers, the measure of death-related words was dichotomized because nearly half of the article full texts contained no such words; these words were present in 397 articles (52.2%). The average percentage of health-related words was 4.22 (*SD* = 2.29) and that of social processes-related words was 8.07 (*SD* = 3.13). The average percentage of words longer than six letters, a measure of writing complexity, was 26.30 (*SD* = 4.05).[13] Article full texts included, on average, about 6.53 hyperlinks (*SD* = 4.55).

Variables Obtained from Context Information and Content Metadata

Using the context information collected via the News Diffusion Tracker (NDT), I measured the *total* amount of time (hours) that health news articles appeared in prominent locations on the main page of the NYT website's Health section (top six positions on the upper-left-hand corner of the page; an editorial cue to news values): *M* =

[13] While the writing complexity of article full texts was measured by the use of words greater than six letters, it can also be quantified using other tools such as the Flesch reading ease test (with a lower readability score indicating a greater writing complexity; Flesch, 1948). As expected, the two measures were highly negatively correlated: $r = -.84$, $p < .001$.

17.06, *SD* = 18.92. The amount of time that articles were shown in prominent locations also showed a long-tailed, lognormal distribution (*p*-value from the Shapiro-Wilk test > .84), and substantial inequality among the articles (Gini coefficient = .55).

The presence of images in article full texts, article column (category), and articles' seasonal variations were measured using the content metadata collected via the NDT. An examination of the image URLs embedded in article full texts indicated that, of the 760 articles, 239 (31.4%) contained no images, 493 (64.9%) contained one, 26 (3.4%) contained two, and 2 (0.3%) contained three images. Given the distribution, the original variable was dichotomized: images were present in 521 article full texts (68.6%). Of the 760 articles, 544 appeared in 20 different "columns" (or "categories") that are assigned by the NYT, such as *Well*, *The New Old Age*, *Mind*, and *News Analysis*. Of the article columns, *Well* and *The New Old Age* were predominant, with the number of articles for the two columns being 388 (51.1% of the 760 articles) and 101 (13.3%), respectively. Articles assigned to none of the 20 columns (n = 216) and those assigned to columns other than *Well* or *The New Old Age* (n = 55) were coded as "other" (n = 271; 35.7%). Regarding seasonal or periodic variations in the 760 articles, I measured the articles' (1) publication month and (2) publication day of the week using the online publication timestamp collected by the NDT. Of the 760 articles, 68 (8.9%) were published online in July 2012, 94 (12.4%) in August 2012, 96 (12.6%) in September 2012, 103 (13.6%) in October 2012, 103 (13.6%) in November 2012, 93 (12.2%) in December 2012, 117 (15.4%) in January 2013, and 86 (11.3%) in February 2013. A total of 354 out of the 760 articles (46.6%) were published online on Mondays, 63 (8.3%) on Tuesdays, 125 (16.4%) on Wednesdays, 118 (15.5%) on Thursdays, 67 (8.8%) on Fridays, 17 (2.2%) on

Saturdays, and 16 (2.1%) on Sundays. Given its distribution, the message variation in the articles' online publication day of the week was recoded as follows: Mondays (n = 354; 46.6%), other weekdays (n = 373; 49.1%), and weekends (n = 33; 4.3%).

Human-Rated Variables

Subjective (or perceived) message features were measured using data from the message evaluation survey described in Chapter 3. For each rating item (e.g., usefulness), survey respondents' evaluations were aggregated (averaged) across the respondents by article. The same set of items was used for evaluating article teasers and full texts. Emotion-related items were evaluated on a 5-point scale ranging from "not at all" (= 1) to "extremely" (= 5), while other items were rated on a 5-point scale ranging from "strongly disagree" (= 1) to "strongly agree" (= 5). The average number of respondents per article was 20.1 (SD = 4.5) for teasers, and 20.1 (SD = 4.4) for full texts. Note that 759 out of the 760 NYT health news articles were evaluated due to an unexpected programming error in the website for the message evaluation survey, as reported in Chapter 3.

Article teaser. Descriptive statistics of the aggregate responses to discrete emotion items were as follows: *pride* (M = 1.40, SD = .35), *amusement* (M = 1.62, SD = .43), *contentment* (M = 1.57, SD = .36), *hope* (M = 1.86, SD = .65), *anger* (M = 1.55, SD = .52), *fear* (M = 1.64, SD = .49), *sadness* (M = 1.82, SD = .62). A scale of emotional valence (positivity) was created by averaging these items (after reverse-scoring the *anger*, *fear*, and *sadness* items): α = .87, M = 2.78, SD = .38. The mean of the aggregate emotional arousal was 1.46 (SD = .22). The perceived novelty scale was based on three items: *newness* (M = 3.22, SD = .48), *unusualness* (M = 2.96, SD = .50), and *surprise* (M = 2.14, SD = .45). The novelty scale was constructed by averaging the three items: α

= .85, *M* = 2.77, *SD* = .42. The average of controversiality was 2.93 (*SD* = .57), and that of usefulness was 3.43 (*SD* = .44).

Article full text. As with article teasers, a scale of emotional valence (positivity) was based on the following discrete emotion items: *pride* (*M* = 1.61, *SD* = .46), *amusement* (*M* = 1.69, *SD* = .47), *contentment* (*M* = 1.82, *SD* = .44), *hope* (*M* = 2.26, *SD* = .68), *anger* (*M* = 1.72, *SD* = .62), *fear* (*M* = 1.84, *SD* = .55), *sadness* (*M* = 2.19, *SD* = .73). The emotional valence scale was constructed by averaging these items (with the *anger*, *fear*, and *sadness* items reverse-scored): α = .87, *M* = 2.80, *SD* = .43. The average of emotional arousal was 1.50 (*SD* = .23). As with article teasers, novelty-related items included: *newness* (*M* = 3.37, *SD* = .44), *unusualness* (*M* = 2.96, *SD* = .48), and *surprise* (*M* = 2.40, *SD* = .43). A novelty scale was created by averaging these items: α = .84, *M* = 2.91, *SD* = .39. The average aggregate evaluations for the perceived controversiality and usefulness were 2.95 (*SD* = .59) and 3.84 (*SD* = .34), respectively.

Analysis

Multiple linear regression models were estimated using the ordinary least squares (OLS) method to test hypotheses related to news attractability (H1 to H4) and virality (H5 to H9). Specifically, the multiple regression model of news attractability can be written as

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \text{(Equation 4-1)}$$

where $\log y_i$ is the logged total number of selections for article *i*, β_0 is the intercept, x_{ik} indicates explanatory variables for article *i* (e.g., message features of article teasers), and ε_i is the error term for article *i* that represents the combined effect on $\log y_i$ of all factors other than the observed x_{ik} variables (Wooldridge, 2009).

Similarly, the multiple regression model of news virality can be expressed as

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma \log z_i + \varepsilon_i \qquad \text{(Equation 4-2)}$$

where $\log y_i$ is the logged total number of retransmissions for article *i* (i.e., the logged sum of retransmissions via email, Facebook, and Twitter), β_0 is the intercept, x_{ik} denotes predictors (e.g., message features of article full texts). As with the news attractability model, the error term ε_i represents all factors other than the right-hand side variables in Equation 4-2 that affect $\log y_i$ (Wooldridge, 2009).

Note that Equation 4-2 includes the logged total number of selections, $\log z_i$, as a predictor variable and estimates its effect (γ) on the logged total retransmission-frequency, $\log y_i$. As discussed in Chapter 2, this is a crucial part of model specification when predicting virality using observational data where the sheer number of news retransmissions is confounded by the number of news selections (see Godes et al., 2005). To address the same issue, alternatively, one may create a proportion variable by dividing the total number of retransmissions by that of selections, taking its logarithm ($= \log \frac{y_i}{z_i}$), and regressing it on the right-hand side variables in Equation 4-2 (except $\log z_i$). This alternative specification is a special case of the one shown in Equation 4-2, in the sense that it is formally equivalent to constraining γ to be 1 in Equation 4-2. By constraining γ to be 1 and moving $\log z_i$ to the left-hand side, Equation 4-2 can be rearranged as below:

$$\log y_i - \log z_i = \log \frac{y_i}{z_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \text{(Equation 4-3)}$$

In sum, the crucial difference between the alternative model and the one employed in this dissertation is whether γ is constrained to be equal to 1 or is freely estimated. Given the lack of empirical evidence about γ, I opted to estimate γ rather than making a strong constraint on its parameter.

Structural equation modeling (SEM; Bollen, 1989; Kline, 2010) was used to answer RQ1, which addresses how message features differentially relate to (1) news propagations via email ($\log y_{i1}$) and (2) those via social media ($\log y_{i2}$). As shown in Figure 4-4, a structural model was constructed with the following specifications: (1) two dependent variables (i.e., $\log y_{i1}$ and $\log y_{i2}$) are regressed on the same predictors shown in Equation 4-2, (2) exogenous (predictor) variables are correlated, and (3) the residuals (errors) of the two dependent variables are allowed to be correlated (i.e., estimating a partial correlation between the dependent variables after controlling for the common predictors). This specification makes the structural model just-identified (i.e., fully saturated model). Multiple equations included in the model were estimated simultaneously (Bollen, 1989; Kline, 2010).

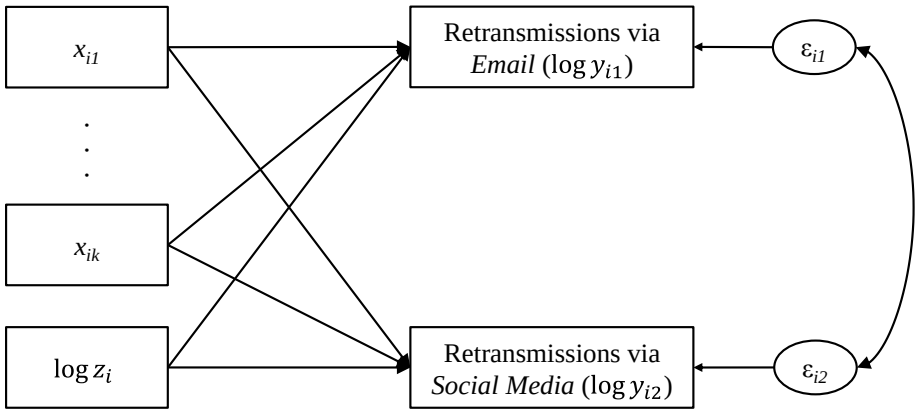

Figure 4-4. Message Effects on News Virality by Retransmission Channels
Definitions and notations of exogenous (predictor) variables (i.e., $x_{i1} \cdots x_{ik}$, $\log z_i$) are the same as those in Equation 4-2. Correlations among exogenous variables are included in the model but not shown here for brevity.

SEM was preferred to estimating separate regression models for the two dependent variables because it takes into consideration the residual correlation between the dependent variables using the full covariance matrix. Moreover, the SEM approach

which analyzes covariance structure makes it possible to statistically compare the effects of message features on the two dependent variables, which is central to answering RQ1.

Parameters in Equations 4-1 and 4-2 were estimated using the OLS method (Wooldridge, 2009), and those of the SEM were estimated using the full information maximum likelihood method (Bollen, 1989; Kline, 2010). In addition to total selection- and retransmission-frequency variables, the following predictor variables were log-transformed in all analyses reported below because of their distributional characteristics: the number of hours shown in prominent locations, the total number of positive and negative emotion words (for article teasers only), words related to health and social processes, and the number of hyperlinks embedded in article full texts.

Throughout the linear regression models examined in this chapter, unstandardized coefficients are reported. Missing data were handled with listwise deletion (Allison, 2002; Enders, 2010).[14]

Results

Predicting News Attractability

Table 4-1 presents results from bivariate and multiple OLS regression analyses of the total volume of news attractability. The final multiple regression model (Model 2) included an interaction effect between (1) the emotional positivity evoked by article

[14] There were two sources of missing data with respect to the analyses of the statistical models shown in Equations 4-1, 4-2, and Figure 4-4. First, as mentioned in Chapter 3, one of the 760 health news articles was dropped from the message evaluation survey due to an unexpected technical error. Thus, all variables related to perceived message features were missing for the article. Second, the NYT API provided no information about the number of selections (views) for another article.

teasers and (2) the mention of diseases or bad health conditions in the teasers. The final model explained about 37% of the total variance in attractability with the content factors explaining about 17% and the context factors accounting for about 20%. Results reported below are based on Model 2 unless otherwise noted.

Informational Utility and Attractability

The results revealed that consistent with H1-1, health news articles presenting efficacy information in their teasers were more frequently viewed by readers than those without such information, unstandardized b = .34, 95% CI [.09, .59]. However, the perceived usefulness of article teasers was not predictive of attractability. Thus, H1-2 was rejected.

Negativity Bias and Attractability

H2-1 which predicted a positive association between (1) the negativity of emotional responses induced by teasers and (2) attractability was not supported by the results (Model 1 in Table 4-1). Rather, the effect of the valence of evoked emotions was moderated by the mention of diseases or bad health conditions (Model 2 in Table 4-1).

Table 4-1. Message Effects on News Attractability

	Bivariate Regression	Multiple Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Content Factors (*df* = 18)			$\Delta R^2 = .17^{***}$
Efficacy Information Present	$.35^{**}$ (.13)	$.30^{*}$ (.13)	$.34^{**}$ (.13)
Usefulness	−.01 (.12)	.03 (.11)	.02 (.11)
Emotional Positivity (Responses)	$.29^{*}$ (.14)	.11 (.15)	$.65^{**}$ (.22)
Expressed Positivity (Words)	.02 (.03)	−.01 (.03)	−.01 (.03)
Controversiality	.14 (.09)	$.18^{*}$ (.09)	$.25^{**}$ (.09)
Emotional Arousal (Responses)	$.74^{**}$ (.23)	.31 (.20)	.31 (.20)
Expressed Emotionality (Words) [a]	$.23^{*}$ (.09)	$.16^{+}$ (.08)	$.16^{*}$ (.08)
Novelty	$-.22^{+}$ (.12)	−.19 (.12)	$-.23^{*}$ (.12)
Diseases / Bad Health Conditions Mentioned	$-.26^{*}$ (.11)	$-.30^{**}$ (.11)	$-.27^{*}$ (.11)
Positivity (Responses) × Diseases			$-.85^{***}$ (.25)
Professional Sources Mentioned	$-.33^{**}$ (.12)	$-.28^{**}$ (.10)	$-.27^{**}$ (.10)
Death-Related Words Present	.02 (.17)	.08 (.15)	.04 (.15)
Health Words [a]	$.19^{*}$ (.09)	−.01 (.08)	.02 (.08)
Social-Processes Words [a]	$.17^{*}$ (.08)	.03 (.07)	.03 (.07)
Word Count	$.03^{***}$ (.01)	.01 (.01)	.01 (.01)
Writing Complexity (Words > 6 Letters)	$.04^{*}$ (.02)	.01 (.02)	.01 (.01)
Context Factors (*df* = 10)			$\Delta R^2 = .20^{***}$
Total Hours Shown in Prominent Locations [a]	$.48^{***}$ (.03)	$.44^{***}$ (.03)	$.44^{***}$ (.03)
Final Model R^2		$.36^{***}$	$.37^{***}$

Note. *N* = 758 for the multiple regression models (Model 1 & 2). Dependent variables were log-transformed. Cell entries are unstandardized OLS regression coefficients (*b*) with standard errors (*se*) in parentheses. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix B). Emotional Positivity (Responses) was mean-centered before entry (Model 2). All variance inflation factors (VIFs) for Model 2 < 2.35. [a] Log-transformed. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

As shown in Figure 4-5, news articles whose teasers evoked more positive emotional responses were more frequently selected when there was no mention of diseases or bad health conditions in the teasers, $b = .65$, 95% CI [.22, 1.07]. When article teasers mentioned diseases or bad health conditions, however, emotional valence was not statistically significantly associated with attractability, $b = -.20$, 95% CI [−.54, .14].

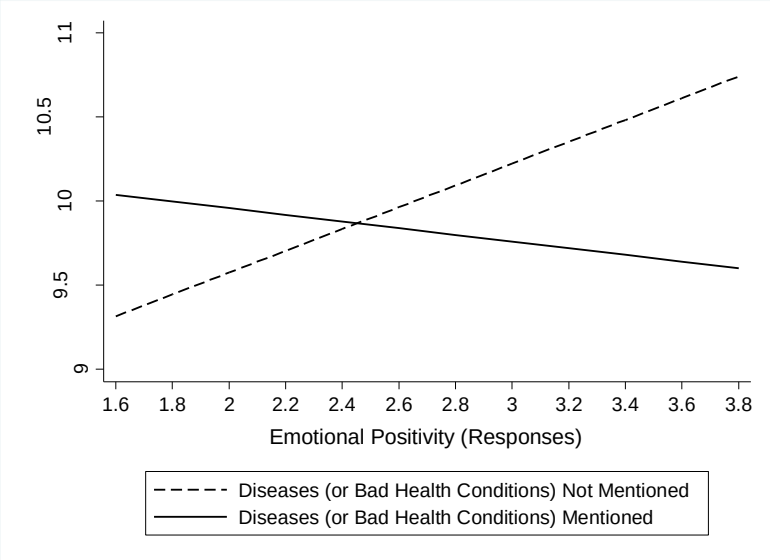

Figure 4-5. The Emotional Positivity (Responses) × Mention of Diseases Interaction Effect on News Attractability

Values in Y-axis are predicted logged total number of selections that are adjusted for explanatory variables in the regression model (Model 2 in Table 4-1).

Inconsistent with H2-2, expressed emotional valence was not significantly associated with attractability. A significant impact of negativity bias on attractability was found for controversiality (which, as expected, was negatively correlated with the positivity of emotional responses, $r = -.39$, $p < .001$). Articles with more controversial teasers were more frequently selected, $b = .25$, 95% CI [.06, .43], providing support for H2-3.

Emotional Evocativeness and Attractability

The relationship between (1) the level of emotional arousal induced by article teasers and (2) attractability was not statistically significant, rejecting H3-1. However, consistent with H3-2, there was a significantly positive association between expressed emotionality and attractability, $b = .16$, 95% CI [.003, .32].

Novelty and Attractability

In contrast to the prediction made by H4, the results revealed a significantly negative relationship between novelty and attractability, $b = -.23$, 95% CI [−.47, −.001].

Control Variables

The results revealed a significant effect of an editorial cue to news values on attractability, such that the longer health news articles were displayed in prominent locations on the main page of the NYT's Health section (in hours), the more frequently they were viewed, $b = .44$, 95% CI [.37, .51].

There was an overall tendency for news articles mentioning diseases or bad health conditions in their teasers to be less frequently viewed than those without such terms (Model 1 in Table 4-1). While the mention of diseases or bad health conditions interacted with the valence of emotional responses (as described earlier), it was negatively associated with attractability overall (Model 2 in Table 4-1): the unstandardized coefficient (*b*) for its simple main-effect term (i.e., when the emotional valence variable was held at its mean) was −.27, 95% CI [−.48, −.06]. The results also revealed that articles mentioning professional sources in their teasers invited a smaller number of selections than those not mentioning such sources, $b = -.27$, 95% CI [−.47, −.07].[15]

[15] With respect to the final multiple regression model of the total volume of news attractability (Model 2 in Table 4-1), there are two additional variables that may affect the total number of selections: topical area and the presence of visual images. These two variables were excluded from the main analyses reported above because they were measured by content-analyzing article full texts. However, one may posit that the topical area of a full article can be noticed (or guessed) by NYT readers, based solely on the article's teaser text. Regarding the presence of images, the NYT does provide thumbnail images along with teaser texts for some articles in some interfaces or contexts, about which this dissertation has no data. However, one may assume that articles' teaser texts are more likely to be presented with a thumbnail image if their full texts contain one or more image(s) – the articles whose full texts include no image cannot be presented with a thumbnail image along with their teaser texts. Taken together, I added the two variables (i.e.,

Predicting News Virality

Results from bivariate and multiple OLS regression analyses of the total volume of news virality are shown in Table 4-2. The multiple regression model explained about 86% of the total variance in virality. About 49% of the total variance was explained by content factors, about 5% by context factors, and additionally, about 32% by a single factor, the logged total number of news selections.

Informational Utility and Virality

Consistent with H5-1, health news articles presenting efficacy information were more frequently shared than those with no such information, $b = .13$, 95% CI [.02, .24]. Articles providing more useful content were more frequently retransmitted, $b = .50$, 95% CI [.36, .64], supporting H5-2.

Positivity Bias and Virality

The results also supported H6-1 which predicted a positive relationship between the positivity of emotional responses and news retransmissions, $b = .19$, 95% CI [.07, .32]. Inconsistent with H6-2, news virality was not significantly associated with expressed emotional valence. Article controversiality (which was negatively associated with the positivity of emotional responses, $r = -.42$, $p < .001$) was not predictive of virality, rejecting H6-3.

---

article topic and the presence of images in article full texts) to the final regression model of news attractability and re-analyzed the data. Results in Table 4-1 remained virtually unchanged with these additional covariates. Findings pertaining to the newly added variables were as follows. First, there was a significant association between article topic and attractability, F (2, 726) = 4.96, $p < .01$. Articles related to "diseases and health conditions" tended to be more frequently viewed than those about "health policy and health care system," $b = .27$, 95% CI [−.02, .57], $p = .07$, and "other" articles, $b = .41$, 95% CI [.13, .69], $p < .01$. Second, the presence of images in article full texts (i.e., a proxy indicator of the presence of thumbnail images in article teasers) had a marginally significantly positive effect on news attractability, $b = .26$, 95% CI [−.02, .53], $p = .07$.

Table 4-2. Message Effects on News Virality

	Bivariate Regression	Multiple Regression
	b (se)	*b (se)*
Content Factors (*df* = 25)		$\Delta R^2 = .49^{***}$
Efficacy Information Present	$.63^{***}$ (.12)	$.13^{*}$ (.06)
Usefulness	$.99^{***}$ (.15)	$.50^{***}$ (.07)
Emotional Positivity (Responses)	$.57^{***}$ (.12)	$.19^{**}$ (.06)
Expressed Positivity (Words)	$.09^{**}$ (.03)	.01 (.01)
Controversiality	.11 (.09)	−.01 (.05)
Emotional Arousal (Responses)	$.95^{***}$ (.23)	.10 (.10)
Expressed Emotionality (Words)	$.08^{*}$ (.03)	.02 (.01)
Novelty	$.35^{**}$ (.14)	.05 (.06)
Exemplification	$.44^{***}$ (.12)	.03 (.06)
Credibility Statements		
1	−.30 (.21)	−.01 (.11)
2+ with no opposing statements	$.65^{**}$ (.20)	−.05 (.11)
2+ with opposing statements	$.68^{**}$ (.24)	−.15 (.13)
Topic (Reference = Health Policy)		
Disease / Health Conditions	.16 (.16)	−.02 (.08)
Other	−.33 (.21)	−.01 (.09)
Writing Style – 1st Person Point of View	−.02 (.14)	.05 (.07)
Death-Related Words Present	−.09 (.11)	−.04 (.05)
Health Words [a]	.07 (.11)	−.01 (.05)
Social-Processes Words [a]	$.48^{***}$ (.13)	.05 (.06)
Word Count × 10^{-2}	$.19^{***}$ (.01)	$.03^{***}$ (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	−.16 (.13)	$.20^{**}$ (.07)
(Writing Complexity)^2		$-.17^{+}$ (.10)
Images Present	$.78^{***}$ (.11)	.03 (.07)
Number of Hyperlinks [a]	$.49^{***}$ (.07)	$.06^{+}$ (.04)
Context Factors (*df* = 10)		$\Delta R^2 = .05^{***}$
Total Hours Shown in Prominent Locations [a]	$.48^{***}$ (.03)	$.04^{*}$ (.02)
Selection (*df* = 1)		$\Delta R^2 = .32^{***}$
Total Number of Selections [a]	$.92^{***}$ (.02)	$.84^{***}$ (.02)
Final Model R^2		$.86^{***}$

Note. *N* = 758 for the multiple regression model. Dependent variables were log-transformed. Cell entries are unstandardized OLS regression coefficients (*b*) with standard errors (*se*) in parentheses. Writing Complexity was mean-centered. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix C). All variance inflation factors (VIFs) for the multiple regression model < 3.30. [a] Log-transformed. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Emotional Evocativeness and Virality

Emotional evocativeness, either measured in terms of the emotional arousal induced by article full texts or expressed emotionality (the use of emotion-related words), was not predictive of news retransmissions. Thus, H7-1 and H7-2 were rejected.

Novelty, Exemplification, and Virality

Neither novelty nor exemplification was significantly associated with news propagations, rejecting H8 and H9.

Control Variables

There was a significant effect of the logged total number of selections (views) on the total volume of news virality, b = .84, 95% CI [.80, .88]. An editorial cue to news values – the logged total hours an article was shown in prominent locations on the NYT Health section's main page – was also a significant predictor of news retransmissions, b = .04, 95% CI [.003, .08]. Article length was positively associated with virality, b = .03, 95% CI [.01, .05]. There was a marginally significant positive relationship between the logged number of hyperlinks and virality, b = .06, 95% CI [−.006, .13], p = .07. Finally, the results also revealed a marginally significant negative quadratic effect of writing complexity on retransmissions (i.e., an inverted U-curve pattern), b = −.17, 95% CI [−.37, .02], p = .08. News articles were more frequently shared when their level of writing complexity was moderate compared to when it was relatively low or high.

Retransmission Channels and Virality

Table 4-3 presents SEM results pertaining to RQ1 about how message features differentially relate to virality between two types of retransmission channels: email and social media (Facebook and Twitter). The structural model was just-identified (fully saturated), meaning that the model exactly reproduced the observed covariance matrix (i.e., model fit is perfect). The residual correlation between the logged email- and logged social-media-retransmissions was .44, $p < .001$, which is the partial correlation between the two variables after controlling for a common set of regressors.

An omnibus test of the null hypothesis that all coefficients are identical between the two regression equations (i.e., one for email-related virality and the other for social media-related virality) was statistically significant, Wald χ^2 (36) = 295.23, $p < .001$, indicating that overall, the coefficients differed between the two types of retransmission platforms. Focusing on the nine focal message features (i.e., variables concerning H5 to H9), an omnibus test also rejected the null hypothesis that all coefficients for the nine variables are the same between the two equations, Wald χ^2 (9) = 73.62, $p < .001$.

Specifically, the presence of efficacy information had a significantly positive effect on email-based propagations, $b = .19$, 95% CI [.07, .32], but not on those made through social media, $b = .002$, 95% CI [−.11, .12]. The difference between these two coefficients was statistically significant, Wald χ^2 (1) = 8.30, $p < .01$. The effect of the perceived usefulness on email-related virality, $b = .66$, 95% CI [.49, .82], was significantly greater than that on social media-related virality, $b = .22$, 95% CI [.07, .37], with a one degree of freedom Wald χ^2 test being 26.59, $p < .001$.

Table 4-3. Message Effects on News Virality by Retransmission Channels

	Retransmission Channel	
	Email	Social Media
	b (se)	*b (se)*
Content Factors (*df* = 25)		
Efficacy Information Present	.19** (.06)	.002 (.06)
Usefulness	.66*** (.08)	.22** (.08)
Emotional Positivity (Responses)	.17* (.07)	.26*** (.07)
Expressed Positivity (Words)	.01 (.02)	.01 (.02)
Controversiality	−.03 (.06)	.06 (.05)
Emotional Arousal (Responses)	−.02 (.11)	.34** (.11)
Expressed Emotionality (Words)	.03+ (.02)	.02 (.02)
Novelty	.17* (.07)	−.16* (.07)
Exemplification	−.005 (.06)	.12* (.06)
Credibility Statements		
1	.04 (.12)	−.06 (.11)
2+ with no opposing statements	−.03 (.12)	−.03 (.11)
2+ with opposing statements	−.11 (.15)	−.16 (.14)
Topic (Reference = Health Policy)		
Disease / Health Conditions	.01 (.09)	−.03 (.08)
Other	.03 (.11)	−.04 (.10)
Writing Style – 1st Person Point of View	.12 (.08)	−.08 (.07)
Death-Related Words Present	−.08 (.06)	−.005 (.05)
Health Words [a]	.04 (.06)	−.05 (.06)
Social-Processes Words [a]	.02 (.07)	.05 (.07)
Word Count × 10^{-2}	.05*** (.01)	.02 (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.25*** (.08)	.13+ (.07)
(Writing Complexity)^2	−.11 (.12)	−.27* (.11)
Images Present	−.04 (.08)	.11 (.07)
Number of Hyperlinks [a]	.08* (.04)	.03 (.04)
Context Factors (*df* = 10)		
Total Hours Shown in Prominent Locations [a]	.04* (.02)	.03 (.02)
Selection (*df* = 1)		
Total Number of Selections [a]	.87*** (.02)	.79*** (.02)
R^2	.83***	.83***
Residual Correlation	.44***	

Note. *N* = 758. Dependent variables were log-transformed. Cell entries are unstandardized regression coefficients (*b*) with standard errors (*se*) in parentheses. Writing Complexity was mean-centered. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix D). [a] Log-transformed. + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Coefficients for variables related to emotional valence (i.e., positivity of emotional responses and expressed positivity) did not differ between the two regression models. Effects of controversiality were also not different. Emotional arousal had a statistically significant effect on social media retransmissions, $b = .34$, 95% CI [.13, .55], but its impact on email retransmissions was not significant, $b = -.02$, 95% CI [−.24, .21], with the coefficient difference being statistically significant, Wald χ^2 (1) = 9.33, $p < .01$. Expressed emotionality did not have a differential effect on the two outcome variables.

Novelty had an *opposite* effect on email-based and social media-based news propagations, Wald χ^2 (1) = 19.60, $p < .001$. It was significantly *positively* associated with email-related virality, $b = .17$, 95% CI [.03, .31], but significantly *negatively* related to social media-related virality, $b = -.16$, 95% CI [−.29, −.03]. The difference between exemplification effects on the two news-sharing outcomes was marginally significant, Wald χ^2 (1) = 3.44, $p = .06$. While the presence of exemplars did not have a significant impact on email retransmissions, $b = -.005$, 95% CI [−.13, .12], it was significantly positively associated with social media retransmissions, $b = .12$, 95% CI [.00002, .23].

Summary

The results presented in this chapter suggest that message features play a significant role in boosting the total volume of news attractability and virality. Aggregate behavioral data on the total frequency with which NYT health news articles were (1) viewed (selected) and (2) shared through multiple online communication channels (email, Facebook, and Twitter) were measured in a natural setting. The total volume of news attractability and virality was analyzed in relation to message features while controlling

for other content characteristics and context factors. Major findings of this chapter can be summarized as follows.

The results from the message effects model of the total volume of attractability indicated that selections (views) of health news articles increased when the articles' teaser texts (1) presented efficacy information, (2) provided more controversial content, and (3) used more emotion-related words (either positive or negative), all of which support this dissertation's hypotheses. However, contrary to expectations, news articles were more frequently viewed when their teasers presented familiar (or usual) rather than novel, unusual, or surprising content. The results further revealed that articles whose teasers evoked positive emotional responses were more frequently selected than those with teasers that induced negative feelings, given that there was no mention of diseases or bad health conditions in the teasers, which is also inconsistent with the proposed hypothesis. When it comes to control variables, it is worth noting that the total hours for which articles were displayed in prominent locations on the NYT Health section's main page (i.e., an editorial cue to news values) exerted a strong impact on the total volume of attractability.

Consistent with the research hypotheses, the results from the message effects model of the total volume of virality showed that news propagations were positively associated with (1) informational utility (i.e., the presence of efficacy information and perceived article usefulness) and (2) positivity of emotional responses. For control factors, as with the message effects model of attractability, an editorial cue to news values had a significant effect on virality. Articles shown in prominent positions on the NYT's Health section for a longer period of time were more frequently retransmitted,

although the magnitude of the editorial cue effect was smaller than that observed from the attractability model (probably because the effect was largely mediated by the volume of news selections).

The results of this chapter further identified significant differences in the effects of message features on news propagations between two types of online news-sharing channels (i.e., email vs. social media). Content characteristics pertaining to informational utility were more strongly associated with email-based retransmissions than with social media-based retransmissions. The direction of the impact of novelty was opposite between email- and social media-related virality. Novel articles were more frequently shared through email, whereas familiar ones were more often circulated via social media. The results also indicated a different role played by exemplification. While the use of exemplars had no impact on email-based news retransmissions, it was significantly associated with an increase in news retransmissions via social media.

The message effects models examined in this chapter serve as baseline models of news attractability and virality throughout this dissertation, in the sense that they center on the question of how message characteristics relate to the *total* volume of news selections and retransmissions, while being mute on the temporal dynamics of cumulative processes by which the final diffusion outcomes are reached. Specifically, the message effects models leave unexamined the role of social influence (indicated by public signals about news popularity) and its interactions with message features over the course of news diffusion. More precisely, the social influence effects are *untestable* by the message effects models which predict the *total* frequency of selections and that of retransmissions. As discussed in Chapter 3, popularity information about an article (e.g., making the

"most-emailed" list) displayed on the NYT website's Health section at a certain time point is automatically generated using the article's diffusion data observed at the same time point (e.g., the frequency with which the article has been shared via email in the last 24 hours). That is, popularity information-related data about an article (e.g., amount of time shown on the "most-emailed" list) in any time interval is *contemporaneously endogenous* to the frequency of news selections or retransmissions in the same time interval. The endogeneity makes it impossible to specify the former as a causal predictor of the latter with the form of the diffusion data analyzed in this chapter.

Taken together, as an extension of the baseline message effects models, Chapter 5 proposes and tests temporal dynamics models that examine the role of public signals about news popularity and their interactions with message features in shaping subsequent news selections and retransmissions. Specifically, Chapter 5 uses a pooled time-series cross-sectional form of the NYT data to investigate the interplay of social influence and content characteristics in health news diffusion.

CHAPTER 5

TEMPORAL DYNAMICS OF ATTRACTABILITY AND VIRALITY

Overview

Chapter 4 examined baseline message effects models that center on how message features relate to the total volume of news selections and retransmissions. While the models shed light on the message-level drivers of online news diffusion, they do not address aggregate-level mechanisms that underlie the observed relationships between content characteristics and the total volume of news attractability and virality.

This chapter extends the baseline message effects models in Chapter 4 by examining temporal dynamics models of news attractability and virality that focus on the diffusion processes by which the total number of selections and that of retransmissions are reached. Specifically, the temporal dynamics models investigate how social influence and its interactions with message features shape the longitudinal processes that underlie the observed effects of the message features on the total volume of news attractability and virality over the full course of news diffusion (i.e., 30 days).

In this chapter, I first test how public signals about the popularity of New York Times (NYT) health news articles at a certain point in time (i.e., social influence cues) affect the selections and retransmissions of the articles at a later point in time (i.e., cumulative advantage effects; Muchnik et al., 2013; Salganik et al., 2006; Salganik & Watts, 2008, 2009a). Then, I examine how focal message features of this dissertation moderate the social influence-driven cumulative advantage effects on news attractability and virality. Finally, as a supplementary analysis, I explore how message features relate

to news articles' early popularity (indicated by the event of first-time appearance on the "most-viewed" or "most-emailed" list).

News attractability and virality in this chapter are indicated by the number of times that health news articles have been (1) viewed and (2) shared in a given time interval, respectively. Social influence is represented by the amount of time that news articles are displayed on the "most-viewed" (or "most-emailed") list on the NYT's website in a given time interval. Since NYT features the "most-emailed" list on its website but not the corresponding list for news sharing via social media, this chapter examines email-specific and social media-specific virality outcomes separately. The analysis of the latter virality outcome reveals how public signals about news popularity in terms of email-based news forwarding produce a carryover (or cross-channel) impact on retransmissions through social media (Facebook and Twitter).

As with Chapter 4, data on article teasers and full texts are used for the analysis of temporal dynamics of news attractability and virality, respectively. Focal message features examined in this chapter are the same as those in Chapter 4.

Hypotheses

Drawing on the review of theoretical and empirical literature in Chapter 2, I predict that social influence produces cumulative advantage effects on the temporal dynamics of news attractability and virality, such that news articles which stay longer on the "most-viewed" and "most-emailed" lists invite more frequent subsequent selections and retransmissions, respectively. I further hypothesize that the cumulative advantage effects are stronger for news stories having the focal message features of this dissertation

(i.e., synergetic interaction effects between social influence and the focal message features).

Temporal Dynamics of News Attractability

The social influence-driven cumulative advantage effect on temporal dynamics of news attractability is hypothesized as follows:

H1: The longer articles stay on the "most-viewed" list in an earlier time interval, the more frequently they will be viewed in a later time interval.

I offer a series of hypotheses which predict that the cumulative advantage effect produced by social influence (i.e., popularity information in terms of news selections) is stronger for news articles containing the focal message features of this dissertation.

H2-1: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles that present efficacy information in their teasers than those without efficacy information.

H2-2: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers provide more useful content.

H2-3: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers evoke more negative emotional responses.

H2-4: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers contain more negative emotion words.

H2-5: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers provide more controversial content.

H2-6: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers evoke more emotional arousal.

H2-7: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers contain more emotion words.

H2-8: The social influence-driven cumulative advantage effect on subsequent news selections will be stronger for articles whose teasers provide more novel content.

Temporal Dynamics of News Virality

I hypothesize that public signals about news popularity generate a cumulative advantage effect on email-based news propagation as follows:

H3: The longer articles stay on the "most-emailed" list in an earlier time interval, the more frequently they will be shared via email in a later time interval.

With respect to the interaction of social influence and message features, I predict that the cumulative advantage effect generated by the social influence cue (i.e., popularity information in terms of email-forwarding) is stronger for news articles with focal message features of this dissertation.

H4-1: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that present efficacy information than those without efficacy information.

H4-2: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that provide more useful content.

H4-3: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that evoke more positive emotional responses.

H4-4: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that contain more positive emotion words.

H4-5: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that provide less controversial content.

H4-6: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that evoke more emotional arousal.

H4-7: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that contain more emotion words.

H4-8: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that provide more novel content.

H4-9: The social influence-driven cumulative advantage effect on subsequent news retransmissions via email will be stronger for articles that present exemplars than those without exemplars.

I further posit a carryover effect of popularity information about the news-forwarding via email on subsequent news propagation through social media, based on the results from Chapter 4 that demonstrated strong associations among the retransmission measures for email, Facebook, and Twitter.

H5: The longer articles stay on the "most-emailed" list in an earlier time interval, the more frequently they will be shared via social media in a later time interval.

Finally, I hypothesize that the cumulative advantage effect on subsequent social-media retransmissions is more pronounced for news articles containing the focal message features of this dissertation.

H6-1: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that present efficacy information than those without efficacy information.

H6-2: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that provide more useful content.

H6-3: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that evoke more positive emotional responses.

H6-4: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that contain more positive emotion words.

H6-5: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that provide less controversial content.

H6-6: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that evoke more emotional arousal.

H6-7: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that contain more emotion words.

H6-8: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that provide more novel content.

H6-9: The social influence-driven cumulative advantage effect on subsequent news retransmissions via social media will be stronger for articles that present exemplars than those without exemplars.

Method

The unit of analysis and the article sample for the models of news attractability and virality in this chapter are identical to those in Chapter 4. For the analysis of the temporal dynamics models, however, I used pooled time-series cross-sectional (TSCS) data (i.e., panel data) of 760 NYT health news articles, where measures on (1) selection and retransmission, (2) social influence cues, and (3) physical locations on the main page of the NYT's Health section for each article were repeatedly recorded over a period of 720 hours (30 days). More specifically, I constructed and analyzed an article-period dataset in which each article has (1) multiple observations on the time-varying variables described above and (2) constant records on time-invariable variables (e.g., message features) across the observations, with the time metric being the article's *age* (rather than *calendar* time), an indication of the number of hours (or days) since its online publication on the NYT website (Singer & Willett, 2003). As detailed in Chapter 3 and 4, the time-varying data on NYT health news articles were kept track of up to their age of 720 hours (30 days), which yielded 547,200 observations for hourly data (= 760 articles × 720 hours) and 22,800 observations for daily data (= 760 articles × 30 days).

Research hypotheses were tested using both hourly and daily data for the following reasons. Recall that the NYT's Most Popular API provides article selection and retransmission count data observed *in the last 24 hours* as of the measurement time (see Chapter 3 for details). This indicates that there is considerable overlap between diffusion data for an article measured at any certain time point (e.g., 28 hours after online publication) and those for the article measured at adjacent time points (e.g., 30 hours after online publication). For example, an article's selection data measured at 28 hours after

its publication are based on NYT readers' news selection behaviors observed between its age of 5 hours and 28 hours, and the data for the same article measured at 30 hours after the publication reflect those observed between its age of 7 hours and 30 hours, which means that the two data actually share the same 22-hour data (i.e., the article's selection count measured between its age of 7 hours and 28 hours). Consequently, the use of hourly data may produce results that are substantially sensitive to the specification of serially correlated error structure. Therefore, I also analyzed daily data to check the sensitivity of the results from hourly data because there is no overlap among diffusion data measured at every 24 hours. While the analysis of daily data discards many observations, it tends to yield more clear-cut results.[16]

Measures

Dependent Variables

News attractability and virality were measured as the number of times that a health news article has been (1) viewed and (2) shared in a given time interval (per hour or per day) by NYT readers, respectively. As described in detail in Chapter 3, the raw time series data for each article consist of repeated observations of time-varying variables at every 15 minutes. The quarter-hourly time series data were collapsed to hourly (daily) data because an examination of a random subset of the raw data indicated that the 15-minute interval is too fine-grained to detect meaningful variations in within-article change over time for the time-varying variables, and the total duration of observation (i.e., 720 hours or 30 days) is relatively too large for the 15-minute time window.

[16] In other words, the autocorrelation among daily data points would be due to the characteristics of the process while that among hourly data points would be a function of overlapping data as well as the characteristics of the process.

As with the measures of the *total* number of selections and that of retransmissions, those observed at each time point also followed a lognormal distribution. The average number of selections for hourly data was 1,809 (*SD* = 11,670), and the average for daily data was 1,823 (*SD* = 12,588).[17] The Shapiro-Wilk tests for both measures failed to reject the null hypothesis of lognormality (both *p*-values > .77). The average number of email retransmissions was 22 (*SD* = 164) for hourly data and 22 (*SD* = 173) for daily data (both *p*-values from the Shapiro-Wilk tests > .77). Finally, the average number of social media retransmissions (Facebook and Twitter) was 12 (*SD* = 82) for hourly data and 12 (*SD* = 87) for daily data (both *p*-values from the Shapiro-Wilk tests > .77). As with Chapter 4, all dependent variables were natural-log-transformed. Figure 5-1 depicts temporal trends of logged numbers of (1) selections, (2) email retransmissions, and (3) social media retransmissions which suggest that overall the diffusion indicators are exponentially decreasing over time.

[17] Note that the mean scores are almost identical between the hourly and daily data. This is because the NYT API provides count data for audience selections that happened "in the last 24 hours" as of each observation time (see Chapter 3 for details). The same holds true for other time-varying variables.

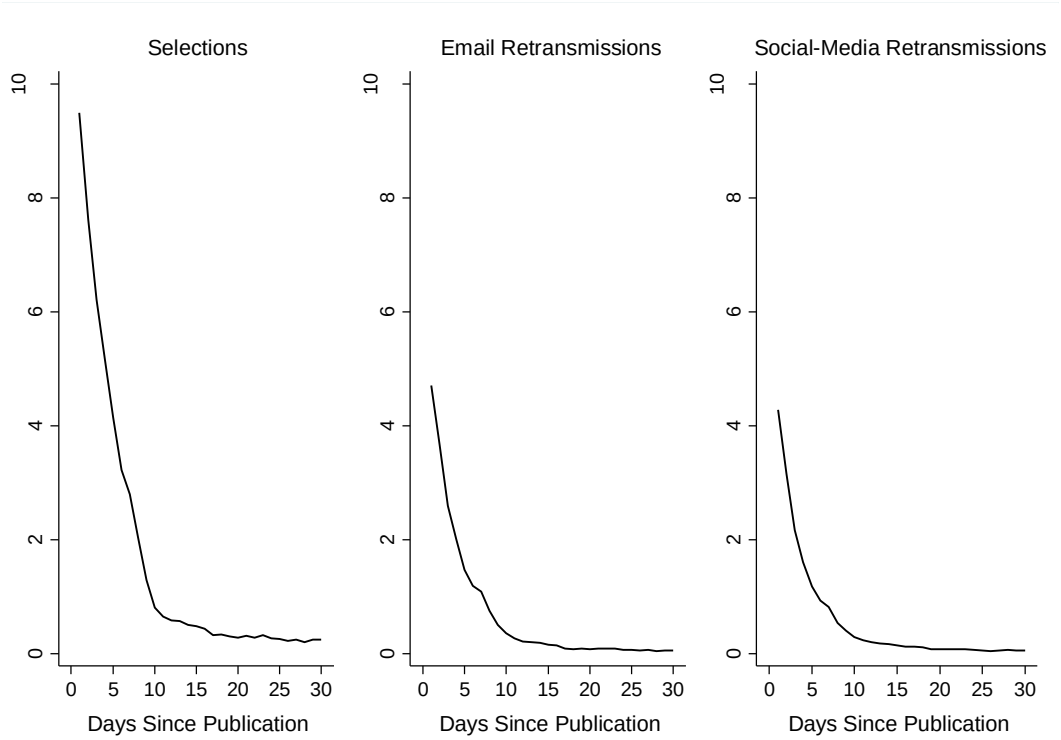

Figure 5-1. Daily Trends of News Attractability and Virality
Values are daily averages of logged number of (1) selections (Left), (2) email retransmissions (Middle), and (3) social media retransmissions (Right).

Social Influence Cues

As described in Chapter 3, the News Diffusion Tracker (NDT) collected data on whether a given article was shown on the "most-viewed" and "most-emailed" lists on the NYT website concurrently to gathering the selection and sharing count for the article. Using the raw data, the number of hours that an article was shown on each of the news popularity lists *in the last 24 hours* as of each observation time was calculated (i.e., each measure ranges from 0 to 24) to ensure that the metrics of the social influence-related variables are compatible to those of diffusion-related outcomes. The mean hours displayed on the "most-viewed" list was 1.86 (*SD* = 5.88) for hourly data and 1.86 (*SD* = 5.92) for daily data. With regard to the "most-emailed" list, the average was 1.76 (*SD* =

5.75) for hourly data and 1.76 (*SD* = 5.76) for daily data. All four measures followed a lognormal distribution and all *p*-values from the Shapiro-Wilk tests were greater than .27, and thus, were natural-log-transformed.

Message Features

Details about the measures and descriptive statistics of focal message feature variables are reported in Chapter 3 and 4.

Contextual Features

As detailed in Chapter 3, the NDT fetched data about whether an article was shown in prominent locations on the NYT Health section's main page (i.e., top six positions in the upper-left-hand corner of the page). For the same reason as for the social influence-related variables, I calculated the number of hours that an article was displayed in the prominent locations in the last 24 hours as of each measurement time using the raw data: *M* = .62, *SD* = 3.25 for hourly data; *M* = .62, *SD* = 3.42 for daily data. Due to its distributional characteristics, this variable was also log-transformed.

Analysis

Hypotheses about the role of social influence and message features in the temporal dynamics of news attractability (H1 and H2) and virality (H3 to H6) were tested by estimating fixed effects linear regression models for the pooled time-series cross-sectional (TSCS) data of 760 NYT health news articles over the 720-hour (30-day) period (Allison, 2009; Wooldridge, 2010). For the fixed effects models, I estimated standard errors that are robust to (i.e., consistent with) autocorrelated, cross-sectionally dependent, and heteroskedastic disturbance terms using the Driscoll-Kraay estimator (Driscoll & Kraay, 1998).

Fixed Effects Model

The temporal dynamics model of news attractability for testing H1 can be expressed as the following standard linear unobserved effects model (Wooldridge, 2010).

$$y_{it} = \beta_0 + \beta_1 x_{i(t-1)1} + \beta_2 x_{i(t-1)2} + \gamma_1 z_{i1} + \cdots + \gamma_k z_{ik} + \lambda_1 v_{i1} + \cdots + \lambda_m v_{im} + \theta w_t + c_i + \varepsilon_{it} \quad \text{(Equation 5-1)}$$

where y_{it} is the logged number of selections for article i at time t (t = 1, 2, ⋯, 720 for hourly data; t = 1, 2, ⋯, 30 for daily data), β_0 is the intercept, $x_{i(t-1)1}$ is the logged number of hours displayed in prominent locations on the NYT Health section's main page for article i at time $t-1$ (i.e., an editorial cue to news values; lagged), and $x_{i(t-1)2}$ indicates the logged hours shown on the "most-viewed" list for article i at time $t-1$ (i.e., a social influence cue; lagged). Focal message features of teaser text for article i are denoted by z_{ik}, and all other time-invariant variables for article i (e.g., control message features of teaser text and time-constant contextual features) are represented by v_{im}. Equation 5-1 also includes w_t, the logged article age at time t as a control for the effect of time since online publication (i.e., article age) on news selections (Leskovec et al., 2009; see also Figure 5-1). Note that there are two error terms in Equation 5-1: (1) c_i, a time-invariant error component, is an unobserved heterogeneity term which represents all unobserved (unmeasured) variables affecting y_{it}, and (2) ε_{it} is an idiosyncratic error term that changes over time t and affects y_{it}.

Adding a set of interaction terms between the lagged social influence cue and focal message features yields the following equation for testing H2.

$$y_{it} = \beta_0 + \beta_1 x_{i(t-1)1} + \beta_2 x_{i(t-1)2} + \gamma_1 z_{i1} + \cdots + \gamma_k z_{ik} + \delta_1 x_{i(t-1)2} \cdot z_{i1} + \cdots + \delta_k x_{i(t-1)2} \cdot z_{ik}$$

$$+\lambda_1 v_{i1} + \cdots + \lambda_m v_{im} + \theta w_t + c_i + \varepsilon_{it} \qquad \text{(Equation 5-2)}$$

where the product terms, $x_{i(t-1)2} \cdot z_{ik}$, represent interactions between the lagged social influence cue and focal message features of teaser text for article *i*.

To obtain unbiased coefficient estimates, I used the fixed effects (FE) estimator (Allison, 2009; Wooldridge, 2009, 2010). Details about the FE estimation procedure are described below, with the estimation of Equation 5-2 (for testing H2) as an example.

For the FE model specification, Equation 5-2 can first be rearranged as

$$y_{it} = \beta_0 + \beta_1 x_{i(t-1)1} + \beta_2 x_{i(t-1)2} + \delta_1 x_{i(t-1)2} \cdot z_{i1} + \cdots + \delta_k x_{i(t-1)2} \cdot z_{ik} + \theta w_t + \alpha_i + \varepsilon_{it} \qquad \text{(Equation 5-3)}$$

where $\alpha_i = \gamma_1 z_{i1} + \cdots + \gamma_k z_{ik} + \lambda_1 v_{i1} + \cdots + \lambda_m v_{im} + c_i$. That is, α_i represents the combined effect of *all* stable, time-invariant features of article *i*, both observed ($\gamma_1 z_{i1} \cdots \gamma_k z_{ik}, \lambda_1 v_{i1} \cdots \lambda_m v_{im}$) and unobserved ($c_i$), that affect y_{it}. The individual heterogeneity term α_i is generally referred to as a "fixed effect" in the sense that α_i is "fixed" over time *t* (Wooldridge, 2009, 2010). As with other regression models, FE model assumes that the idiosyncratic error ε_{it} is independent of everything else in the right-hand side variables in Equation 5-3. However, it allows for any correlations between α_i and time-varying predictors, which would be a reasonable specification particularly in the context of the current analysis of observational data, where, for instance, $x_{i2(t-1)}$ is not randomly assigned, but is instead an observed variable. By doing so, the FE model can estimate partial effects of the time-varying predictors while "controlling for" α_i which stands for the effects of all time-constant variables that are both observed and unobserved (Allison, 2009; Wooldridge, 2009, 2010).

Note, however, that one cannot directly control for α_i in Equation 5-3, because it is unobservable (the unobserved heterogeneity term, c_i, is a component of α_i). The FE model handles this issue by eliminating α_i in its estimation process. A consistent estimate of the FE model shown in Equation 5-3 is obtained by (1) "time-demeaning" Equation 5-3 using the "within" transformation (or, the mean deviation method), and (2) performing a pooled ordinary least squares (OLS) regression on the "time-demeaned" data (Allison, 2009; Wooldridge, 2009, 2010). Specifically, first, averaging Equation 5-3 over time *t* for each article *i* yields

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \beta_2 \bar{x}_{i2} + \delta_1 \bar{x}_{i2} \cdot z_{i1} + \cdots + \delta_k \bar{x}_{i2} \cdot z_{ik} + \theta \bar{w} + \alpha_i + \bar{\varepsilon}_i \quad \text{(Equation 5-4)}$$

where $\bar{y}_i = \frac{\sum_{t=1}^{T} y_{it}}{T}$, and the other variables in the right-hand side of Equation 5-4 similarly indicate article-specific means. Note that β_0, z_{ik}, and α_i remain the same here because they are time-constant variables. Then, if we subtract Equation 5-4 from Equation 5-3 (i.e., "time-demeaning"; calculating deviations from article-specific means), we get the following "estimating equation" of the FE model where the article-level fixed effect α_i (as well as the intercept β_0) has been removed and only time-demeaned variables are present (Wooldridge, 2009, 2010):

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{i(t-1)1} + \beta_2 \ddot{x}_{i(t-1)2} + \delta_1 \ddot{x}_{i(t-1)2} \cdot z_{i1} + \cdots + \delta_k \ddot{x}_{i(t-1)2} \cdot z_{ik} + \theta \ddot{w}_t + \ddot{\varepsilon}_{it} \quad \text{(Equation 5-5)}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, and similarly for the right-hand side variables. Because α_i has been eliminated by the within transformation, estimating Equation 5-5 using the OLS method

provides consistent estimates of the FE model specification.[18] The fact that Equation 5-5 consists of only time-demeaned variables indicates that the FE model uses each article as its "own control" (Allison, 2009, p. 1) to obtain unbiased and consistent coefficient estimates for time-varying predictors. By discarding between-article variation and focusing only on within-article variation over time, the FE model controls for all stable observed and unobserved article characteristics that can potentially confound the relationships between y_{it} and the time-varying explanatory variables (Allison, 2009; Wooldridge, 2009, 2010).

Similarly, H1 (i.e, the effect of social influence on subsequent news selections) was examined using the following estimating equation obtained by the within-transformation procedure detailed above.

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{i(t-1)1} + \beta_2 \ddot{x}_{i(t-1)2} + \theta \ddot{w}_t + \ddot{\varepsilon}_{it} \qquad \text{(Equation 5-6)}$$

The temporal dynamics model of news virality for testing H3 and H5 can be represented by the following FE model.

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{i(t-1)2} + \beta_3 x_{i(t-1)3} + \theta w_t + \alpha_i + \varepsilon_{it} \qquad \text{(Equation 5-7)}$$

[18] Performing a pooled OLS without appropriate transformations of Equation 5-3 yields biased results because OLS assumes that the composite error term (i.e., $\alpha_i + \varepsilon_{it}$) is uncorrelated with the time-variant predictors in Equation 5-3, which is inconsistent with the FE model specification that permits correlations between α_i and the time-varying variables. An alternative estimation method to the FE model in this regard is the random effects (RE) model which (1) assumes that α_i is statistically independent of (i.e., uncorrelated with) all the other explanatory variables in Equation 5-3 and (2) uses a feasible generalized least squares estimator (Allison, 2009; Wooldridge, 2009, 2010). In other words, the RE model is a special case of the FE model (Allison, 2009), in the sense that the former imposes an orthogonality restriction on the relationship between α_i and the other predictors in Equation 5-3. A Hausman test provides a statistical test comparing the FE and RE models under the null hypothesis that the orthogonality assumption imposed by the RE model is valid (Hausman, 1978). For all models analyzed in this chapter, the Hausman tests rejected the null hypothesis, suggesting that RE models should be rejected in favor of FE models.

where y_{it} is the logged number of email retransmissions for article *i* at time *t* (H3; for H5, it denotes the logged number of social media retransmissions for article *i* at time *t*), and x_{it1} is the logged number of selections for article *i* at time *t* (see Chapter 4 for the rationale for modeling the *contemporaneous* effect of news selection count on retransmission measures). The notation $x_{i(t-1)2}$ indicates the logged number of hours shown in prominent locations on the main page of the NYT Health section for article *i* at time *t*−1 (i.e., an editorial cue to news values; lagged), and $x_{i(t-1)3}$ is the logged number of hours displayed on the "most-emailed" list for article *i* at time *t*−1 (i.e., a social influence cue; lagged). The definitions of w_t and α_i are the same as those for the attractability models. Using the within transformation procedure detailed earlier, the model shown in Equation 5-7 was tested with the following estimating equation.

$$\ddot{y}_{it} = \beta_0 + \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{i(t-1)2} + \beta_3 \ddot{x}_{i(t-1)3} + \theta \ddot{w}_t + \ddot{\varepsilon}_{it} \qquad \text{(Equation 5-8)}$$

Tests of interaction effects between social influence and message features on news virality (i.e., H4 and H6) were based on the following FE model.

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{i(t-1)2} + \beta_3 x_{i(t-1)3} + \delta_1 x_{i(t-1)3} \cdot z_{i1} + \cdots + \delta_k x_{i(t-1)3} \cdot z_{ik} + \theta w_t + \alpha_i + \varepsilon_{it} \qquad \text{(Equation 5-9)}$$

where $x_{i(t-1)3} \cdot z_{ik}$ indicates the interactions between the social influence cue and focal message features of full text for article *i*. As with all models described above, the model was estimated using the following equation.

$$\ddot{y}_{it} = \beta_0 + \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{i(t-1)2} + \beta_3 \ddot{x}_{i(t-1)3} + \delta_1 \ddot{x}_{i(t-1)3} \cdot z_{i1} + \cdots + \delta_k \ddot{x}_{i(t-1)3} \cdot z_{ik} + \theta \ddot{w}_t + \ddot{\varepsilon}_{it} \qquad \text{(Equation 5-10)}$$

Robust Standard Errors: The Driscoll-Kraay Estimator

For the FE regression coefficients, I estimated robust standard errors that are consistent with autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals using the method proposed by Driscoll and Kraay (1998). The conventional FE regression models usually focus on the case of short-panel data (i.e., small number of time points T relative to sample size N) with an assumption that the model errors (i.e., ε_{it} in the equations shown earlier) are independent over time and over cross-sectional units (e.g., survey respondents). In other words, the models assume that the errors are (1) serially uncorrelated, Cor (ε_{it}, ε_{is}) = 0, where $t \neq s$, and (2) cross-sectionally (or spatially) uncorrelated, Cor (ε_{it}, ε_{jt}) = 0, where $i \neq j$. However, the present data take a form of relatively long panels with a large sample size: T = 720 for hourly data (T = 30 for daily data) and N = 760. Thus, estimating the conventional FE models for the present data would violate the assumption about the disturbance terms and threaten the validity of statistical results (Sarafidis & Wansbeek, 2011; Wooldridge, 2010). Rather, it would be reasonable to assume that the errors are both serially and cross-sectionally correlated: Cor (ε_{it}, ε_{is}) $\neq$ 0, where $t \neq s$; Cor (ε_{it}, ε_{jt}) $\neq$ 0, where $i \neq j$; respectively. An examination of the data, using the Wooldridge test for serial correlation (Wooldridge, 2010) and Pesaran's test for cross-sectional dependence (Pesaran, 2004), suggests evidence of both temporal and spatial dependence in the errors (ε_{it}) for the models analyzed below. Taken together, I estimated standard errors of the FE regression coefficients that are robust to serial autocorrelation and cross-sectional dependence in the model residuals using the Driscoll-Kraay method (1998). The Driscoll-Kraay estimator provides standard errors that are

adjusted for (i.e., consistent with) serially correlated, cross-sectionally dependent, and heteroskedastic model residuals (ε_{it}).

The number of lags for which ε_{it} is allowed to be serially correlated (= m) was specified using the following "rule of thumb" formula (Hoechle, 2007) based on the procedure proposed by Newey and West (1994).

$$m = \left[4(T/100)^{2/9}\right] \quad \text{(Equation 5-11)}$$

where T indicates the number of time points (T = 720 for hourly data, 30 for daily data), and the square brackets indicate a floor function. This heuristic method suggested ε_{it} to be autocorrelated up to (1) six lags for hourly data and (2) three lags for daily data.

Other Statistical Notes

For the FE models examined below, unstandardized regression coefficients are reported. Besides the sources of missing data reported in Chapter 4, there were occasions where the NYT's Most Popular API data were missing primarily due to unexpected connectivity issues to the NYT API (e.g., server maintenance). However, throughout the statistical analyses reported below, the maximum rate of listwise missing observations was about 1.45% (see the Results section below). Thus, as with Chapter 4, missing data were handled with listwise deletion (Allison, 2002; Enders, 2010).

Results

Predicting News Attractability

Table 5-1 presents results from fixed effects (FE) regression models of the volume of news attractability. The analyses of hourly and daily data revealed almost the same pattern of results.

Consistent with H1, the results indicated a strong lagged effect of social influence on subsequent news selections (Model 1s in Table 5-1). The longer articles stayed on the "most-viewed" list, the more frequently they were viewed at a later point in time: unstandardized $b_{\text{hourly\_data}} = 1.12$, 95% CI [1.02, 1.21], $b_{\text{daily\_data}} = .85$, 95% CI [.52, 1.18]. The results further revealed a positive interaction effect between social influence and the presence of efficacy information (Model 2s in Table 5-1), $b_{\text{hourly\_data}} = .07$, 95% CI [.02, .11], $b_{\text{daily\_data}} = .07$, 95% CI [−.006, .14], $p = .07$, providing overall support for H2-1. The social influence effect was stronger for news articles presenting efficacy information in their teasers, $b_{\text{hourly\_data}} = 1.16$, 95% CI [1.07, 1.26], $b_{\text{daily\_data}} = .90$, 95% CI [.60, 1.21], than those without such information, $b_{\text{hourly\_data}} = 1.10$, 95% CI [1.01, 1.19], $b_{\text{daily\_data}} = .83$, 95% CI [.50, 1.17]. Other focal message features, however, did not significantly moderate the social influence effect, rejecting H2-2 to H2-8.

With regard to control factors (Model 2s in Table 5-1), the results showed that the logged number of hours displayed in prominent locations on the main page of the NYT Health section (i.e., an editorial cue to news value) was positively associated with a future increase in attractability, $b_{\text{hourly\_data}} = .46$, 95% CI [.35, .57], $b_{\text{daily\_data}} = .32$, 95% CI [.08, .57]. The results revealed that attractability decayed with time (see also Figure 5-1), $b_{\text{hourly\_data}} = -1.62$, 95% CI [−1.78, −1.47], $b_{\text{daily\_data}} = -1.81$, 95% CI [−2.52, −1.11].

Table 5-1. The Impact of Social Influence and Focal Message Features on News Attractability

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Viewed List[a, b]	2.35*** (.12)	1.12*** (.05)	1.10*** (.05)	1.94*** (.21)	.85*** (.16)	.83*** (.16)
MV List × Efficacy Information [c]			.07** (.02)			.07+ (.04)
Hours Shown in Prominent Locations [a, b]	2.93*** (.12)	.46*** (.06)	.46*** (.06)	2.28*** (.20)	.32* (.12)	.32* (.12)
Time Since Online Publication [a]	−2.38*** (.14)	−1.62*** (.08)	−1.62*** (.08)	−2.74*** (.30)	−1.81*** (.34)	−1.81*** (.34)
Within R^2		.66***	.66***		.61***	.61***
N		541,095	541,095		21,995	21,995

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to six lags for hourly data and three lags for daily data. All variance inflation factors (VIFs) for Model 2 < 2.11 for hourly data (< 2.21 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Predicting News Virality: Email Retransmissions

Results from the FE models of email-based news retransmissions are shown in Table 5-2. The results did not differ by the use of hourly or daily data. As shown under Model 1s in Table 5-2, the results provide evidence for a strong lagged impact of social influence on subsequent news propagations via email, which is consistent with H3. The duration of staying on the "most-emailed" list was significantly associated with a future increase in email-related virality, $b_{\text{hourly\_data}}$ = .64, 95% CI [.61, .66], $b_{\text{daily\_data}}$ = .46, 95% CI [.39, .54].

Consistent with H4-1, the results indicated a significant interaction effect between social influence and the presence of efficacy information, such that the social influence effect was stronger for news articles presenting efficacy information in their full texts: $b_{\text{hourly\_data}}$ = .06, 95% CI [.04, .08], $b_{\text{daily\_data}}$ = .08, 95% CI [.02, .13]. The social influence effect was also greater for news articles (1) providing more useful content, $b_{\text{hourly\_data}}$ = .15, 95% CI [.11, .19], $b_{\text{daily\_data}}$ = .22, 95% CI [.11, .33], and (2) evoking more positive emotional responses, $b_{\text{hourly\_data}}$ = .05, 95% CI [.03, .08], $b_{\text{daily\_data}}$ = .09, 95% CI [.04, .14]. Thus, H4-2 and H4-3 were also supported. However, inconsistent with H4-4 to H4-9, other focal message features did not significantly alter the social influence effect on subsequent email-based news retransmissions.

Table 5-2. The Impact of Social Influence and Focal Message Features on News Virality (Email Retransmissions)

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Emailed List[a, b]	1.31*** (.07)	.64*** (.01)	.60*** (.01)	1.00*** (.12)	.46*** (.04)	.41*** (.04)
ME List × Efficacy Information [c]			.06*** (.01)			.08** (.03)
ME List × Usefulness [c]			.15*** (.02)			.22*** (.05)
ME List × Positivity (Responses) [c]			.05*** (.01)			.09*** (.02)
Selection Count [a]	.39*** (.02)	.16*** (.004)	.16*** (.004)	.39*** (.05)	.16*** (.01)	.16*** (.01)
Hours Shown in Prominent Locations [a, b]	1.57*** (.06)	.28*** (.03)	.30*** (.03)	1.18*** (.10)	.24*** (.04)	.26*** (.04)
Time Since Online Publication [a]	−1.12*** (.09)	−.34*** (.02)	−.35*** (.02)	−1.27*** (.18)	−.29*** (.05)	−.29*** (.05)
Within R^2		.82***	.82***		.74***	.74***
N		540,390	539,694		21,995	21,966

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to six lags for hourly data and three lags for daily data. All variance inflation factors (VIFs) for Model 2 < 2.45 for hourly data (< 2.47 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). * $p < .05$, ** $p < .01$, *** $p < .001$.

Combining the three significant interaction effects together, the results indicated that, as a function of the three focal message features, the unstandardized FE regression coefficients for the social influence cue ranged from .52, 95% CI [.48, .56] to .73, 95% CI [.70, .76] for hourly data, and the corresponding coefficients ranged from .29, 95% CI [.18, .41] to .60, 95% CI [.51, .69] for daily data. The lower bound coefficient ($b_{\text{hourly\_data}}$ = .52; $b_{\text{daily\_data}}$ = .29) quantifies the social influence effect for news articles (1) presenting no efficacy information, (2) providing less useful content (scored at one standard deviation [*SD*] below the mean [*M*] of the perceived usefulness variable), and (3) evoking less positive emotions (at $M - 1SD$ for the positivity rating scale). The upper bound coefficient ($b_{\text{hourly\_data}}$ = .73; $b_{\text{daily\_data}}$ = .60) indicates the social influence effect for articles (1) presenting efficacy information, (2) providing more useful content (at $M + 1SD$ for the perceived usefulness variable), and (3) evoking more positive emotions (at $M + 1SD$ for the positivity rating scale). Detailed results on the decomposition of the interaction effects are presented in Appendix E.

Time-varying control variables were also significantly associated with news propagations through email (see Model 2s in Table 5-2). The number of selections in a given time interval was positively associated with that of email retransmissions measured in the same time interval, $b_{\text{hourly\_data}}$ = .16, 95% CI [.15, .17], $b_{\text{daily\_data}}$ = .16, 95% CI [.14, .18]. There was also a lagged effect of an editorial cue to news value, such that the duration shown in prominent locations on the NYT Health section's main page was positively associated with a subsequent increase in email-related virality, $b_{\text{hourly\_data}}$ = .30, 95% CI [.24, .35], $b_{\text{daily\_data}}$ = .26, 95% CI [.18, .35]. As with the case of news

attractability, news propagations via email decreased with time (see also Figure 5-1), $b_{\text{hourly\_data}} = -.35$, 95% CI [−.39, −.31], $b_{\text{daily\_data}} = -.29$, 95% CI [−.39, −.20].

Predicting News Virality: Social Media Retransmissions

As with previous FE results reported above, FE regression analyses of social media-based news retransmissions (i.e., Facebook and Twitter) yielded an almost identical pattern of results either when using hourly or daily data (see Table 5-3). The lagged social influence cue, indicated by the logged time shown on the "most-emailed" list, had a significant carryover effect on subsequent news retransmissions via social media (see Model 1s in Table 5-3), $b_{\text{hourly\_data}} = .43$, 95% CI [.40, .45], $b_{\text{daily\_data}} = .35$, 95% CI [.30, .40]. Thus, H5 was supported.

As shown in Model 2s in Table 5-3, the results further revealed that the carryover effect of social influence was significantly stronger for news articles (1) presenting efficacy information, $b_{\text{hourly\_data}} = .02$, 95% CI [.01, .03], $b_{\text{daily\_data}} = .03$, 95% CI [.002, .07], (2) providing more useful content, $b_{\text{hourly\_data}} = .04$, 95% CI [.01, .07], $b_{\text{daily\_data}} = .09$, 95% CI [.02, .15], (3) evoking more positive emotional responses, $b_{\text{hourly\_data}} = .08$, 95% CI [.06, .10], $b_{\text{daily\_data}} = .13$, 95% CI [.09, .16], (4) using more positive emotion words, $b_{\text{hourly\_data}} = .02$, 95% CI [.01, .02], $b_{\text{daily\_data}} = .02$, 95% CI [.01, .04], and (5) presenting exemplars, $b_{\text{hourly\_data}} = .03$, 95% CI [.01, .04], $b_{\text{daily\_data}} = .05$, 95% CI [.03, .07]. These significant interaction effects provide support for H6-1 to H6-4, and H6-9. Other focal message features were not significant moderators of the social influence cue's carryover effect, which rejects H6-5 to H6-8.

Table 5-3. The Impact of Social Influence and Focal Message Features on News Virality (Social Media Retransmissions)

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Emailed List[a, b]	1.07*** (.07)	.43*** (.01)	.40*** (.01)	.82*** (.11)	.35*** (.02)	.30*** (.02)
ME List × Efficacy Information [c]			.02** (.01)			.03* (.02)
ME List × Usefulness [c]			.04* (.02)			.09* (.03)
ME List × Positivity (Responses) [c]			.08*** (.01)			.13*** (.02)
ME List × Positivity (Words) [c]			.02*** (.004)			.02** (.01)
ME List × Exemplification [c]			.03*** (.01)			.05*** (.01)
Selection Count [a]	.33*** (.02)	.15*** (.005)	.15*** (.005)	.34*** (.05)	.14*** (.01)	.14*** (.01)
Hours Shown in Prominent Locations [a, b]	1.43*** (.05)	.39*** (.03)	.40*** (.03)	1.05*** (.08)	.31*** (.04)	.33*** (.04)
Time Since Online Publication [a]	−1.00*** (.07)	−.31*** (.02)	−.31*** (.02)	−1.09*** (.18)	−.20*** (.03)	−.20*** (.04)
Within R^2		.78***	.78***		.70***	.70***
N		539,217	538,521		21,995	21,966

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to six lags for hourly data and three lags for daily data. All variance inflation factors (VIFs) for Model 2 < 3.02 for hourly data (< 3.04 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). * $p < .05$, ** $p < .01$, *** $p < .001$.

The combination of the five significant interaction effects showed that the carryover effect of social influence, indicated by the unstandardized FE coefficients for the logged hours shown on the "most-emailed" list, varied as a function of the five focal message features. Specifically, the coefficients ranged from .32, 95% CI [.29, .35] to .52, 95% CI [.49, .56] for hourly data, and the corresponding coefficients ranged from .18, 95% CI [.11, .25] to .51, 95% CI [.45, .57] for daily data (see Appendix E for detailed results). The lower bound coefficient ($b_{\text{hourly\_data}}$ = .32; $b_{\text{daily\_data}}$ = .18) measures the carryover effect of social influence for news articles (1) lacking efficacy information and exemplification, and (2) having relatively low scores on continuous focal message features (at $M - 1SD$ for perceived usefulness, positivity of emotional responses, and expressed positivity). The upper bound coefficient ($b_{\text{hourly\_data}}$ = .52; $b_{\text{daily\_data}}$ = .51) quantifies the corresponding carryover effect for articles (1) presenting efficacy information and exemplars and (2) having relatively high scores on the continuous message features (at $M + 1SD$ for perceived usefulness, positivity of emotional responses, and expressed positivity).

The results also revealed significant effects of time-varying control variables (Model 2s in Table 5-3), with patterns similar to those identified by the FE models predicting selections and email retransmissions. There was a significant and positive relationship between the logged number of selections and that of social-media retransmissions, $b_{\text{hourly\_data}}$ = .15, 95% CI [.14, .16], $b_{\text{daily\_data}}$ = .14, 95% CI [.11, .17]. Social media-related virality was positively associated with an editorial cue to news values (i.e., the lagged measure of the logged hours displayed in prominent locations on the NYT Health section's main page), $b_{\text{hourly\_data}}$ = .40, 95% CI [.34, .46], $b_{\text{daily\_data}}$ = .33,

95% CI [.25, .41]. Finally, social media-based news propagations decreased as time passed (see also Figure 5-1), $b_{\text{hourly\_data}} = -.31$, 95% CI [−.35, −.27], $b_{\text{daily\_data}} = -.20$, 95% CI [−.27, −.13].[19]

Robustness Tests

As mentioned in the Methods section, the number of lags of serial correlation for the fixed effects (FE) regression models was chosen using a heuristic formula (Equation 5-11). While this lag-length selection method is rooted in empirical research (Newey & West, 1994), it does not use information about the present data except for the number of time points and may yield a number of lags that tends to be small (Hoechle, 2007). Thus, I reexamined all the FE models with the specification of the lag-length based on the Cumby-Huizinga test for autocorrelated errors (Cumby & Huizinga, 1992), which is applicable to panel data (Baum & Schaffer, 2013). The results remained almost unchanged as compared with the alternative method used to specify a model for serial correlation in the FE regression residuals. Full results based on the alternative autocorrelation specification are presented in Appendix F.

Ancillary Analysis: Predicting Early Popularity of Health News Articles

Results from the temporal dynamics models of the volume of news attractability and virality revealed strong effects of social influence on subsequent news selections and

[19] It appears that the moderating effects of focal message features on the relationship between social influence and news propagations differ by type of retransmission channels (i.e., email vs. social media). However, statistical tests (e.g., the Wald tests used in Chapter 4) of the differences in the interaction effects are not feasible for the current data. This is because (1) the two FE models here are not based on two independent samples, and (2) unlike the structural equation model in Chapter 4, covariance structure is not considered when estimating these models.

propagations (and even stronger effects for articles containing certain message features). Health news articles that stayed longer on the "most-viewed" list were more frequently viewed at a later point in time, and those that stayed longer on the "most-emailed" list triggered more frequent subsequent social media retransmissions as well as email propagations. That is, the results suggest that once articles become initially popular and make the news popularity lists (i.e., visible to news consumers), if they stay on the lists longer, it is more likely that they become even more popular later, providing evidence for the social influence-driven cumulative advantage effects (Muchnik et al., 2013; Salganik et al., 2006; Salganik & Watts, 2008, 2009a).

With respect to the findings, a follow-up question may arise as to what predicts *early popularity* of health news articles (in terms of attractability and virality) which begets further popularity of the articles. Given the evidence of strong effects of popularity information (i.e., social influence cue) which is essentially *endogenous* (and generative or uncontrollable) to the temporal processes of news diffusion, it would be important to examine whether and how *exogenous* (and controllable) factors such as message features affect the endogenous driver of attractability and virality. One way to answer this question is to test a model which predicts from a list of exogenous factors *whether* and *when* an article makes the news popularity lists (i.e., "most-viewed" and "most-emailed" lists) *for the first time*.

Taken together, in this section, I address this research question by conducting an ancillay analysis of event occurrence (i.e., event history analysis or survival anlalysis; Allison, 2014; Singer & Willett, 2003) where an "event" indicates an article's first-time appearance on the news popularity lists over the course of their lifecycle. Focal and

control message features (i.e., time-constant variables) tested in the event history analysis are identical to those in Chapter 4.

Method

The unit of analysis is (1) the article teaser for an attractability-related event history model and (2) the article full text for a virality-related model. The article sample consists of 760 NYT health news articles. I analyzed pooled time-series cross-sectional (TSCS) data where binary indicators of (1) whether an article appears on the "most-viewed" or "most-emailed" list (i.e., dependent variables) and (2) whether an article is displayed in prominent locations on the main page of the NYT's Health section were repeatedly measured over a period of 720 hours. An article-period dataset was used with the time metric being the articles' age in terms of the number of hours since their online publication.[20] The dataset included (1) multiple observations on the time-varying binary indicators for each article and (2) a set of time-invariant variables (e.g., the article's message features) that had constant records across the observations (Allison, 2014). Unlike the temporal dynamics models, however, the current article-period dataset consisted of each article's time-series records until (1) an event occurred to the article (i.e., making the news popularity list) or (2) the article was right-censored, which means that study observations were terminated before the article experienced the event (Allison, 2014; Singer & Willett, 2003).

[20] Only hourly data were used for the event history analysis, because time-varying variables of interest here (i.e., whether an article is displayed on the popularity list and whether an article is shown in prominent locations) do not have the same measurement issue (i.e., overlapped data) as those of the temporal dynamics models. As detailed in Chapter 3, there is no overlap among the hourly-measured data used for the current event history models.

Measures

An "event" was defined as an article's first-time appearance on the "most-viewed" list (for the attractability-related event history model) or the "most-emailed" list (for the virality-related model). As detailed in Chapter 3, the News Diffusion Tracker (NDT) kept track of whether an article was shown on the news popularity lists over a period of 720 hours (after the article's online publication). Of the 760 NYT health news articles, 614 (81%) made the "most-viewed" list and 566 (74%) made the "most-emailed" list. In other words, by the end of the observation time (i.e., 720 hours after online publication), about 19% were right-censored regarding the attractability-related event, meaning that they did not experience the event of making the "most-viewed" list for the first time. About 26% were right-censored with regard to the "most-emailed" list. For the articles shown on the news popularity list, an average "age" (i.e., hours since online publication) at which they made a first-time appearance was about 10 hours for the "most-viewed" list ($n = 614$) and about 12 hours for the "most-emailed" list ($n = 566$). For news articles making the "most-viewed" list at least once during their lifetime (i.e., $n = 614$), the Pearson correlation between (1) the time (hours) to their first-time appearance on the "most-viewed" list (log-transformed) and (2) their total selection count (log-transformed; see Chapter 4) was $-.47$, $p < .001$. For articles shown on the "most-emailed" list at least once over the course of their lifecycle (i.e., $n = 566$), the correlation between (1) the logged hours to their first-time appearance on the "most-emailed" list and (2) their logged total number of email retransmissions (see Chapter 4) was $-.62$, $p < .001$.[21]

[21] It should be noted that the correlation coefficients reported here tend to be *underestimated* because the coefficients were calculated based only on articles making the news popularity list at least once while excluding those never shown on the list (i.e., right-censored cases) which tend to

A binary indicator of whether an article was displayed in prominent locations on the main page of the NYT Health section (i.e., an editorial cue to news value), another time-varying variable, was also measured at each of the 720 observation time points (see Chapter 3 for details). Of the 760 articles, 224 (29%) were never shown in prominent locations.

Time-invariant variables – message features and other contextual features – were the same as those used in Chapter 4. Details about the measures and descriptive statistics of these variables are described in Chapters 3 and 4.

Analysis

To address the research question as to the associations between message features and the event of articles' making a first-time appearance on news popularity lists, I estimated Cox regression models using the partial likelihood method which can handle right-censoring and time-varying explanatory variables (Allison, 2014; Cox, 1972; Singer & Willett, 2003). A Cox regression model for the event of first-time appearance on the "most-viewed" list can be expressed as

$$r_i(t) = r_0(t)\exp\{\beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma z_i(t-1)\} \qquad \text{(Equation 5-12)}$$

where $r_i(t)$ denotes an instantaneous rate that the event occurs to article *i* (which has not yet experienced the target event) at time *t* (Allison, 2014; Singer & Willett, 2003).[22]

---

be less frequently viewed (or emailed) than those appearing on the list. When including the right-censored cases and assigning the "age" of 720 hours (i.e., the observation end time) as time to their first-time appearance on the popularity list, the correlation was −.74 for the "most-viewed" list-related relationship and −.79 for the "most-emailed" list-related one.

[22] More formally, $r_i(t)$, the article *i*'s *continuous-time* event rate at time *t* (also known as the hazard function) can be defined as follows (Allison, 2014; Singer & Willett, 2003):

$$r_i(t) = \lim_{\Delta t \to 0} \frac{\Pr(t < T < t + \Delta t | T \geq t)}{\Delta t}$$

Equation 5-12 shows that the instantaneous event rate for article *i* at time *t* is a product of two factors: (1) a baseline event rate, $r_0(t)$, when all other variables in the right-hand side of Equation 5-12 is 0, and (2) an exponentiated linear function of time-invariant variables for article *i* (x_{ik}; e.g., message features) and a lagged time-variant variable $z_i(t-1)$, a binary indicator of whether article *i* is shown in prominent locations on the NYT Health section's main page at time $t-1$ (i.e., an editorial cue to news values). Taking the logarithm of both sides of Equation 5-12, the Cox model can be rearranged as follows:

$$\log r_i(t) = \log r_0(t) + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma z_i(t-1) \qquad \text{(Equation 5-13)}$$

That is, the Cox model specifies that the logged event rate for article *i* at time *t* is a linear function of the time-constant predictors and the lagged time-variant predictor (along with the logged baseline event rate).

The Cox regression model was estimated using the method of partial likelihood (PL; Cox, 1972). The PL estimation method provides consistent estimates of β_k and γ coefficients while allowing for any functional form or shape of baseline event rate; in other words, $r_0(t)$ can be any function of time *t*. More precisely, $r_0(t)$ is eliminated when constructing partial likelihoods for observed events, which makes it possible to estimate the Cox model without having to specify $r_0(t)$'s functional form (for details about the PL estimation method, see Allison, 2014; Cox, 1972; Singer & Willett, 2003). In sum, the Cox model is more robust than parametric event history models (e.g., gamma, lognormal, and Weibull models), in the sense that it yields consistent coefficient estimates for explanatory variables (i.e., β_k and γ), regardless of the $r_0(t)$'s actual

---

where *T* is article *i*'s event time which is a nonnegative continuous random variable. That is, $r_i(t)$ is the probability that article *i*'s event time occurs in the infinitesimally small interval between time *t* and $t + \Delta t$ (i.e., as the interval width, Δt, approaches 0), conditional upon the article having survived to time *t* (i.e., the beginning of the interval), divided by the interval width.

functional form (Efron, 1977). Parametric models, on the contrary, assume a particular functional form[23] for $r_0(t)$ which is potentially inaccurate (Singer & Willett, 2003).

Similarly, a Cox regression model for the early news popularity in terms of making a first-time appearance on the "most-emailed" list can be written as

$$\log r_i(t) = \log r_0(t) + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma_1 z_{i1}(t-1) + \gamma_2 z_{i2}(t) \quad \text{(Equation 5-14)}$$

where $z_{i2}(t)$ denotes whether article *i* appears on the "most-viewed" list at time *t*.

When examining the Cox models with PL estimation, time-varying predictors were dealt with the "episode splitting" method (Allison, 2014), and tied data (i.e., articles having the same event times) were handled using the approximation method proposed by Efron (1977). Standard errors were adjusted for article clusters (Lin & Wei, 1989).

As a model summary statistic, I reported generalized R^2 using the following formula (Allison, 2010; Cox & Snell, 1989; Magee, 1990):

$$Generalized\ R^2 = 1 - \exp\left(\frac{-G^2}{n}\right) \quad \text{(Equation 5-15)}$$

where G^2 is the model likelihood ratio χ^2, and *n* is the number of articles in the model. It should be noted that unlike the usual R^2 for linear regression models, the generalized R^2 does not quantify the faction of variation in the outcome variable explained by model predictors. Instead, it measures the improvement of the full model with the predictors over the baseline model with no predictors (i.e., magnitude of the association between the predictors and the outcome variable, which ranges from 0 to 1). As mentioned above, I estimated robust standard errors that are adjusted for article clusters. This robust variance estimation method, however, employs a log-pseudolikelihood as a maximization criterion,

[23] For example, the Weibull model specifies that $\log r_0(t)$ in Equation 5-13 is a linear function of logged time ($\log t$).

rather than the standard log-likelihood on which the generalized R^2 measure is based. Thus, I used a conventional variance estimation method to calculate the generalized R^2.

In the Results section below, I report unstandardized Cox regression coefficients and provide interpretation in terms of event ratios by exponentiating the coefficients (or relative event rates; Allison, 2014; Singer & Willett, 2003). As with previous analyses, I handled missing data with listwise deletion (Allison, 2002; Enders, 2010). Focal and control message features that were log-transformed are the same as those in Chapter 4.

Results

Table 5-4 presents results from bivariate and multiple Cox regression of the event of news articles' first-time appearance on the "most-viewed" list. Articles with more controversial teasers were more likely to experience the event earlier, unstandardized b = .18, 95% CI [.003, .36]. The exponentiation of the Cox regression coefficient (= exp[b]) yielded an event ratio (relative event rate) of 1.20 with its 95% confidence interval ranges from 1.003 to 1.43, which means that each 1-unit increase in the controversiality score was associated with about 20% increase in the rate of the event of making a first-time appearance on the "most-viewed" list. Effects of other focal message features were not statistically significant, although the directions of their effects were largely consistent with those on the total volume of news attractability (Table 4-1 in Chapter 4).

There was a significant lagged effect of an editorial cue to news values on the event rate of first-time appearance on the "most-viewed" list, b = 1.04, 95% CI [.84, 1.24]. The event rate for articles displayed in prominent locations on the NYT Health section's main page in an earlier time interval was about 2.83 times higher than the rate for those not shown in such places, exp(b) = 2.83, 95% CI [2.32, 3.47].

Table 5-4. Message Effects on the First-Time Appearing on the "Most-Viewed" List

	Bivariate Cox Regression	Multiple Cox Regression
	b (se)	*b (se)*
Efficacy Information Present	.03 (.10)	.06 (.13)
Usefulness	−.05 (.08)	−.03 (.10)
Emotional Positivity (Responses)	−.07 (.11)	−.06 (.15)
Expressed Positivity (Words)	−.002 (.02)	−.04 (.03)
Controversiality	.19** (.07)	.18* (.09)
Emotional Arousal (Responses)	.27 (.17)	.10 (.19)
Expressed Emotionality (Words) [a]	.11 (.07)	.06 (.08)
Novelty	−.18+ (.10)	−.04 (.12)
Diseases / Bad Health Conditions Mentioned	−.06 (.08)	−.22* (.11)
Professional Sources Mentioned	−.18* (.09)	−.18+ (.10)
Death-Related Words Present	.13 (.16)	.12 (.18)
Health Words [a]	.24*** (.07)	.001 (.08)
Social-Processes Words [a]	.18** (.06)	.12+ (.07)
Word Count	.03*** (.01)	.002 (.01)
Writing Complexity (Words > 6 Letters)	.03** (.01)	.01 (.01)
Shown in Prominent Locations [b]	1.11*** (.09)	1.04*** (.10)
Generalized (Cox-Snell) R^2		.32

Note. N = 109,652 for the multiple Cox regression model. Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix G). [a] Log-transformed. [b] Lagged. + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Regarding control variables, articles mentioning diseases or bad health conditions in their teasers were less likely to make the "most-viewed" list earlier, $b = -.22$, 95% CI [−.43, −.01]. The event rate for articles whose teasers included terms related to diseases or bad health conditions was about 80% of that for those with no such terms, exp(*b*) = .80, 95% CI [.65, .99]. The mention of professional sources and the use of social processes-related words were marginally significant predictors of the event rate.[24]

[24] As with the message effects model of the total volume of attractability (Chapter 4), I checked the robustness of the findings by including as additional covariates (1) topical area and (2) the

Table 5-5 presents bivariate and multiple Cox regression results for the event that health news articles make a first-time appearance on the "most-emailed" list. Message features related to information utility facilitated the event occurrence, b = .24, 95% CI [.01, .46] for efficacy information, b = .75, 95% CI [.46, 1.03] for usefulness. Articles containing efficacy information were about 1.27 times more likely to experience the event earlier than those without such information, exp(b) = 1.27, 95% CI [1.01, 1.59]. The event rate for making a first-time appearance on the "most-emailed" list increased by about 111% in response to each 1-unit increase in the usefulness score, exp(b) = 2.11, 95% CI [1.58, 2.81]. Expressed emotionality was also positively associated with the event rate, although the relationship was marginally statistically significant, b = .05, 95% CI [−.01, .11], exp(b) = 1.05, 95% CI [.99, 1.11]. Other focal message features did not have significant effects on the event rate.

As with the event history model of the first-time appearance on the "most-viewed" list, the results revealed a significant lagged effect of an editorial cue to news values, b = .49, 95% CI [.28, .69]. The event of making a first-time appearance on the "most-emailed" list was about 1.62 times more likely to occur to news articles displayed in prominent locations on the main page of the NYT Health section earlier in time than those not featured in such positions, exp(b) = 1.62, 95% CI [1.33, 1.99]. Articles shown on the "most-viewed" list in a given time interval were also more likely to make a first-time appearance on the "most-emailed" list in the same time interval, b = 2.11, 95% CI [1.87, 2.35], exp(b) = 8.23, 95% CI [6.48, 10.45].

---

presence of visual images (in article full texts). Results reported in Table 5-4 remained almost unchanged with this additional control. Effects of the two covariates were not significant.

Table 5-5. Message Effects on the First-Time Appearing on the "Most-Emailed" List

	Bivariate Cox Regression	Multiple Cox Regression
	b (se)	*b (se)*
Efficacy Information Present	.21* (.10)	.24* (.11)
Usefulness	.72*** (.13)	.75*** (.15)
Emotional Positivity (Responses)	.19+ (.10)	.21 (.13)
Expressed Positivity (Words)	.04+ (.02)	−.02 (.03)
Controversiality	.18* (.07)	−.13 (.11)
Emotional Arousal (Responses)	.66*** (.18)	−.17 (.19)
Expressed Emotionality (Words)	.06* (.03)	.05+ (.03)
Novelty	.16 (.11)	.19 (.13)
Exemplification	.38*** (.09)	−.01 (.11)
Credibility Statements		
1	−.55*** (.14)	.03 (.19)
2+ with no opposing statements	.38** (.12)	.08 (.19)
2+ with opposing statements	.42** (.16)	−.02 (.23)
Topic (Reference = Health Policy)		
Disease / Health Conditions	−.02 (.11)	−.03 (.14)
Other	−.27+ (.16)	−.21 (.19)
Writing Style – 1st Person Point of View	.20* (.09)	.21+ (.13)
Death-Related Words Present	.13 (.08)	−.10 (.09)
Health Words [a]	.08 (.09)	−.13 (.11)
Social-Processes Words [a]	.42*** (.10)	.09 (.15)
Word Count × 10^{-2}	.15*** (.01)	.11*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	−.17+ (.09)	.26+ (.15)
(Writing Complexity)^2		−.37+ (.20)
Images Present	.14 (.09)	−.22 (.15)
Number of Hyperlinks [a]	.53*** (.07)	.20** (.07)
Shown in Prominent Locations [b]	1.07*** (.09)	.49*** (.10)
Shown on the "Most-Viewed" List	2.26*** (.10)	2.11*** (.12)
Generalized (Cox-Snell) R^2		.67

Note. N = 144,967 for the multiple Cox regression model. Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Writing Complexity was mean-centered. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix H). [a] Log-transformed. [b] Lagged. + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

With respect to control message features, the event rate of first-time appearance on the "most-emailed" list was higher for (1) longer articles, b = .11, 95% CI [.08, .14], exp(b) = 1.11, 95% CI [1.08, 1.15], and (2) articles containing more hyperlinks, b = .20,

95% CI [.06, .33], exp(*b*) = 1.22, 95% CI [1.06, 1.40]. Writing style (i.e., whether an article was written in a first-person point of view) showed a marginally significant positive relationship with the event rate. Finally, the results revealed a marginally significant negative quadratic effect of writing complexity.

Summary

This chapter examined how public signals about news popularity (i.e., social influence cues) and message features jointly influence temporal dynamics of the volume of news attractability and virality by analyzing pooled time-series cross-sectional data. The results suggest that social influence plays a central role in triggering subsequent news selections and retransmissions.

More importantly, the results provide support for the notion that health news stories containing certain message features making the stories inherently attractable and viral produce stronger social influence-driven cumulative advantage effects. Specifically, the presence of efficacy information amplified the cumulative advantage effects both on news selections and retransmissions. While there was a strong tendency for news articles that stayed longer on the "most-viewed" and "most-emailed" lists to invite more frequent subsequent selections and email-based propagations, respectively, this pattern was even more pronounced for articles containing efficacy information in their teasers (for selections) and full texts (for email-based propagations). Similarly, usefulness and positivity of emotional responses strengthened the cumulative advantage effects of the email-related social influence cue (i.e., the amount of time shown on the "most-emailed" list) on subsequent news sharing via email.

The results also showed that the impact of the email-related social influence indicator went beyond email retransmissions to include carryover effects on news propagations made through social media such as Facebook and Twitter. In addition to the message features that enhanced the social influence-driven cumulative advantage effects on email-based news sharing (i.e., efficacy information, usefulness, and positivity of emotional responses), expressed positivity and exemplification also boosted the social influence effects on news propagations via social media.

Given the finding that the duration of staying on news popularity lists was positively associated with subsequent news selections and propagations (and even more so for articles containing certain message features such as efficacy information), this chapter further investigated what makes health news articles initially popular in terms of making a first-time appearance on the "most-viewed" list and the "most-emailed" list. The results revealed that the controversiality of teasers facilitated articles to make the "most-viewed" list earlier. However, other focal message features were not significantly associated with early news attractability. An editorial cue to news values (i.e., article placement in prominent locations) was a strong predictor of the first-time appearance on the "most-viewed" list. With regard to early news virality, information utility-related message features (i.e., efficacy information and usefulness) promoted the event of first-time appearance on the "most-emailed" list. Expressed emotionality was marginally significantly associated with an increase in the rate of the event.

Taken together, the results reported in this chapter shed light on the interplay of social influence and message features in the temporal processes by which the final news diffusion outcomes (i.e., the total number of selections and that of retransmissions) are

reached. The positive association between the presence of efficacy information and the total volume of news attractability (Chapter 4) can be explained by the finding that articles presenting efficacy information triggered more frequent subsequent selections once they made the "most-viewed" list (social influence cue) than those shown on the list for the same amount of time but providing no efficacy information. Although the presence of efficacy information was not a significant predictor of *early* news attractability, one can speculate that its synergetic interaction with the social influence cue might have resulted in the observed pattern that news articles presenting efficacy information in their teasers were more frequently selected over the full course of online news diffusion. In the case of the controversiality of article teasers, one can conclude from the results of this chapter that articles with more controversial teasers prompted more frequent selections (Chapter 4), because such articles were more likely to make the "most-viewed" list earlier (i.e., became initially popular), which in turn produced social influence-driven cumulative advantage effects (i.e., inviting more frequent subsequent selections; no interaction between social influence and controversiality).

When it comes to news virality, the Chapter 4 results revealed positive associations between the total volume of news retransmissions and three focal message features (efficacy information, usefulness, and positivity of emotional responses). The positive relationships, established over the full news diffusion process, can be accounted for by the current chapter's findings that the three focal message features (1) facilitated the event of making a first-time appearance on the "most-emailed" list and (2) strengthened the cumulative advantage effects of making (and staying on) the list on subsequent news propagations.

CHAPTER 6

PERSISTENCE OF ATTRACTABILITY AND VIRALITY

Overview

So far, this dissertation has examined how the *volume* of news selections and retransmissions is affected by message characteristics of health news articles and social influence. Yet another important dimension of attractability and virality is their *persistence* or *lifespan* (Asur et al., 2011; Berger & Iyengar, 2013, p. 577; Berger & Schwartz, 2011). That is, audience selections and retransmissions of all health news stories grow and fade as time passes, and these communication behaviors stop happening at some time or other.

Then, what makes some health news articles get selected and shared for a longer time period than others? To answer this question, this chapter builds on the same basic framework as used for predicting the temporal dynamics of the volume of news attractability and virality (Chapter 5): the interplay of content characteristics and social influence. Pooled time-series cross-sectional data on audience selections and retransmissions of 760 New York Times (NYT) health news articles are analyzed to examine how message features and social influence jointly shape the persistence of news attractability and virality. With regard to the lifespan of news sharing, as in Chapter 5, the impact of social influence (indicated by the amount of time that articles are shown on the "most-emailed" list) and its interactions with focal message features are examined separately for two types of retransmission channels (i.e., email and social media). By doing so, this chapter explores whether social influence exerts carryover effects on the

persistence of news propagations via social media as it does for their volume. Event history analyses (Allison, 2014; Singer & Willett, 2003) are conducted where an "event" is defined as the "termination" of an article's life in terms of its attractability or virality (i.e., the article no longer being read or shared, respectively).

Research Questions

As discussed in Chapter 2, there is little theoretical or empirical literature in this area. Thus, effects of message features and social influence in driving the persistence of attractability and virality are posed as exploratory research questions.

RQ1: How do (1) message features, (2) social influence, and (3) their interactions affect the persistence of audience news selections?

RQ2: How do (1) message features, (2) social influence, and (3) their interactions affect the persistence of audience news retransmissions via email?

RQ3: How do (1) message features, (2) social influence, and (3) their interactions affect the persistence of audience news retransmissions via social media?

Method

An article teaser is the unit of analysis for the persistence model of news selections, and an article's full text is for that of news retransmissions. The article sample comprises 760 New York Times (NYT) health news articles. Pooled time-series cross-sectional (TSCS) data were used for event history analyses.

In this chapter, an article is defined as having stopped being selected if it is *not* selected (viewed) for *two consecutive days*. Similarly, the termination of retransmission

is operationalized as having occurred when an article is *not* shared for *two consecutive days*. The retransmission termination is measured separately for news propagations via email and social media (Facebook and Twitter).[25] Given that the termination events are defined in terms of two consecutive days of non-selection and non-retransmission, I analyzed daily rather than hourly TSCS data. The daily TSCS data used in this chapter included repeated observations over the course of 30 days on the following variables: (1) binary indicators of whether an event of "termination" occurs to an article with regard to selections and retransmissions, respectively, (2) the number of hours that an article is shown on the "most-viewed" or "most-emailed" list, and (3) the number of hours that an article is displayed in prominent locations on the main page of the NYT's Health section. An article-period dataset was constructed and analyzed with article age (i.e., the number

[25] Unlike the event of making a first-time appearance on popularity lists examined in Chapter 5, the operationalization of the termination of selections and retransmissions is necessarily arbitrary to some extent. It is possible that an article stops getting selected or retransmitted for a certain period of time (e.g., for a day), but resumes being viewed or forwarded later (e.g., on the next day). Moreover, while this dissertation kept track of up to 30 days of selections and propagations, an article can be viewed or shared at any time after 30 days since its online publication. Thus, I defined the event using a cutoff criterion based on empirical as well as conceptual considerations: two consecutive days of non-selection (for attractability) and those of non-retransmission (for virality). Conceptually, I posited that the duration of non-selection and non-sharing should be long enough to treat an article being terminated in getting read or shared. Given that the NYT is a daily newspaper, I considered a one-day a *minimum* length in this regard. From an empirical perspective, however, exploration of the data indicated that using a one-day of non-selection or non-sharing as a cutoff treats a non-negligible number of articles as being terminated which actually were read and shared later in time. Taken together, I opted to use a "two-consecutive-day" of no selection or sharing as a cutoff for defining the event of termination. The average ratio of (1) the cumulative number of news selections up to the event time (i.e., the time point with no selections in the past 24 hours) to (2) the total number of news selections (i.e., 30-day aggregate count) was about .99. Further increases in the duration of non-selection (e.g., three-day or longer) resulted in decreased increments in the ratio. Similar pattern was observed for retransmission data. The corresponding average ratio when using a two-day as a cutoff was .99 for email propagations, and .97 for social media propagations. Analyses with varying cutoff criteria (e.g., one-day and three-day) did not significantly change the results reported in this chapter.

of days since online publication) as a time metric. The article-period dataset included a set of time-invariant variables (e.g., message features) in addition to the time-variant variables listed above. The dataset, as with the event history models in Chapter 5, included each article's time-series data until (1) the article experienced the event (i.e., selection- or retransmission-termination) or (2) the article was right-censored, meaning that the event did not occur to the article by the end of the observation period (Allison, 1984, 2010; Singer & Willett, 2003).

Measures

About 3.3% (n = 25) of the 760 NYT health news articles were right-censored, indicating that these articles had constantly been selected (viewed) over the 30-day period. Of the 760 articles, about 1.2% (n = 9) and 0.9% (n = 7) were right-censored with regard to news propagations via (1) email and (2) social media (Facebook and Twitter), respectively. The strength of the relationship between the persistence and the volume of news attractability was moderate. The Pearson correlation between (1) the time (days) to the termination of article selection (i.e., article's lifespan; log-transformed) and (2) the total number of selections (log-transformed; see Chapter 4) was .54 ($p < .001$). The corresponding relationship for news virality was a little stronger, but it was not so large as to conclude that the two metrics are virtually identical. The Pearson correlation between (1) the number of days to the retransmission termination (log-transformed) and (2) the total retransmission count (log-transformed; see Chapter 4) was .66 ($p < .001$) and .64 ($p < .001$), respectively, for email propagations and social media propagations.[26]

[26] The reported correlation coefficients are based on data where right-censored cases were treated as missing observations. The coefficients were somewhat increased if the right-censored cases

Other time-variant and time-invariant variables analyzed in this chapter were identical to those described in Chapters 3 to 5 (e.g., social influence cues, and message features, etc.).

Analysis

The research question was examined using Cox regression analyses (Allison, 2014; Cox, 1972; Singer & Willett, 2003). Key features of the Cox regression models are detailed in Chapter 5. A Cox regression model for the event of selection termination can be written as

$$\log r_i(t) = \log r_0(t) + \beta_1 x_{i1}(t-1) + \beta_2 x_{i2}(t-1) + \gamma_1 z_{i1} + \cdots + \gamma_k z_{ik} + \delta_1 x_{i2}(t-1) \cdot z_{i1} + \cdots \delta_m x_{i2}(t-1) \cdot z_{im} \quad \text{(Equation 6-1)}$$

where $r_i(t)$ is an instantaneous event (or hazard) rate that selections of article *i* are terminated at time *t* (assuming article *i* has not yet experienced the event earlier), $r_0(t)$ is a baseline hazard rate, $x_{i1}(t-1)$ is the logged number of hours that article *i* is displayed in prominent locations on the main page of the NYT's Health section at time $t-1$, and $x_{i2}(t-1)$ indicates the logged number of hours that article *i* appears on the "most-viewed" list at time $t-1$ (i.e., a social influence cue). Time-invariant variables (e.g., message characteristics) are denoted by z_{ik}. The notation $x_{i2}(t-1) \cdot z_{im}$ indicates a set of interactions between (1) the social influence cue and (2) *m* number of focal message features.

Similarly, a Cox regression model for the termination of retransmissions (i.e., either email- or social media-based news sharing) can be represented by the following equation:

---

were assigned "30" (days) for the lifespan variables: Pearson correlation was .58 for selection, .68 for email retransmission, and .66 for social media retransmission (all *p*-values < .001).

$$\log r_i(t) = \log r_0(t) + \beta_1 x_{i1}(t-1) + \beta_2 x_{i2}(t-1) + \gamma_1 z_{i1} + \cdots + \gamma_k z_{ik} + \delta_1 x_{i2}(t-1) \cdot z_{i1} + \cdots \delta_m x_{i2}(t-1) \cdot z_{im} \quad \text{(Equation 6-2)}$$

where $r_i(t)$ is an instantaneous event (hazard) rate at time *t* for article *i*'s termination in terms of getting retransmitted (via email or social media, respectively), and $x_{i2}(t-1)$ denotes the logged number of hours that article *i* is shown on the "most-emailed" list at time $t-1$.

As with Chapter 5, all Cox models were tested using partial likelihood estimation with robust standard errors (Lin & Wei, 1989). Time-varying explanatory variables and tied data were handled with the "episode splitting" method (Allison, 2014) and the Efron's approximation method (Efron, 1977), respectively. Generalized R^2 was calculated as a model summary statistic (Allison, 2010; Cox & Snell, 1989; Magee, 1990). More details about these statistical decisions are described in Chapter 5.

Unstandardized and exponentiated Cox regression coefficients are reported (Allison, 2014; Singer & Willett, 2003). Regarding the Cox regression results, it should be noted that unlike the event history models in Chapter 5, *negative* unstandardized coefficients (or, equivalently, exponentiated coefficients *less than one*) indicate *positive associations* between predictors and the *persistence* of news attractability and virality because the event rate in the current event history models, $r_i(t)$, is indicative of the hazard of an article's termination in triggering selections and retransmissions. Missing data were handled with listwise deletion (Allison, 2002; Enders, 2010). The same set of focal and control message features as in Chapters 4 and 5 was log-transformed.

Results

Predicting the Persistence of News Attractability

Results from bivariate and multiple Cox regressions models of the persistence of news attractability (RQ1) are presented in Table 6-1. The results revealed a positive association between the positivity of emotional responses toward article teasers and the persistence of news attractability (Model 2 in Table 6-1), unstandardized $b = -.32$, 95% CI [−.58, −.06]. Each 1-unit increase in the positivity of emotional responses was associated with about 27% decrease in the hazard of selection termination, exp(b) = .73, 95% CI [.56, .94]. Expressed emotionality was also positively associated with the persistence of news articles in triggering selections (Model 1 in Table 6-1), $b = -.17$, 95% CI [−.31, −.03]. For each 1-unit increase in the expressed emotionality, the hazard of selection termination went down by about 16%, exp(b) = .84, 95% CI [.73, .97].

The results also identified a significant lagged effect of social influence on the persistence of attractability (Model 1 in Table 6-1), $b = -.26$, 95% CI [−.34, −.19]. Each 1-unit increase in the logged number of hours on the "most-viewed" list in an earlier time interval was associated with about 23% decrease in the hazard of selection termination, exp(b) = .77, 95% CI [.71, .83].

Table 6-1. The Impact of Social Influence and Message Features on the Persistence of News Attractability

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.15 (.10)	.07 (.11)	.07 (.11)
Usefulness	−.02 (.08)	−.001 (.10)	.002 (.10)
Emotional Positivity (Responses)	−.36*** (.10)	−.32* (.13)	−.32* (.13)
Expressed Positivity (Words)	−.02 (.02)	−.002 (.02)	−.002 (.02)
Controversiality	.09 (.06)	−.09 (.08)	−.08 (.08)
Emotional Arousal (Responses)	−.29+ (.15)	−.22 (.17)	−.22 (.17)
Expressed Emotionality (Words) [a]	−.25*** (.06)	−.17* (.07)	−.25** (.09)
Novelty	.18* (.09)	.14 (.11)	.14 (.11)
Diseases / Bad Health Conditions Mentioned	.19** (.07)	−.005 (.10)	−.0004 (.10)
Professional Sources Mentioned	.14+ (.08)	.03 (.08)	.03 (.08)
Death-Related Words Present	.11 (.11)	−.07 (.12)	−.07 (.12)
Health Words [a]	.09 (.06)	.10 (.07)	.10 (.07)
Social-Processes Words [a]	−.11+ (.05)	−.02 (.06)	−.02 (.06)
Word Count	−.01 (.004)	−.0002 (.01)	−.0004 (.01)
Writing Complexity (Words > 6 Letters)	−.01 (.01)	−.01 (.01)	−.01 (.01)
Hours Shown in Prominent Locations [a, b]	−.24** (.09)	−.14 (.09)	−.15 (.09)
Hours Shown on the Most-Viewed (MV) List [a, b]	−.30*** (.04)	−.26*** (.04)	−.27*** (.04)
MV List × Expressed Emotionality (Words) [c]			−.11* (.05)
Generalized (Cox-Snell) R^2		.22	.22

Note. N = 5,998 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix I). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

More important, the results showed a significant and synergetic interaction effect between the social influence cue and the expressed emotionality (Model 2 in Table 6-1), $b = -.11$, 95% CI [−.22, −.01], exp(b) = .89, 95% CI [.80, .99]. As shown in Figure 6-1, while there was a decrease in the hazard of selection termination (i.e., increase in the persistence of selections) for news articles that stayed on the "most-viewed" list for a longer duration, this pattern was more pronounced (i.e., sharper decrease in the hazard)

for articles whose teasers included more emotion words: $b_{social\_influence} = -.20$, 95% CI [−.29, −.11], exp($b_{social\_influence}$) = .82, 95% CI [.75, .89] for article teasers with "low" expressed emotionality (scored at one standard deviation [*SD*] below the mean [*M*]) and $b_{social\_influence} = -.33$, 95% CI [−.44, −.23], exp($b_{social\_influence}$) = .72, 95% CI [.65, .80] for those with "high" expressed emotionality (at $M + 1SD$).[27]

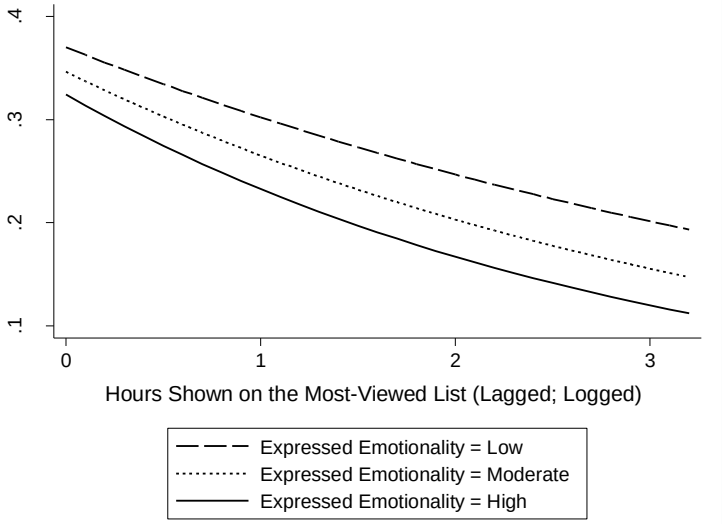

Figure 6-1. The Social Influence × Expressed Emotionality Interaction Effect on the Persistence of News Attractability

Values in Y-axis are predicted event rates (i.e., hazards of selection termination) that are adjusted for explanatory variables in the Cox regression model (Model 2 in Table 6-1). Three values of expressed emotionality: *Low* = $M - 1SD$; *Moderate* = M; *High* = $M + 1SD$ (where *M* and *SD* are, respectively, the mean and the standard deviation of the expressed emotionality score).

Predicting the Persistence of News Virality

Retransmissions via Email

Table 6-2 presents results from bivariate and multiple Cox regression analyses of the persistence of email-based retransmissions (RQ2).

[27] As a robustness check, I tested (1) topical area and (2) the presence of images (in full texts) as additional covariates (see Chapters 4 and 5). Results were almost identical to those in Table 6-1. The two covariates were not significantly associated with the persistence of news attractability.

Table 6-2. The Impact of Social Influence and Message Features on the Persistence of News Virality (Email Retransmissions)

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.36*** (.08)	−.21* (.09)	−.21* (.09)
Usefulness	−.70*** (.11)	−.29* (.13)	−.29* (.13)
Emotional Positivity (Responses)	−.25** (.09)	−.11 (.11)	−.11 (.11)
Expressed Positivity (Words)	−.02 (.02)	.0002 (.02)	.003 (.02)
Controversiality	.01 (.06)	−.10 (.08)	−.10 (.08)
Emotional Arousal (Responses)	−.38* (.15)	−.08 (.18)	−.09 (.18)
Expressed Emotionality (Words)	−.14*** (.03)	−.07** (.02)	−.20*** (.05)
Novelty	−.29** (.09)	−.17 (.11)	−.17 (.11)
Exemplification	−.23** (.08)	.05 (.09)	.06 (.09)
Credibility Statements			
1	−.11 (.12)	−.03 (.16)	−.03 (.16)
2+ with no opposing statements	−.47*** (.11)	−.09 (.16)	−.09 (.16)
2+ with opposing statements	−.17 (.16)	.20 (.20)	.20 (.20)
Topic (Reference = Health Policy)			
Disease / Health Conditions	−.32** (.11)	−.27* (.12)	−.28* (.12)
Other	−.28+ (.15)	−.35* (.15)	−.36* (.15)
Writing Style – 1st Person Point of View	−.13 (.09)	.03 (.11)	.04 (.11)
Death-Related Words Present	.07 (.07)	.16* (.08)	.17* (.08)
Health Words [a]	.02 (.08)	−.003 (.09)	.002 (.09)
Social-Processes Words [a]	−.41*** (.09)	−.04 (.10)	−.03 (.10)
Word Count × 10^{-2}	−.08*** (.01)	−.08*** (.01)	−.08*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.07 (.09)	.05 (.12)	.06 (.12)
Images Present	−.27*** (.08)	−.13 (.13)	−.13 (.12)
Number of Hyperlinks [a]	−.18*** (.05)	−.15** (.06)	−.15** (.06)
Hours Shown in Prominent Locations [a, b]	−.41** (.15)	−.23+ (.14)	−.24+ (.14)
Hours Shown on the Most-Emailed (ME) List [a, b]	−.73*** (.06)	−.60*** (.06)	−.65*** (.07)
ME List × Expressed Emotionality (Words) [c]			−.14** (.05)
Generalized (Cox-Snell) R^2		.39	.40

Note. N = 6,290 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix J). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Persistence of email-related news virality was positively associated with message features related to informational utility, specifically the presence of efficacy information and perceived usefulness, $b = -.21$, 95% CI [−.39, −.03] and $b = -.29$, 95% CI [−.54, −.05], respectively (Model 2 in Table 6-2). The hazard of termination of email-based news retransmissions for articles presenting efficacy information was about 81% of that for those without such information, exp(*b*) = .81, 95% CI [.68, .97]. The hazard of termination decreased by about 25% in response to a 1-unit increase in the usefulness score, exp(*b*) = .75, 95% CI [.58, .96]. There was also a positive relationship between expressed emotionality and the persistence of email retransmissions (Model 1 in Table 6-2), $b = -.07$, 95% CI [−.12, −.02]. For each 1% increase in expressed emotionality (i.e., the percentage of emotion words in article full texts), the article's hazard of termination of email-based news forwarding went down by about 7%, exp(*b*) = .93, 95% CI [.89, .98].

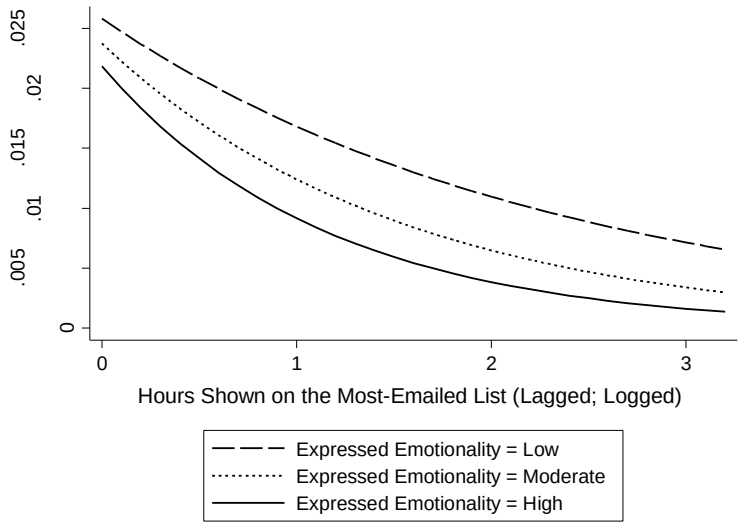

Figure 6-2. The Social Influence × Expressed Emotionality Interaction Effect on the Persistence of News Virality (Email Retransmissions)

Values in Y-axis are predicted event rates (i.e., hazards of termination of email-based news retransmissions) that are adjusted for explanatory variables in the Cox regression model (Model 2 in Table 6-2). Three values of expressed emotionality: *Low* = *M* – 1*SD*; *Moderate* = *M*; *High* = *M* + 1*SD* (where *M* and *SD* are, respectively, the mean and the standard deviation of the expressed emotionality score).

The results further revealed a significant lagged effect of social influence on the persistence of email-related news virality (Model 1 in Table 6-2), $b = -.60$, 95% CI [−.72, −.47]. For each 1-unit increase in the logged number of hours shown on the "most-emailed" list in an earlier time interval, the hazard of termination dropped by about 45%, exp(b) = .55, 95% CI [.49, .62].

As with the case of the persistence of attractability, social influence and expressed emotionality exerted a significant and synergetic interaction effect on the persistence of email retransmission (Model 2 in Table 6-2), $b = -.14$, 95% CI [−.24, −.05], exp(b) = .87, 95% CI [.79, .95]. As Figure 6-2 shows, the social influence effect was strengthened as expressed emotionality increased: $b_{\text{social\_influence}} = -.43$, 95% CI [−.58, −.28], exp($b_{\text{social\_influence}}$) = .65, 95% CI [.56, .76] for articles with "low" expressed emotionality (at $M - 1SD$) and $b_{\text{social\_influence}} = -.87$, 95% CI [−1.12, −.62], exp($b_{\text{social\_influence}}$) = .42, 95% CI [.33, .54] for those with "high" expressed emotionality (at $M + 1SD$).

With respect to control variables, the topical area of news articles had a significant effect on the persistence of email-based news propagations (Model 2 in Table 6-2). Articles about (1) diseases and health conditions and (2) other subjects (e.g., public health and global news) tended to be shared via email for a longer period of time than those related to health policy or health care system, $b = -.28$, 95% CI [−.51, −.05], exp(b) = .76, 95% CI [.60, .95] and $b = -.36$, 95% CI [−.65, −.07], exp(b) = .70, 95% CI [.52, .93], respectively. The results also showed that the presence of death-related words in article full texts facilitated the termination of email retransmissions, $b = .17$, 95% CI [.01, .33], exp(b) = 1.19, 95% CI [1.01, 1.39]. Article length and the logged number of hyperlinks embedded in article full texts were positively associated with the persistence

of news propagations via email, $b = -.08$, 95% CI [−.11, −.06], exp(b) = .92, 95% CI [.90, .95] and $b = -.15$, 95% CI [−.26, −.04], exp(b) = .86, 95% CI [.77, .96], respectively. Article placement in prominent locations on the main page of the NYT's Health section (i.e., an editorial cue to news values) was positively associated with the persistence of email-related news virality, but the relationship was marginally statistically significant.

Retransmissions via Social Media

Bivariate and multiple Cox regression results pertaining to the persistence of news retransmissions via social media (RQ3) are shown in Table 6-3. Content valence-related message features were positively associated with the lifespan of social media-based news propagations (Model 2 in Table 6-3): $b = -.26$, 95% CI [−.48, −.03] for positivity of emotional responses, $b = -.05$, 95% CI [−.10, −.01] for expressed positivity, and $b = -.19$, 95% CI [−.37, −.02] for controversiality. The hazard of termination of social media retransmissions went down by (1) about 23% in response to each 1-unit increase in the emotional positivity rating, exp(b) = .77, 95% CI [.62, .97], and (2) about 5% with each 1% increase in expressed positivity, exp(b) = .95, 95% CI [.90, .99]. An approximate 18% drop in the hazard of termination was associated with each 1-unit increase in the controversiality score, exp(b) = .82, 95% CI [.69, .98]. As with the persistence of selections and email-based retransmissions, news propagations through social media were more likely to persist when articles included more emotion words (see Model 2 in Table 6-3), $b = -.09$, 95% CI [−.14, −.05]. For each 1% increase in expressed emotionality, the hazard of an article no longer inviting social media retransmissions decreased by about 9%, exp(b) = .91, 95% CI [.87, .96].

Table 6-3. The Impact of Social Influence and Message Features on the Persistence of News Virality (Social Media Retransmissions)

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.26** (.08)	−.01 (.09)	−.01 (.09)
Usefulness	−.41*** (.11)	−.14 (.12)	−.14 (.12)
Emotional Positivity (Responses)	−.42*** (.09)	−.25* (.11)	−.26* (.11)
Expressed Positivity (Words)	−.08*** (.02)	−.05* (.02)	−.05* (.02)
Controversiality	.04 (.06)	−.20* (.09)	−.19* (.09)
Emotional Arousal (Responses)	−.58*** (.16)	−.24 (.18)	−.24 (.18)
Expressed Emotionality (Words)	−.13*** (.03)	−.09*** (.02)	−.09*** (.02)
Novelty	−.18+ (.10)	−.04 (.12)	−.03 (.12)
Exemplification	−.23** (.07)	−.02 (.09)	−.32+ (.19)
Credibility Statements			
1	.29* (.13)	.20 (.17)	.20 (.17)
2+ with no opposing statements	−.07 (.12)	.05 (.17)	.04 (.17)
2+ with opposing statements	.15 (.17)	.29 (.21)	.27 (.21)
Topic (Reference = Health Policy)			
Disease / Health Conditions	−.23* (.11)	−.25* (.12)	−.24+ (.12)
Other	−.19 (.14)	−.23 (.15)	−.22 (.15)
Writing Style – 1st Person Point of View	−.14+ (.08)	.08 (.11)	.08 (.11)
Death-Related Words Present	.13+ (.07)	.17* (.08)	.42** (.14)
Health Words [a]	.11 (.08)	−.02 (.09)	−.03 (.09)
Social-Processes Words [a]	−.44*** (.09)	−.05 (.10)	−.06 (.10)
Word Count × 10^{-2}	−.08*** (.01)	−.05*** (.01)	−.05*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.26** (.10)	.32** (.12)	.33** (.12)
Images Present	−.42*** (.08)	−.35** (.12)	−.34** (.12)
Number of Hyperlinks [a]	−.13** (.05)	−.16* (.06)	−.16* (.06)
Hours Shown in Prominent Locations [a, b]	−.74*** (.21)	−.53** (.19)	−.55** (.19)
Hours Shown on the Most-Emailed (ME) List [a, b]	−.64*** (.06)	−.50*** (.06)	−.59*** (.09)
ME List × Exemplification [c]			−.34* (.17)
ME List × Death-Related Words Present [c]			.28* (.11)
Generalized (Cox-Snell) R^2		.37	.38

Note. N = 5,855 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Effects of the following variables are not shown here for brevity: *Article Category*, *Publication Month*, and *Publication Day of the Week* (full results are reported in Appendix K). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Social influence also had a significant lagged impact on the persistence of social media-based propagations (Model 1in Table 6-3), $b = -.50$, 95% CI [−.62, −.39]. Each 1-unit increase in the logged hours shown on the "most-emailed" list in an earlier time interval was associated with about a 40% decrease in the hazard of articles no longer getting shared through social media, exp(b) = .60, 95% CI [.54, .68].

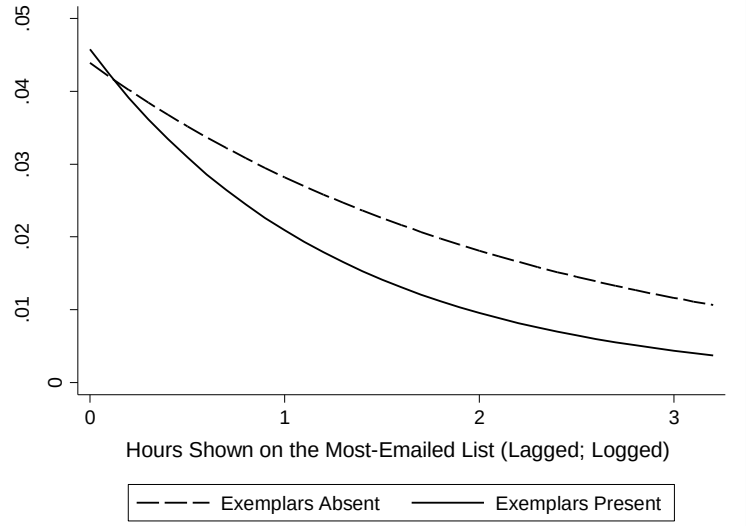

Figure 6-3. The Social Influence × Exemplification Interaction Effect on the Persistence of News Virality (Social Media Retransmissions)

Values in Y-axis are predicted event rates (i.e., hazards of termination of news retransmissions through social media) that are adjusted for explanatory variables in the Cox regression model (Model 2 in Table 6-3).

The social influence effect was further moderated by two message characteristics (Model 2 in Table 6-3): exemplification, $b = -.34$, 95% CI [−.67, −.01], exp(b) = .71, 95% CI [.51, .99], and the presence of death-related words, $b = .28$, 95% CI [.06, .51], exp(b) = 1.33, 95% CI [1.06, 1.66]. Specifically, exemplification strengthened the social influence effect on reducing the hazard of termination of social media retransmissions (i.e., increasing the persistence of social media-related virality). As illustrated in Figure 6-3, the duration of staying on the "most-emailed" list was associated with a sharper decrease in the hazard of termination when news articles presented exemplars in their full

texts, $b = -.93$, 95% CI [−1.27, −.58], exp(b) = .40, 95% CI [.28, .56], compared to when they contained no exemplars, $b = -.59$, 95% CI [−.77, −.41], exp(b) = .56, 95% CI [.47, .67].[28]

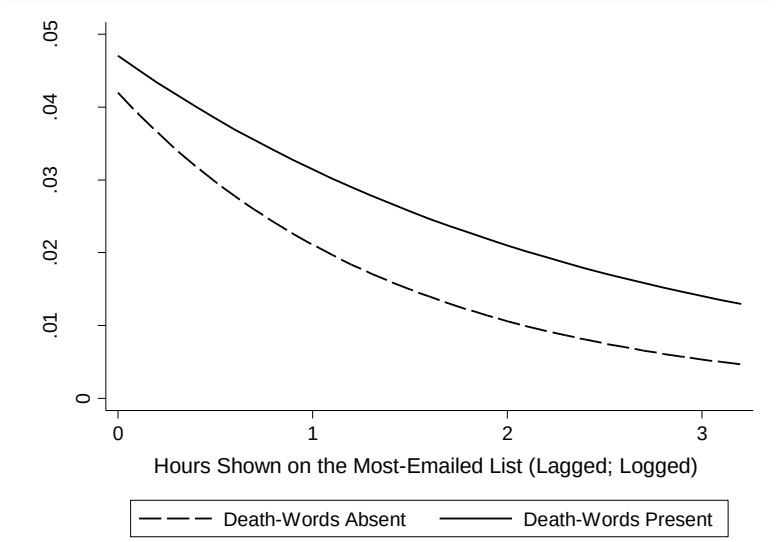

Figure 6-4. The Social Influence × Death-Related Words Interaction Effect on the Persistence of News Virality (Social Media Retransmissions)
Values in Y-axis are predicted event rates (i.e., hazards of termination of news retransmissions through social media) that are adjusted for explanatory variables in the Cox regression model (Model 2 in Table 6-3).

The presence of death-related words played an opposite role to exemplification. As shown in Model 1 in Table 6-3, using death-related words was positively associated with the hazard of termination of news propagations via social media, $b = .17$, 95% CI [.01, .33], exp(b) = 1.18, 95% CI [1.01, 1.39], and further undermined the effect of social influence on reducing the hazard. As depicted in Figure 6-4, the pattern of an association between the social influence cue and the decline in the hazard of termination was less

[28] It should be noted that the coefficient estimates for simple main effects of social influence (i.e., when exemplars are present vs. absent) reported here are those for news articles without death-related words in article full texts, because, as shown in Model 2 in Table 6-3, the social influence factor was also allowed to interact with the presence of death-related words. The difference between the two simple main-effect coefficients, however, is invariant to the choice of the reference category of the death-words variable.

pronounced for articles mentioning death-related words, $b = -.30$, 95% CI [−.45, −.15], exp(b) = .74, 95% CI [.63, .86], than those with no such words, $b = -.59$, 95% CI [−.77, −.41], exp(b) = .56, 95% CI [.47, .67].[29]

With respect to control variables (Model 2 in Table 6-3), social media-based news retransmissions were more likely to persist when articles (1) were longer, $b = -.05$, 95% CI [−.07, −.02], exp(b) = .95, 95% CI [.93, .98], (2) presented images, $b = -.34$, 95% CI [−.58, −.10], exp(b) = .71, 95% CI [.56, .91], (3) included more hyperlinks, $b = -.16$, 95% CI [−.29, −.03], exp(b) = .85, 95% CI [.75, .97], and (4) were displayed in prominent locations on the main page of the NYT Health section for a longer period of time, $b = -.55$, 95% CI [−.93, −.17], exp(b) = .58, 95% CI [.39, .84]. On the other hand, news articles written in a more complex way (i.e., using a greater proportion of complex words) were more likely to facilitate the termination of news propagations through social media, $b = .33$, 95% CI [.09, .56], exp(b) = 1.39, 95% CI [1.10, 1.75].

Summary

In sum, the results of this chapter suggest that social influence and message features jointly shape the persistence of news attractability and virality. The analysis of behavioral measures of audience selections and retransmissions of 760 New York Times (NYT) health news articles revealed that the persistence and volume of these diffusion

[29] Similar to the case of the interaction effect between the social influence cue and exemplification, the coefficient estimates presented here quantify simple main effects of social influence (i.e., when death-related words were mentioned vs. not mentioned) when news articles presented no exemplars. The difference between the two simple main-effect coefficients reported here remains the same with the alternative specification of the reference category of the exemplification (i.e., articles with exemplars).

indicators were not very strongly correlated with each other, and similarly, their significant predictors were also somewhat different. While this chapter's analyses offer new insights into the notion of the persistence or sustainability of news attractability and virality, it is important to note that the analyses were conducted as an exploratory proof-of-concept effort and thus the results should be interpreted accordingly.

Audience selections of news articles were more likely to persist when their teasers (1) evoked positive emotional responses, (2) used more emotion words (i.e., expressed emotionality), and (3) the articles were shown on the "most-viewed" list for a longer period of time (i.e., a social influence cue). It was further found that the positive association between the social influence cue and selection persistence was stronger for article teasers characterized by higher expressed emotionality.

Interestingly, expressed emotionality was also positively associated with the persistence of news propagations via both email and social media (Facebook and Twitter), while its impact on the volume of news virality (either in terms of email or social media forwarding) was not statistically significant (Chapter 4). That is, expressed emotionality was a common message feature that drove news articles to continue to be selected and shared (both via email and social media).

As with the case of the total volume of email-based news sharing (Chapter 4), news articles likely to invite audience retransmissions for a longer period of time were characterized by message features related to informational utility: the presence of efficacy information and perceived article usefulness. The persistence of email-related news virality was also enhanced by social influence and its synergetic interaction with expressed emotionality. Quite consistent with the results on the volume of social media-

related virality (Chapter 4), its persistence was less explained by informational utility-related content features. Rather, it was more significantly and positively associated with message features pertaining to the valence of article content: the positivity of emotional responses and that of emotion words. The results also showed that articles providing more controversial content were more likely to be shared through social media for a longer period of time, while controversiality was not predictive of the total volume of news retransmissions via either social media or email (Chapter 4). Similar to the findings from the temporal dynamics model of the volume of social media-related virality (Chapters 5), the results identified a significant interaction effect between social influence and exemplification, such that the social influence effect on enhancing the persistence of social media-based news sharing was stronger for articles that presented exemplars.

Finally, with regard to control features, it is worth noting that the length of article full texts and the number of hyperlinks in article full texts were positively associated with the persistence of news retransmissions thorough both email and social media. The mention of death-related words in article full texts, on the other hand, was associated with an increase in the hazard that articles no longer triggered news propagations via both types of retransmission channels. Moreover, it further weakened the positive link between social influence and the lifespan of social media-based news retransmissions.

CHAPTER 7

DISCUSSION AND CONCLUSION

What makes media messages more attractable and viral? Why and in what ways do certain messages invite more frequent selections and social propagations than others? Decades of research on selective exposure and social diffusion have identified factors driving people's choices among media messages and their decisions on what to share with their social networks. Yet there are important theoretical and empirical questions that still remain unanswered. Most research has focused either only on content features or only on social influence as drivers of information diffusion, but not on both together. Similarly, little research has examined message selections and retransmissions simultaneously. Little attention has been paid to how digital content-sharing channels such as email and social media affect what people share. Furthermore, little is known about what shapes the persistence (as opposed to the volume) of information diffusion.

By examining how content characteristics and social influence shape the volume and persistence of audience message selections and retransmission, this dissertation fills the gaps in the literature and provides a more comprehensive basis for understanding what makes media messages more attractable and viral. Using a computational social science method (Lazer et al., 2009; Parks, 2014), this dissertation collects aggregate behavioral measures of audience selections and social retransmissions of 760 New York Times (NYT) health news stories in real time and in an automated manner. The aggregate behavioral data are examined in relation to the articles' content and context data, collected by API-based software, web-scraping, content analysis, and a message

evaluation survey. This dissertation's analyses identify message-level ingredients of the volume, as well as the persistence, of news attractability and virality, and further shed light on the interplay of social influence and content characteristics in driving the temporal dynamics of health news diffusion. The analyses also offer insight into how news retransmission channels (email vs. social media) shape what health news goes viral.

Summary and Discussion of Key Findings

Tables 7-1 and 7-2 present central findings from the analyses of the volume and persistence of news selections and retransmissions. Results are summarized and discussed in light of (1) message features predicting attractability and virality, (2) news retransmission channels and virality, (3) social influence and its interaction with message features, (4) the volume versus persistence of attractability and virality, and (5) exogenous and endogenous drivers of attractability and virality.

Message Features Predicting Attractability and Virality

Informational Utility

The results indicate strong support for the notion that informational utility impacts what health news people choose to read and retransmit afterwards, which is consistent with previous research (Berger & Milkman, 2012; Hart et al., 2009; Thorson, 2008).

Table 7-1. Summary of Key Findings I

	Volume of Selections and Retransmissions						
	Total				Temporal Dynamics		
	View	Share			View	Share	
		Total	Email	SM		Email	SM
Message Features							
Efficacy Information	1.41	1.14	1.21		n/a	n/a	n/a
Usefulness		1.40	1.56	1.16	n/a	n/a	n/a
Positive Emotional Responses	1.63[a]	1.18	1.16	1.25	n/a	n/a	n/a
Expressed Positivity (Words)					n/a	n/a	n/a
Controversiality	1.33				n/a	n/a	n/a
Emotional Arousal				1.17	n/a	n/a	n/a
Expressed Emotionality (Words)	1.20				n/a	n/a	n/a
Novelty	0.82		1.14	0.88	n/a	n/a	n/a
Exemplification	n/a			1.12	n/a	n/a	n/a
Social Influence (SI)							
Public Signals about Popularity	n/a	n/a	n/a	n/a	6.94	2.45	1.73
SI × Message Features							
× Efficacy Information	n/a	n/a	n/a	n/a	1.12	1.11	1.03
× Usefulness	n/a	n/a	n/a	n/a		1.19	1.05
× Positive Emotional Responses	n/a	n/a	n/a	n/a		1.08	1.13
× Expressed Positivity (Words)	n/a	n/a	n/a	n/a			1.11
× Expressed Emotionality (Words)	n/a	n/a	n/a	n/a			
× Exemplification	n/a	n/a	n/a	n/a	n/a		1.05
Editorial Cue to News Values							
Article Placement	3.82	1.13	1.14		1.65	1.38	1.54

Note. Cell entries indicate expected increases when predictors change from *low* (= *absence* or *M* − *SD* for dichotomous or continuous variables, respectively) to *high* (= *presence* or *M* + *SD* for dichotomous or continuous variables, respectively). Specifically, each value denotes the *ratio* of (1) the expected number of selections (retransmissions) when a given predictor is *high* to (2) that when it is *low*. Thus, predictors with values greater than one and those less than one represent, respectively, positive and negative associations between the predictors and volume outcomes. Percent changes can be obtained by subtracting one from cell entries (x) and then multiplying 100 (i.e., $100 \times [x - 1]$). All associations reported in this table are statistically significant ($p < .05$). Values for social influence (SI) are SI effects when associated moderators are *low*. For interaction terms, each value indicates an expected increase in the SI effect when a given moderator changes from *low* to *high*. Multiplication of cell entries yields a combined impact of predictors (e.g., effects of four message features on email-based retransmissions is 2.50 [= 1.21 × 1.56 × 1.16 × 1.14]). All predictors are continuous variables, except efficacy information and exemplification (which are dichotomous variables). Temporal dynamics models are based on hourly data. As detailed in Chapter 5, daily data show virtually identical results. [a] conditional effect (when article teasers do not mention diseases or bad health conditions).

Table 7-2. Summary of Key Findings II

	Early Popularity		Hazard of Termination of Selections and Retransmissions		
	View	Email	View	Share	
				Email	SM
Message Features					
Efficacy Information		1.27		0.81	
Usefulness		1.66		0.82	
Positive Emotional Responses			0.79		0.80
Expressed Positivity (Words)					0.83
Controversiality	1.20				0.79
Emotional Arousal					
Expressed Emotionality (Words)			0.76	0.54	0.75
Novelty					
Exemplification					
Social Influence (SI)					
Public Signals about Popularity	n/a	n/a	0.56	0.31	0.19
SI × Message Features					
× Efficacy Information	n/a	n/a			
× Usefulness	n/a	n/a			
× Positive Emotional Responses	n/a	n/a			
× Expressed Positivity (Words)	n/a	n/a			
× Expressed Emotionality (Words)	n/a	n/a	0.69	0.29	
× Exemplification	n/a	n/a	n/a		0.39
Editorial Cue to News Values					
Article Placement	2.83	1.62			0.34

Note. Cell entries indicate expected increases in event rates when predictors change from *low* (= *absence* or $M - SD$ for dichotomous or continuous variables, respectively) to *high* (= *presence* or $M + SD$ for dichotomous or continuous variables, respectively). Specifically, each value denotes the *ratio* of (1) the expected event rate when a given predictor is *high* to (2) that when it is *low*. An "event" refers to (1) an article's first-time appearance on the "most-viewed" ("most-emailed") list for the early popularity model, and (2) an article's termination of selections (retransmissions) for the hazard model. Predictors with values greater than one and those less than one represent, respectively, positive and negative associations between the predictors and early popularity. The opposite is true for the associations between the predictors and *persistence* outcomes because dependent variables here are the *hazards of selection- and retransmission- termination*. Percent changes can be obtained by subtracting one from cell entries (x) and then multiplying 100 (i.e., $100 \times [x - 1]$). All associations reported in this table are statistically significant ($p < .05$). Values for social influence (SI) are SI effects when associated moderators are *low*. For interaction terms, each value indicates an expected increase in the SI effect when a given moderator changes from *low* to *high*. Multiplication of cell entries yields a combined impact of predictors. All predictors are continuous variables, except (1) efficacy information, (2) exemplification, and (3) editorial cues to news values for the early popularity model (which are dichotomous variables).

Health news stories presenting efficacy information were more frequently viewed and shared, and those perceived as more useful were also more likely to go viral. The presence of efficacy information and perceived usefulness also made news articles reach viral status early on in the course of news diffusion, such that they facilitated articles to make the "most-emailed" list earlier. Furthermore, these two message features were positively associated with the persistence of email-based news retransmissions. To my knowledge, this is the first study to demonstrate that efficacy information, which has been shown to be a content feature that enhances the persuasiveness of health message (Witte & Allen, 2000), also makes messages more attractable and viral.

Content Valence

Unlike the prediction of this dissertation that negativity bias operates in news selections and positivity bias drives retransmissions, the results suggest that positivity looms larger in deciding both what to read and what to share. In agreement with previous findings (Alhabash et al., 2013; Berger & Milkman, 2012; Kim et al., 2013), health news stories evoking more positive emotional responses were more viral. Furthermore, positive articles, either in terms of induced or expressed emotions, continued to get shared through social media (i.e., Facebook and Twitter) for a longer period of time than those with negative sentiment.

While controversiality extended the lifespan of news articles in terms of inviting social media-based retransmissions, it was unrelated to the volume of virality, which is inconsistent with this dissertation's prediction. The null effect of controversiality on virality might be explained by a recent study finding that controversial content produces both *interest* and *discomfort* simultaneously (Z. Chen & Berger, 2013), although the

study focuses on conversation likelihood as a final outcome variable. Specifically, Chen and Berger (2013) suggest that controversial messages increase interest but at the same time they also increase discomfort, especially when personal identity is disclosed, as it was in this dissertation's case (i.e., news retransmissions via email and social media reveal personal identity). Thus, it is plausible that the two countervailing psychological states evoked by controversial health news stories led to the observed null impact of controversiality on news propagations. Future research might test psychological factors that mediate or moderate the relationship between controversiality and retransmissions of health news articles.

With regard to attractability, news articles were more frequently selected (viewed) when there was no mention of diseases or bad health conditions in their teasers. Assuming that article teasers including terms related to diseases or unhealthy statuses tend to be perceived as more negative,[30] this finding can be interpreted as showing that positivity bias, rather than negativity bias, operates in news selections. This interpretation is further supported by the significant interaction effect that positive articles were more attractable when their teasers did not mention diseases or bad health conditions. That is, emotional positivity invited more frequent selections for article teasers exhibiting positivity in terms of another dimension of content valence (i.e., no explicit mention of disease-related terms; cf. Heath, 1996). The results further showed that news articles whose teasers induced positive emotional reactions were also more persistent in terms of getting read.

---

[30] The study data supports this speculation. Article teasers mentioning diseases or bad health conditions were rated as significantly less positive than those without such terms, t (757) = 5.93, $p < .001$.

As mentioned earlier, the observed positivity bias in audience news selections is at odds with previous research findings that content negativity drives selective exposure. One reason for this inconsistency might be the difference in topical domains chosen for theory testing. Many message stimuli employed in the past studies were about politics (Donsbach, 1991; Meffert et al., 2006) and crimes or accidents (Knobloch, Hastall, et al., 2003; Zillmann et al., 2004), while this dissertation focused exclusively on health news. Compared to news about politics, crimes, and accidents, health news might be more self-focused and more directly linked to individual well-being. In fact, this appeared to be the case especially for the health news articles examined in this dissertation. As reported in Chapter 4, about 68.8% of the articles were about individuals' diseases and health conditions. Therefore, it may be that people avoid negative or bad news stories if they cover such self-oriented health topics. This line of reasoning is also consistent with the finding of a recent study that examines message-level predictors of selective exposure to health information (Kim et al., 2013). In their study, Kim and colleagues found that smokers are more likely to choose tobacco control messages (introduced as brief summaries of health videos) evoking positive – rather than negative – feelings (Kim et al., 2013).

Controversiality was the only negativity-related content feature that boosted attractability both in terms of its volume and persistence. In line with previous studies suggesting that controversy- or conflict-oriented news frames draw more audience attention (e.g., Zillmann et al., 2004; see also Cappella, 2002; Cappella & Jamieson, 1997), health news stories with more controversial teasers received more frequent selections. Moreover, the controversiality effect manifested itself from an early stage of

news diffusion, such that articles with more controversial teasers made the "most-viewed" list earlier. Taken together, the results suggest that it is controversiality (a specific component of negativity) – rather than overall content negativity – that enhances news attractability.

Emotional Evocativeness

Health news articles using more emotion words (i.e., high expressed emotionality) in their teasers triggered more frequent selections, and emotionally arousing articles were more frequently retransmitted via social media, which is mostly consistent with previous findings that emotional evocativeness boosts both attractability (Zillmann et al., 2004) and virality (Berger & Milkman, 2012). Expressed emotionality was also positively associated with the lifespan of both news attractability and virality. Health news articles continued to get selected and shared, either via email or social media, for a longer time period when they used more emotion words in their teasers and full texts, respectively.

The observed association between emotional evocativeness and virality, together with the positivity-virality link, can be further discussed in relation to the role of discrete emotions in driving news propagations. Recent empirical studies (Berger, 2011; Berger & Milkman, 2012) reveal that while positively valenced messages are overall more viral, discrete emotions with varying levels of *physiological* arousal impact virality differently. They found that independent of the valence effect, emotions characterized by high physiological arousal (e.g., amusement and anger) increase virality, whereas those of low physiological arousal (e.g., contentment and sadness) decrease it. Given the recent findings, I conducted bivariate correlation analyses to investigate how the logged total number of news retransmissions (see Chapter 4 for details) relates to each specific

emotion examined in the previous studies.[31] The analyses revealed no evidence that high- versus low- physiological-arousal emotions have opposite relationships with virality. Instead, discrete emotions were associated with virality in ways that are consistent with the Chapter 4 results based on a single scale of emotional valence (positivity). The logged total number of news propagations was significantly positively associated with positive emotions ($r = .16$, $p < .001$ for amusement; $r = .19$, $p < .001$ for contentment), whereas its relationships with negative emotions were significantly negative ($r = -.07$, $p < .05$ for anger; $-.13$, $p < .05$ for sadness). In sum, the results suggest that emotional positivity – but not physiological arousal alongside – boosts content virality (and emotional arousal for social media-specific virality), in so far as the content is health news.

Novelty and Exemplification

The results revealed a negative relationship between novelty and attractability, which runs counter to this dissertation's prediction and previous literature (J. H. Lee, 2008; Shoemaker et al., 1987; Shoemaker & Cohen, 2006). As with the case of the valence-attractability link, topical difference and associated psychological factors might explain the discrepancy between the present and past findings. Little research has used health messages to examine how novelty affects attractability. Instead, for example, an experimental study identifying a causal path from novelty to audience news selection employed *crime* news stories as stimuli (J. H. Lee, 2008). On the other hand, recall that

[31] Bivariate analyses were conducted here because the inclusion of discrete emotions as predictors in the message effects model of the total volume of virality (Chapter 4) produced a near extreme multicollinearity issue (recall that I created a single scale of emotional positivity because of a high internal consistency among discrete emotion items).

more than two thirds of health news articles examined in this dissertation are about diseases and health conditions, which are presumably more self-oriented topics. Thus, because the information addresses self-oriented issues such as diseases and health conditions, it may be that individuals choose familiar health information in defense of certainty, rather than new, unusual, deviant, or surprising information that is potentially threatening. On the contrary, individuals may still seek out unusual or surprising messages because such messages are appraised as more interesting (Silvia, 2005, 2008), but only if the messages are about relatively other-focused topics such as urban legends and crimes, rather than self-focused topics. Another possibility is that news articles with novel teasers invite less frequent selections because novelty in this context undermines persuasiveness (which is positively associated with attractability; Kim et al., 2013). Individuals may consider novel health information as unpersuasive when the information is embedded in short texts like teasers because there is little room to convey supporting reasons or evidence in such brief texts.[32] Future work might examine psychological mechanisms that underlie the negative association between the novelty and attractability of health news, and how they operate differentially across topical domains.

With regard to virality, novelty was unrelated to the total number of news retransmissions, which is inconsistent with previous studies (Berger, 2013; Kim et al., 2013; Loewenstein & Heath, 2009). Further analyses showed that the non-significant relationship emerged because novelty was positively associated with news propagations via email, whereas it was negatively related to those through social media. The results also showed that health news articles presenting exemplars were more frequently shared

[32] As reported in Chapter 4, the average word-count of article teasers was 33.26 (*SD* = 7.42).

via social media. Detailed discussions about the effects of novelty and exemplification on news virality are provided in relation to the role of retransmission channels (email vs. social media) in the section below.

News Retransmission Channels and Virality

The results indicate that online news retransmission channels such as email (i.e., narrowcasting) and social media (i.e., broadcasting) significantly affect what news people share with their social networks. This is consistent with recent theoretical and empirical works demonstrating that a news propagator's consideration of target audience plays a significant role in deciding what to share (Barasch & Berger, 2014; Falk et al., 2013; Falk et al., 2012).

Specifically, message features related to informational utility (i.e., efficacy information and usefulness) were more closely tied to news retransmissions via email than those via social media, both in terms of their volume and persistence. On the other hand, emotion-related content characteristics played a larger role in boosting social media-specific virality. Emotional arousal invited more frequent social media-based news propagations, while it was unrelated to email-based retransmissions. Positive news articles (either in terms of emotional responses or expressed positivity) lasted longer in terms of getting shared through social media, but not in terms of email-based news propagations. These findings are overall consistent with recent theorizing and empirical evidence (Barasch & Berger, 2014) that narrowcasting triggers social sharing of useful content by activating other-focus (i.e., message recipients), whereas broadcasting ignites social propagation of self-enhancing or self-presentational content (e.g., emotionally positive and arousing content) by boosting self-focus (i.e., messenger or sharer).

Novelty played an opposite role in news retransmissions via email and in those via social media. Novel health news stories were more frequently forwarded via email, which is in agreement with previous research (Berger, 2013; Kim et al., 2013). However, novelty was negatively associated with the total volume of social media-based propagations. The opposite role of novelty in the context of the two retransmission channels might also be due to the differences in how people perceive their target audience when deciding whether to pass along health news stories. Compared to email-forwarding, message recipients of social media-based retransmissions (i.e., Twitter "followers" or Facebook "friends") tend to be larger in size and more diverse in terms of demographics, preferences, and relationship strengths (Barasch & Berger, 2014; Berger & Milkman, 2012). Thus, sharing health news that is (1) unusual or surprising and (2) closely tied to individual well-being with large and heterogeneous audience members might be considered detrimental to enhancing a positive self-view (or at least unclear as to whether it would be helpful to self-enhancement) because doing so could annoy or offend someone in the sharer's social networks (Barasch & Berger, 2014; De Angelis et al., 2012; Hennig-Thurau et al., 2004; Sundaram et al., 1998). This psychological consideration might have produced the observed pattern that relatively familiar health information was more frequently shared via social media. On the other hand, compared to news retransmissions through social media, email-based sharing tends to involve a smaller and narrower audience. Perhaps more importantly, sharers usually specify particular recipients when they use email to forward news stories, while it is much less common to do so on social media (albeit possible). That is, people might feel "safer" to share unusual, new, and surprising health news articles (which tend to be interesting and

remarkable in general; Berger, 2013; Silvia, 2005, 2008) via email because they can narrowcast to particular audience members who they think would like the articles. In sum, when it comes to health news, it appears to be email, rather than social media, that ensures high social currency of novel content (Berger, 2013) because sharers have more control of targeting specific audience and thus have more information about their audience members (e.g., backgrounds and preferences). This psychological mechanism might have underlain the observed positive association between novelty and email-based retransmissions of health news stories.

Health news articles presenting exemplars – delivery vehicles of health messages (Cappella, 2006; Kim et al., 2012) – were more frequently shared through social media, while exemplification was unrelated to the volume of email-based propagations. This retransmission-channel difference might be due to the aforementioned psychological tendency that assuming a larger audience (broadcasting) leads news propagators to focus more on themselves than recipients (i.e., self-enhancement motivation), compared to when deciding what to share through email (narrowcasting). That is, exemplification might boost social media-based retransmissions because story-like messages have high social currency when people communicate with a large audience (i.e., self-enhancing content; Berger, 2013), but not necessarily so when assuming a smaller and narrower audience.

In sum, the results underscore the significant role of retransmission channels in shaping the relationship between content characteristics and virality. While this dissertation provided some explanations as to why email- and social media-based news retransmissions make a difference in what goes viral, they are speculative rather than

empirically grounded, given the lack of data concerning the social psychology of such effects. Therefore, more research is warranted to examine psychological mechanisms that underlie and determine the impact of news retransmission channels in health contexts, including the role of narrowcasting- and broadcasting-related news sharing motivations (Barasch & Berger, 2014).

Social Influence and Its Interactions with Message Features

Analyses of temporal dynamics of health news diffusion highlight a crucial role of social influence in boosting the volume and persistence of attractability and virality, which is consistent with prior research (Messing & Westwood, 2012; Muchnik et al., 2013; Salganik et al., 2006; Salganik & Watts, 2008, 2009a). Public signals about news popularity (i.e. social influence cues) produced cumulative advantage effects (DiPrete & Eirich, 2006; Salganik & Watts, 2009a), such that news articles shown for a longer time on the "most-viewed" ("most-emailed") list on the main page of the NYT's Health section (1) triggered more frequent subsequent selections (retransmissions) and (2) were more persistent in terms of getting read (shared). These findings suggest that news consumption and propagation are essentially "social" communication behaviors in the emerging media landscape (Napoli, 2011; Rainie & Wellman, 2012; Williams & Delli Carpini, 2011). The results also indicate strong support for the notion that the source of social influence extends beyond one's real-world relationships (Katz & Lazarsfeld, 2006) to include anonymous or impersonal others whose aggregate behaviors are represented in the form of sheer numbers (Cialdini, 2003; Mutz, 1998; Salganik et al., 2006) in the context of health news exposure and sharing.

The results further demonstrated that social influence interacted with certain message characteristics in a synergistic manner to increase news attractability and virality. While public signals about popularity produced cumulative advantage effects, health news articles with certain message features generated even stronger social influence effects than those staying on the news popularity list (i.e., "most-viewed" or most-emailed" list) for the same amount of time but lacking (or having a lower level of) such features. Specifically, the presence of efficacy information generated stronger social influence-driven cumulative advantage effects on subsequent news selections and retransmissions (both email- and social media-based). News articles (1) perceived as more useful and (2) evoking more positive emotional responses benefited more from the social influence effects on subsequent news propagations through email and social media. For the temporal dynamics of the volume of social media-specific virality, the magnitude of social influence effects was also enlarged by expressed positivity and exemplification. The results further indicated that expressed emotionality strengthened social influence effects on the persistence of news attractability and email-related virality, and that exemplification enhanced the role of social influence in extending the lifespan of news retransmissions through social media.

In sum, the analyses of this dissertation shed light on the interplay of focal message features and social influence over the course of health news diffusion, which underlies the overall effects of central content characteristics on the total volume of news attractability and virality. The results suggest that while audience news selection and retransmission behaviors are strongly influenced by popularity information (i.e., indicators of what others read and share), those communication behaviors are not simply

imitative but instead are also based on another important consideration: *message features*. Health news consumers were more likely to select or retransmit articles (1) when many others had viewed or shared the articles earlier *and* (2) when the articles had certain message features, rather than merely depending on and imitating others' behaviors. Specifically, it should be noted that social influence produces significantly stronger effects for news articles with message characteristics, most of which are significantly associated with the total volume of news attractability and virality. The close correspondence between (1) message attributes reinforcing social influence effects and (2) those predictive of the total frequency of news selections and retransmissions supports the notion that there are certain features of messages – rooted in biological and/or sociocultural factors – that inherently boost the messages' attractability and virality (i.e., an epidemiological approach to message effects; Berger, 2013; Cappella, 2002; Heath & Heath, 2007; Sperber, 1996; see also Katz, 1976, 1999; Rogers, 2003; Tarde, 1903).

Volume versus Persistence of Attractability and Virality

Exploratory analyses of this dissertation suggest that (1) the volume of news selections and retransmissions and (2) their persistence tap into related but different dimensions of health news diffusion (Asur et al., 2011; Berger & Schwartz, 2011). The volume and persistence measures were correlated moderately (attractability) or somewhat strongly (virality), and their predictors were somewhat dissimilar.

Social influence and some message features (e.g., efficacy information, emotional valence, and exemplification) shaped the persistence of news selections and propagations in a similar way to their effects on the volume of those communication behaviors. At the same time, however, the results also identified message-level predictors that are relatively

unique to the lifespan of news attractability and virality: expressed emotionality and controversiality. Expressed emotionality was positively associated with the persistence of news selections and propagations (both via email and social media) but unrelated to their volume (except attractability). It further interacted synergistically with social influence cues to extend the lifespan of attractability and email-specific virality, but the interaction effect was not significant on their volume. Similarly, controversial health news articles were more long-lived in terms of inviting social media-based retransmissions, while controversiality was unrelated to the volume of social media-related virality. Taken together, while expressed emotionality and controversiality do not necessarily make health news articles achieve enormous popularity, they seem to boost the articles' staying power to continue to be read and propagated, and essentially survive longer (Cappella, 2002). It is also worth noting that emotion-laden and controversial content has been considered to have high news value (Harcup & O'Neill, 2001; Shoemaker & Cohen, 2006; Stephens, 2007; see also Cappella & Jamieson, 1997).

As discussed earlier in this dissertation, there is sparse theoretical and empirical research regarding the persistence dimension and its predictors. Hence, the current analyses are exploratory in nature, and the obtained results should be interpreted accordingly. Future research will need to theorize further and examine factors shaping the lifespan of news selections and retransmissions, especially in comparison to those for the volume of the communication behaviors.

Exogenous and Endogenous Drivers of Attractability and Virality

The results of this dissertation can also be discussed in light of exogenous and endogenous drivers of news selections and retransmissions over the course of health news

diffusion. Now that public signals about news popularity (i.e., articles shown on the "most-viewed" or "most-emailed" list) are based on automated aggregations of audience behaviors (i.e., viewing and sharing), the way they impact subsequent news selections and propagations involves bottom-up and generative processes. That is, social influence is an endogenous driver that is uncontrollable (or unmanipulable) in a natural and real-world context. On the other hand, message features and editorial decisions about article positioning (i.e., editorial cues to news values) represent exogenous drivers that are controllable and unaffected by audience selection and retransmission behaviors during the news diffusion process.[33] Of course, the intuitions and experience of editors in article placement may be implicitly tracking the dimensions of news that are studied here, empirically and explicitly leading to placements that reflect exogenous factors.

The exogenous-endogenous driver distinction has important implications for message design for web-based public health communication campaigns (e.g., email health newsletters or framing of health press releases) where messages and their positions on a webpage (or email newsletter) are determined a priori. Specifically, it would be useful to quantify what consequences in message selections and propagations would follow from manipulating the exogenous and controllable factors. Using message effects models in Chapter 4, I estimated predicted increases in the number of news selections (retransmissions) in response to changes in (1) focal message features and (2) editorial

[33] It is possible that articles' positions on the main page of the NYT Health section are affected by their popularity, such that editorial decisions are made to place popular articles in an earlier time interval on prominent locations in a later time interval. I conducted an ancillary analysis to check this possibility using pooled time-series cross-sectional data. Results revealed that whether articles were shown in prominent locations in a given time interval were unaffected by their popularity in an earlier time interval (either in terms of selections or retransmissions).

cues to news values (i.e., editorial decisions about article positioning) that were found to be significantly associated with the total volume of attractability (virality).

Specifications of the predictive analysis are as follows. First, I estimated the predicted total selections and retransmissions when both focal message features and editorial cues to news values are *low*, while other variables in the message effects models (Chapter 4) are held constant ("Baseline"). Second, while everything else remains the same as for the "Baseline" specifications, I obtained corresponding predicted values when focal message features are *high* ("Message Features"). Third, other features being identical to those of "Message Features," I estimated predicted scores when editorial cues to news values are *high* ("Message Features & Editorial Cues"). When estimating predicted total retransmissions, I also included indirect effects of message features and editorial cues that are mediated through the total volume of selections, in addition to their direct effects. Details about variations (i.e., low vs. high) in focal message features and editorial cues are summarized in Table 7-3.

Figure 7-1 presents results from this ancillary analysis. Everything else being equal, health news articles with *high* message features are predicted to invite about 4.25 times more frequent selections than those with *low* message features (55,866 vs. 13,137).[34] Similarly, articles equipped with *high* message characteristics are expected to

[34] Recall that I took the logarithms of the total number of selections (and that of retransmissions) and used them as dependent variables for the message effects models in Chapter 4. Thus, I obtained predicted values for the original scales (i.e., "numbers") by back-transforming the model-based predicted values (i.e., "logged numbers") using the following formula (Wooldridge, 2009):

$$\hat{y} = exp\left(\frac{\hat{\sigma}^2}{2}\right) \cdot exp\left(\widehat{\log y}\right)$$

trigger about 6.34 times more frequent propagations (either via email or social media) than those with *low* message characteristics (1,906 vs. 300).

Table 7-3. Low and High Exogenous Factors for Predictive Analysis

	Low	High
Selections		
Message Features (*Teasers*)	Efficacy Information *Absent* Emotional Positivity = *M* – 1*SD* Diseases / BHC *Mentioned* Controversiality = *M* – 1*SD* Expressed Emotionality* = *M* – 1*SD* Novelty = *M* + 1*SD*	Efficacy Information *Present* Emotional Positivity = *M* + 1*SD* Diseases / BHC *Not Mentioned* Controversiality = *M* + 1*SD* Expressed Emotionality* = *M* + 1*SD* Novelty = *M* – 1*SD*
Editorial Cues to News values	Total Hours Shown in Prominent Locations* = *M* – 1*SD*	Total Hours Shown in Prominent Locations* = *M* + 1*SD*
Retransmissions		
Message Features (*Full Texts*)	*Direct* Effects • Efficacy Information *Absent* • Perceived Usefulness = *M* – 1*SD* • Emotional Positivity = *M* – 1*SD* *Indirect* Effects • Selections* (mediator) = *M*	*Direct* Effects • Efficacy Information *Present* • Perceived Usefulness = *M* + 1*SD* • Emotional Positivity = *M* + 1*SD* *Indirect* Effects • Selections* (mediator) = *M* + 1.45[a]
Editorial Cues to News values	*Direct* Effects • Total Hours Shown in Prominent Locations* = *M* – 1*SD* *Indirect* Effects • Selections* (mediator) = *M*	*Direct* Effects • Total Hours Shown in Prominent Locations* = *M* + 1*SD* *Indirect* Effects • Selections* (mediator) = *M* + 1.34[b]

Note. * Log-transformed. BHC = Bad Health Condition. Means (*M*) and standard deviations (*SD*) are as follows: emotional positivity (responses toward teasers: *M* = 2.78, *SD* = .38; responses toward full texts: *M* = 2.80, *SD* = .43); controversiality (teasers: *M* = 2.93, *SD* = .57); expressed emotionality (teasers: *M* = .88, *SD* = .56); novelty (teasers: *M* = 2.77, *SD* = .42); perceived usefulness (full texts: *M* = 3.84, *SD* = .34); total hours shown in prominent locations (*M* = 2.05, *SD* = 1.52); selections (*M* = 9.95, *SD* = 1.44). [a] Predicted difference in Selections between *low* and *high* message features of article teasers. [b] Predicted difference in Selections between *low* and *high* editorial cues to news values (thus, the value of Selections used in the "Message Features & Editorial Cues" analysis for the predicted total retransmission count = *M* + 1.45 + 1.34).

---

where $\hat{y}$ is a predicted value for the total number of selections (or retransmissions), $\hat{\sigma}^2$ indicates an estimated mean squared error (MSE) of an OLS regression model, and $\widehat{\log y}$ denotes a predicted value for the logged total number of selections (or retransmissions).

The results also indicate that one can expect even further increases in the volume of selections and retransmissions by displaying health messages in prominent locations on a website for a longer time in addition to designing inherently attractable and viral messages. Other things being equal, increasing the amount of time that health news articles stay in prominent locations from *low* to *high* is predicted to invite about 3.82 times more frequent selections (213,170 vs. 55,866) and about 3.46 times more frequent retransmissions (6,599 vs. 1,906), when compared to the "Message Features" results. Taken together, the combination of (1) crafting health news stories with strong message features and (2) showing the stories in prominent locations for a longer time is expected to prompt about 16.23 times more frequent selections and about 21.97 times more frequent propagations, compared to when no such efforts are made.

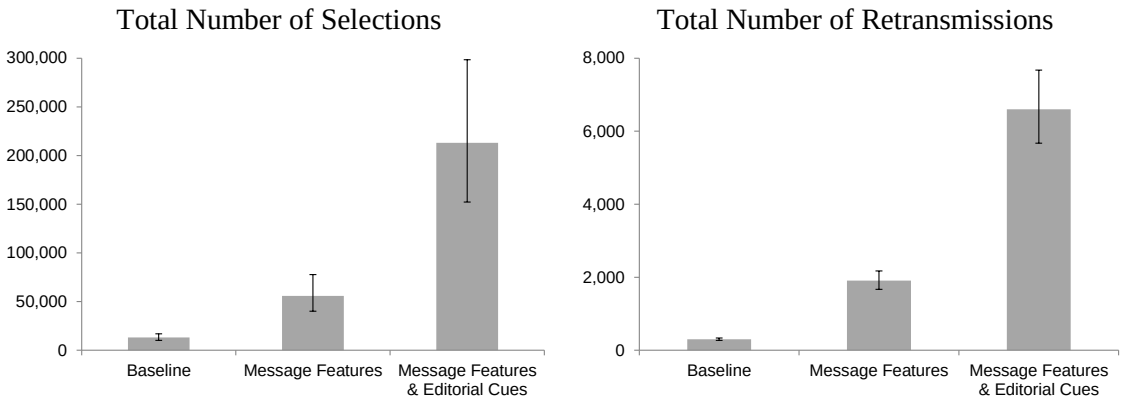

Figure 7-1. Combined Effects of Message Features and Editorial Decisions
Values in bar graphs represent predicted total number of news selections (Left) and that of news retransmissions (Right) along with their 95% confidence intervals. The predicted values are derived from message effects models in Chapter 4.

A related point worthy of further discussion is that editorial cues to news values have strong effects on news attractability and virality (see Figure 7-1 and Tables 7-1 and 7-2). While this finding might be unsurprising given the literature documenting a "position bias" in drawing user attention and igniting click-through behaviors in online

contexts (Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Joachims et al., 2007; Pan et al., 2007), it has important implications for our understanding of news diffusion in the changing media landscape characterized by the increasing use of digital and social media for news consumption (Pew Research Center, 2013a). Consistent with recent research (Cha et al., 2012; Goel et al., 2012; S. Wu, Hofman, Mason, & Watts, 2011), the results suggest that traditional news outlets (i.e., mass media) and their journalistic judgments about news values still play a central role in the social flow of news in the emerging public communication environment by impacting what people read and share with their social networks (Katz, 1961; Katz & Lazarsfeld, 2006).

Limitations and Future Directions

While this dissertation sheds light on how message characteristics and social influence jointly drive news attractability and virality, much more remains to be done to advance this line of research by addressing limitations of the current work. In addition to those already discussed, this dissertation has several other limitations.

This dissertation analyzed NYT health news stories as a study sample. Thus, results reported in this dissertation may not generalize to attractability and virality of health news articles of other news outlets. While this dissertation focused on NYT data primarily because of their measurement quality (e.g., selection and retransmission *count*; see Chapter 2 for details), future research might test the generalizability of the current findings using health news data from other outlets.

This dissertation did not manipulate key independent variables (i.e., message features and social influence) with random assignment but measured them instead. Thus,

despite the efforts to measure and control for potential confounders, this dissertation cannot conclusively rule out the possibility that a causal inference from the observed effects is spurious (Shadish, Cook, & Campbell, 2002). For example, unmeasured message characteristics such as open-ended information presentation (Southwell, 2013) and interest (Berger & Milkman, 2012; Berger & Schwartz, 2011) might explain the observed relationships between (1) message features and (2) attractability and virality. Thus, future research will need to conduct a large-scale web-based experiment (Salganik et al., 2006; Salganik & Watts, 2008, 2009b; Watts, 2011) that manipulates both message features and social influence cues to test their causal impact on news selections and retransmissions in a clearer way.

It should also be noted that this dissertation's message evaluation survey was conducted with an online convenience sample recruited from Amazon's Mechanical Turk. Thus, survey respondents' aggregate assessments of article teasers and full texts are not necessarily representative of those from NYTimes.com readers. In addition, as detailed in Appendix A, there was a low level of agreement among respondents' ratings on the measure of emotional arousal evoked by article texts. In sum, the results pertaining to perceived message features should be interpreted in light of the limitations of the message evaluation survey.

This dissertation examined separately (1) the impact of editorial cues to news values (i.e., articles displayed in prominent positions on the main page of the NYT website's Health section) and (2) that of social influence cues (i.e., articles shown on the "most-viewed" and "most-emailed" lists). While the distinction between editorial and social influence factors is conceptually clear, this dissertation cannot conclusively rule

out the possibility that the observed effects of the two factors are empirically inseparable from those of article accessibility on the NYT website (Berger, 2013; Berger & Heath, 2005; Berger & Schwartz, 2011). That is, one cannot say for certain from the current data whether health news articles shown in prominent locations and popular-news lists are more frequently read and retransmitted because of editorial cues to news values and social influence, or due to the fact that such articles are more accessible on the website. Future work should address this issue by employing an experimental design that makes the editorial, social influence, and accessibility factors independent or orthogonal from each other (e.g., Salganik et al., 2006).

In this dissertation, audience news selections and retransmission behaviors are observed at the aggregate level. There thus remains an important empirical question worth investigating more thoroughly as to whether the observed results are replicated at the individual level (Axelrod, 1997; Epstein, 2006; Gilbert, 2007; Macy & Willer, 2002; Miller & Page, 2007; Schelling, 2006; Vicsek, 2002; Watts, 2007). In addition to the replication tests of the current findings at the individual level, the following two person-level factors warrant further investigation. First, now that there is evidence that sharers' consideration of target audience shapes what they share (Barasch & Berger, 2014; Falk et al., 2013; Falk et al., 2012), it would be important to examine how sharers' presumptions about the target audience's evaluations of content characteristics (i.e., their perception of what recipients would think about the forwarded news) affects their news retransmission decision, and compare its impact with that of their own evaluations. Second, more research is also warranted to test how the relationships between message features and diffusion outcomes (i.e., attractability and virality) are moderated by opinion leadership

which taps into individual differences in motivation and ability to spread messages (Boster, Carpenter, Andrews, & Mongeau, 2012; Boster, Kotowski, Andrews, & Serota, 2011; Katz & Lazarsfeld, 2006; Rogers, 2003; Weimann, 1994).

This dissertation focused on news propagations that take place online. However, people also share news with their social networks face-to-face (Katz, 2006; Katz & Lazarsfeld, 2006). A recent report shows that the most common communication channel through which people receive news from their friends and family is face-to-face word of mouth (about 72%; Pew Research Center, 2013b).[35] Similarly, about 76% of word of mouth about brands (and associated products and services) takes place in face-to-face communication contexts (Keller & Fay, 2012). Future studies can thus shed further light on drivers of news virality by examining face-to-face news sharing as an outcome variable and by testing how communication modalities affect what news people spread (e.g., oral vs. written communication; Berger & Iyengar, 2013).

It is also important to note that while intrinsic message features (e.g., the presence of efficacy information) can also indirectly impact news attractability and virality by shaping perceived or effect-based message features (e.g., perceived usefulness; O'Keefe, 2003), no such indirect effects were tested in this dissertation. Instead, all prediction models examined in this dissertation treated intrinsic and perceived message properties as parallel predictors. The parallel-predictors approach was preferred because this

[35] This might have resulted in the pattern that the frequency of news retransmissions is much smaller relative to that of selections in this dissertation. For a given news article, on average, the total number of news sharing composed about 4% of the total number of selections (i.e., the likelihood of sharing given selection) when using the total retransmission scale of (1) NYT API's email count and (2) social media APIs' Facebook and Twitter count (recall that NYT API's Facebook and Twitter sharing count is a lower bound of the actual one. See Chapters 3 and 4 for details). The average percentage was about 2% when using the total retransmission measure solely based on data from NYT API (i.e., NYT API's email, Facebook, and Twitter count).

dissertation analyzed aggregate-level data where mediating paths are conceptually less clear than individual-level data and it opted to conduct more conservative tests for intrinsic message characteristics. Consequently, this dissertation only estimated the direct effects of intrinsic content features, but not their total effects (i.e., direct + indirect effects; Hayes, 2009; MacKinnon, 2008). This means that effects of intrinsic message features on various forms of news attractability and virality reported throughout this dissertation tend to be underestimated. Perhaps more importantly, this dissertation is mute on potentially theoretically meaningful indirect effects of intrinsic message features. Future work should further examine individual-level pathways that flow from intrinsic message properties to news selections and retransmissions through perceived message properties, which can advance our understanding of message effects on attractability and virality (Cappella, 2006; O'Keefe, 2003). Similarly, a related area for further investigation is the relationship between message features of news articles and editorial decision on the placement of the articles on news websites. Given the finding that article location plays a vital role in triggering news selections and retransmissions, identifying message effects on such editorial decisions might illuminate an important mechanism through which content features drive news diffusion.

Future research might also examine consequences of news propagations. Similar to other studies of message virality (e.g., Berger & Milkman, 2012), this dissertation focused on what drives messengers (sharers) to propagate messages. But retransmitted messages may also have consequences on the recipients of those messages, especially in terms of (1) further information seeking and (2) persuasion and behavior change. Future studies might investigate how content characteristics and relationship strengths affect the

likelihood that recipients of news-retransmission messages expose themselves to full news articles or seek out further relevant information (Pew Research Center, 2013b). Another important area for future work would be to identify how virality relates to persuasiveness (Kim et al., 2013) and how message features and relationship strengths shape the virality-persuasiveness link (Garrett, 2011; van Noort, Antheunis, & van Reijmersdal, 2012).

Conclusion

In conclusion, this dissertation contributes to our understanding of factors driving the volume and persistence of news selections and retransmissions in the emerging public communication environment. Results of this dissertation identify central message features that make health news more attractable and viral. The results also demonstrate that news retransmission channels shape what goes viral, and further show that social influence and its synergistic interactions with message features boost news attractability and virality. This dissertation makes methodological contributions to the literature by examining news propagations that are adjusted for selections and thereby estimating effects of message features and social influence on virality that is not confounded by attractability. It should also be highlighted that the computational social science method developed in this dissertation for automated data collection of time-series behavioral measures of news selections and retransmissions holds promise for future research on news attractability and virality. It is hoped that future work will advance this line of research by further clarifying social psychological mechanisms through which message features, social influence, and news retransmission channels drive health news diffusion.

APPENDICES

Appendix A. Message Evaluation Survey: Validity and Reliability (Chapter 3)

As the main goal of the message evaluation survey is to measure an evaluation (or perception) for each article teaser and full text, an ideal method to achieve the goal would be to use a *representative* and *large* sample of the general population. That is, one can obtain an evaluation score that is more valid (closer to the population value) and reliable with a more representative and larger sample.

Sample Representativeness

Using a convenience sample like the MTurk sample used in the current message evaluation survey can yield evaluation scores that are potentially less valid. The non-representative sample recruited via MTurk might generate less valid (or less accurate) aggregate responses with respect to the perceived message features described above. However, recruiting a representative sample of over 5,000 U.S. adults is costly, especially for the survey like the current one which involves (1) reading three article teasers and three full texts and (2) answering questions about them. Moreover, there is evidence that, compared to more traditionally-used convenience samples such as college students, the MTurk sample is more diverse and more similar to the general population (e.g., Berinsky et al., 2012). All in all, the MTurk sample was considered a realistic compromise between data quality (validity) and efficiency for this dissertation study.

Sample Size

It should also be noted that the aggregate evaluations obtained by the current survey design with each article being rated on average by 20 respondents are less reliable, as compared to when a larger number of respondents per article are sampled. However, while it is clear that one can obtain more reliable aggregate ratings by recruiting more respondents for each article, little is known about optimal cut-off criteria for the *average* number of respondents per article that ensures acceptable reliability, especially for survey studies that measure evaluations of news article texts and further relate aggregate responses to other outcomes of interest. The choice of the expected number of raters per article (derived from the survey design ≈ 20) in this dissertation was thus based on a qualitative review of several previous studies that (1) included evaluations of messages (of various kinds) and (2) analyzed aggregate ratings in relation to other outcome variables. For example, Bakshy and colleagues (2011) assessed subjective features of online

messages (shared via Twitter) by recruiting the MTurk sample with a survey design where each message was rated by on average 11 respondents (range: 3 to 20). Durkin, Biener, and Wakefield (2009) employed 18 raters to evaluate the emotional intensity of antismoking televised ads. Bigsby, Cappella, and Seitz (2013) measured the perceived effectiveness of antismoking televised public service announcements (PSAs) by recruiting on average about 46 adult smokers for each PSA. Considering these and other previous studies, I concluded that designing a message evaluation survey with the expected number of respondents per article being 20 is a workable trade-off between data quality (reliability) and efficiency.

Inter-Respondent Agreement on Message Evaluations

As discussed earlier, a great deal of agreement among respondents' ratings about a target article text is not crucial for obtaining reliable aggregate message evaluations. Rather, the reliability of aggregate evaluations is largely dependent on the number of respondents per article (either article teaser or full text) because the unit of analysis in this dissertation is the article and respondents' ratings are aggregated across the respondents by article and examined in relation to other article-related outcomes.

Nonetheless, it would be informative to examine the inter-respondent agreement on message evaluations because this can help understand the variability of individual evaluations on various rating items and guide interpretation of study results. Before presenting the relevant statistics, it should be noted that a conventional measure of inter-respondent agreement (McGraw & Wong, 1996; Shrout & Fleiss, 1979) was unable to be obtained using the present message evaluation data, due to the complex survey design (1) with a large number of articles and respondents and (2) cross-classified multilevel structure between article- and respondent-level (and varying number of respondents per article).

Therefore, I calculated an approximate inter-respondent agreement coefficient by simplifying the data structure. Specifically, the coefficient was obtained by the following simplified one-way random effects analysis of variance (ANOVA) model which treats the article-factor as a random effect and the respondent-factor as measurement error:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \text{(Equation A-1)}$$

where y_{ij} is an article i's evaluation score given by a respondent j, α_i is an article random effect, and ε_{ij} is an idiosyncratic error term. Then, the intraclass correlation coefficient (ICC) for α_i was estimated using the following equation:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \qquad \text{(Equation A-2)}$$

where σ_{α}^{2} is the between-article variance (i.e., variance component attributable to articles), and σ_{ε}^{2} is the residual variance of the evaluation score y_{ij} in Equation A-1. As the unit of analysis in this dissertation is the article, an approximate measure of inter-respondent agreement was based on aggregate evaluations (i.e., averaged scores by article) rather than a single evaluation (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Finally, the agreement coefficient for the *k* number of respondents per article (ρ_k) was obtained using the Spearman-Brown formula (Shrout & Fleiss, 1979; Winer, Brown, & Michels, 1991):

$$\rho_k = \frac{k\rho}{1 + (k-1)\rho} \qquad \text{(Equation A-3)}$$

where ρ is the ICC estimated from Equations A-1 and A-2. Established measures of inter-rater agreement are typically based on the message evaluation design where *k* (the number of raters per message) is a fixed number (i.e., all messages are evaluated by the same number of raters). However, the current message evaluation survey design introduced random variation into *k* across article texts. Therefore, another approximation was made such that the inter-respondent agreement coefficient was calculated with *k* being 20.1 (i.e., the average number of raters per article in this survey). The analysis revealed that the between-article variance (i.e., α_i's variance) is statistically significantly different from zero for all evaluation items (all *p*-values < .001) and for both article teasers and full texts, suggesting that article texts explained a significant portion of the variance in the evaluation score y_{ij} in Equation A-1. Table A-1 presents approximate inter-respondent agreement measures obtained by the procedures described above. Overall, there was a reasonably high level of agreement across respondents with regard to article evaluations, except for the degree of emotional arousal (evocativeness) that articles induced.

Table A-1. Approximate Inter-Respondent Agreement Coefficients

	Article Teaser	Article Full Text
a) Pride	.73	.79
b) Amusement	.76	.80
c) Contentment	.68	.74
d) Hope	.88	.87
e) Anger	.85	.89
f) Fear	.82	.84
g) Sadness	.88	.90
Positivity Scale (items *a* to *g*) *	.91	.92
h) Newness	.76	.75
i) Unusualness	.77	.76
j) Surprise	.69	.64
Novelty Scale (items *h* to *j*) *	.80	.78
k) Emotional Arousal (Evocativeness)	.19	.16
l) Controversiality	.82	.83
m) Usefulness	.75	.67

* Details about the construction of positivity and novelty scales are discussed in Chapter 4.

Appendix B. Message Effects on News Attractability: Full Results (Chapter 4)

	Bivariate Regression	Multiple Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Content Factors (*df* = 18)			$\Delta R^2 = .17^{***}$
Efficacy Information Present	$.35^{**}$ (.13)	$.30^{*}$ (.13)	$.34^{**}$ (.13)
Usefulness	−.01 (.12)	.03 (.11)	.02 (.11)
Emotional Positivity (Responses)	$.29^{*}$ (.14)	.11 (.15)	$.65^{**}$ (.22)
Expressed Positivity (Words)	.02 (.03)	−.01 (.03)	−.01 (.03)
Controversiality	.14 (.09)	$.18^{*}$ (.09)	$.25^{**}$ (.09)
Emotional Arousal (Responses)	$.74^{**}$ (.23)	.31 (.20)	.31 (.20)
Expressed Emotionality (Words) [a]	$.23^{*}$ (.09)	$.16^{+}$ (.08)	$.16^{*}$ (.08)
Novelty	$-.22^{+}$ (.12)	−.19 (.12)	$-.23^{*}$ (.12)
Diseases / Bad Health Conditions Mentioned	$-.26^{*}$ (.11)	$-.30^{**}$ (.11)	$-.27^{*}$ (.11)
Positivity (Responses) × Diseases			$-.85^{***}$ (.25)
Professional Sources Mentioned	$-.33^{**}$ (.12)	$-.28^{**}$ (.10)	$-.27^{**}$ (.10)
Death-Related Words Present	.02 (.17)	.08 (.15)	.04 (.15)
Health Words [a]	$.19^{*}$ (.09)	−.01 (.08)	.02 (.08)
Social-Processes Words [a]	$.17^{*}$ (.08)	.03 (.07)	.03 (.07)
Word Count	$.03^{***}$ (.01)	.01 (.01)	.01 (.01)
Writing Complexity (Words > 6 Letters)	$.04^{*}$ (.02)	.01 (.02)	.01 (.01)
Article Category (Reference = "Well")			
The New Old Age	-1.35^{***} (.15)	$-.72^{***}$ (.16)	$-.66^{***}$ (.16)
Other	$-.25^{*}$ (.11)	$-.52^{***}$ (.10)	$-.50^{***}$ (.10)
Context Factors (*df* = 10)			$\Delta R^2 = .20^{***}$
Total Hours Shown in Prominent Locations [a]	$.48^{***}$ (.03)	$.44^{***}$ (.03)	$.44^{***}$ (.03)
Publication Month (Reference = July 2012)			
August 2012	.02 (.23)	.13 (.19)	.12 (.19)
September 2012	.02 (.23)	.05 (.19)	.01 (.19)
October 2012	−.30 (.22)	−.29 (.19)	$-.35^{+}$ (.19)
November 2012	$-.37^{+}$ (.22)	−.16 (.19)	−.17 (.18)
December 2012	−.13 (.23)	−.002 (.19)	−.02 (.19)
January 2013	.11 (.22)	.13 (.18)	.10 (.18)
February 2013	.16 (.23)	.19 (.19)	.18 (.19)
Publication Day of the Week (Ref. = Monday)			
Tuesday to Friday	$.29^{**}$ (.11)	.07 (.10)	.06 (.10)
Saturday & Sunday	$.95^{***}$ (.26)	$.63^{**}$ (.23)	$.63^{**}$ (.23)
Final Model R^2		$.36^{***}$	$.37^{***}$

Note. $N = 758$ for the multiple regression models (Model 1 & 2). Dependent variables were log-transformed. Cell entries are unstandardized OLS regression coefficients (*b*) with standard errors (*se*) in parentheses. Emotional Positivity (Responses) was mean-centered before entry (Model 2). All variance inflation factors (VIFs) for Model 2 < 2.35. [a] Log-transformed. $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Appendix C. Message Effects on News Virality: Full Results (Chapter 4)

	Bivariate Regression	Multiple Regression
	b (se)	*b (se)*
Content Factors (*df* = 25)		$\Delta R^2 = .49^{***}$
Efficacy Information Present	$.63^{***}$ (.12)	$.13^{*}$ (.06)
Usefulness	$.99^{***}$ (.15)	$.50^{***}$ (.07)
Emotional Positivity (Responses)	$.57^{***}$ (.12)	$.19^{**}$ (.06)
Expressed Positivity (Words)	$.09^{**}$ (.03)	.01 (.01)
Controversiality	.11 (.09)	−.01 (.05)
Emotional Arousal (Responses)	$.95^{***}$ (.23)	.10 (.10)
Expressed Emotionality (Words)	$.08^{*}$ (.03)	.02 (.01)
Novelty	$.35^{**}$ (.14)	.05 (.06)
Exemplification	$.44^{***}$ (.12)	.03 (.06)
Credibility Statements		
1	−.30 (.21)	−.01 (.11)
2+ with no opposing statements	$.65^{**}$ (.20)	−.05 (.11)
2+ with opposing statements	$.68^{**}$ (.24)	−.15 (.13)
Topic (Reference = Health Policy)		
Disease / Health Conditions	.16 (.16)	−.02 (.08)
Other	−.33 (.21)	−.01 (.09)
Writing Style – 1st Person Point of View	−.02 (.14)	.05 (.07)
Death-Related Words Present	−.09 (.11)	−.04 (.05)
Health Words [a]	.07 (.11)	−.01 (.05)
Social-Processes Words [a]	$.48^{***}$ (.13)	.05 (.06)
Word Count × 10^{-2}	$.19^{***}$ (.01)	$.03^{***}$ (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	−.16 (.13)	$.20^{**}$ (.07)
$(\text{Writing Complexity})^2$		$-.17^{+}$ (.10)
Images Present	$.78^{***}$ (.11)	.03 (.07)
Number of Hyperlinks [a]	$.49^{***}$ (.07)	$.06^{+}$ (.04)
Article Category (Reference = "Well")		
The New Old Age	-1.28^{***} (.16)	.04 (.11)
Other	−.13 (.11)	.09 (.07)
Context Factors (*df* = 10)		$\Delta R^2 = .05^{***}$
Total Hours Shown in Prominent Locations [a]	$.48^{***}$ (.03)	$.04^{*}$ (.02)
Publication Month (Reference = July 2012)		
August 2012	−.19 (.23)	$-.18^{+}$ (.09)
September 2012	.28 (.23)	$.19^{*}$ (.09)
October 2012	.16 (.23)	$.42^{***}$ (.09)
November 2012	$-.61^{**}$ (.23)	$-.23^{**}$ (.09)
December 2012	−.22 (.23)	−.11 (.09)
January 2013	−.04 (.22)	−.08 (.09)
February 2013	.10 (.24)	.01 (.09)
Publication Day of the Week (Reference = Monday)		
Tuesday to Friday	.11 (.11)	$-.13^{**}$ (.05)
Saturday & Sunday	$.93^{***}$ (.27)	−.13 (.11)
Selection (*df* = 1)		$\Delta R^2 = .32^{***}$
Total Number of Selections [a]	$.92^{***}$ (.02)	$.84^{***}$ (.02)
Final Model R^2		$.86^{***}$

Note. $N = 758$ for the multiple regression model. Dependent variables were log-transformed. Cell entries are unstandardized OLS regression coefficients (*b*) with standard errors (*se*) in parentheses. Writing Complexity was mean-centered. All variance inflation factors (VIFs) for the multiple regression model < 3.30. [a] Log-transformed. $^{+}\ p < .10$, $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$.

Appendix D. Retransmission Channels and News Virality: Full Results (Chapter 4)

	Retransmission Channel	
	Email	Social Media
Content Factors (*df* = 25)		
Efficacy Information Present	.19** (.06)	.002 (.06)
Usefulness	.66*** (.08)	.22** (.08)
Emotional Positivity (Responses)	.17* (.07)	.26*** (.07)
Expressed Positivity (Words)	.01 (.02)	.01 (.02)
Controversiality	−.03 (.06)	.06 (.05)
Emotional Arousal (Responses)	−.02 (.11)	.34** (.11)
Expressed Emotionality (Words)	.03+ (.02)	.02 (.02)
Novelty	.17* (.07)	−.16* (.07)
Exemplification	−.005 (.06)	.12* (.06)
Credibility Statements		
1	.04 (.12)	−.06 (.11)
2+ with no opposing statements	−.03 (.12)	−.03 (.11)
2+ with opposing statements	−.11 (.15)	−.16 (.14)
Topic (Reference = Health Policy)		
Disease / Health Conditions	.01 (.09)	−.03 (.08)
Other	.03 (.11)	−.04 (.10)
Writing Style – 1st Person Point of View	.12 (.08)	−.08 (.07)
Death-Related Words Present	−.08 (.06)	−.005 (.05)
Health Words [a]	.04 (.06)	−.05 (.06)
Social-Processes Words [a]	.02 (.07)	.05 (.07)
Word Count × 10^{-2}	.05*** (.01)	.02 (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.25*** (.08)	.13+ (.07)
(Writing Complexity)^2	−.11 (.12)	−.27* (.11)
Images Present	−.04 (.08)	.11 (.07)
Number of Hyperlinks [a]	.08* (.04)	.03 (.04)
Article Category (Reference = "Well")		
The New Old Age	.23+ (.13)	−.37** (.12)
Other	.03 (.08)	.18* (.07)
Context Factors (*df* = 10)		
Total Hours Shown in Prominent Locations [a]	.04* (.02)	.03 (.02)
Publication Month (Reference = July 2012)		
August 2012	−.15 (.11)	−.09 (.10)
September 2012	.24* (.1)	.28** (.10)
October 2012	.50*** (.11)	.43*** (.10)
November 2012	−.26* (.10)	−.13 (.10)
December 2012	−.04 (.11)	−.10 (.10)
January 2013	.03 (.10)	−.11 (.09)
February 2013	.10 (.11)	−.03 (.10)
Publication Day of the Week (Reference = Monday)		
Tuesday to Friday	−.24*** (.06)	.09 (.05)
Saturday & Sunday	−.21 (.13)	.05 (.12)
Selection (*df* = 1)		
Total Number of Selections [a]	.87*** (.02)	.79*** (.02)
R^2	.83***	.83***
Residual Correlation	.44***	

Note. N = 758. Dependent variables were log-transformed. Cell entries are unstandardized regression coefficients (*b*) with standard errors (*se*) in parentheses. Writing Complexity was mean-centered. [a] Log-transformed. + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix E. The Interplay of Social Influence and Message Features in Driving News Virality (Chapter 5)

Table E-1. Social Influence Effects on Email Retransmissions Moderated by Message Features (Hourly Data)

		Emotional Responses - Positivity (L)		Emotional Responses - Positivity (H)	
		Usefulness (L)	Usefulness (H)	Usefulness (L)	Usefulness (H)
Efficacy Information	Absent	.52*** (.02)	.62*** (.01)	.57*** (.02)	.67*** (.01)
	Present	.58*** (.02)	.68*** (.01)	.63*** (.01)	.73*** (.02)

Note. Results are based on the fixed effects regression model shown in Model 2 (hourly data) in Table 5-2 (Chapter 5). Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors (the Driscoll-Kraay estimator) in parentheses for the logged number of hours shown on the "most-emailed" list (i.e., social influence indicator; lagged). For continuous moderators: L = M – 1SD; H = M + 1SD. *** $p < .001$.

Table E-2. Social Influence Effects on Email Retransmissions Moderated by Message Features (Daily Data)

		Emotional Responses - Positivity (L)		Emotional Responses - Positivity (H)	
		Usefulness (L)	Usefulness (H)	Usefulness (L)	Usefulness (H)
Efficacy Information	Absent	.29*** (.06)	.44*** (.04)	.37*** (.04)	.52*** (.03)
	Present	.37*** (.04)	.52*** (.04)	.45*** (.03)	.60*** (.04)

Note. Results are based on the fixed effects regression model shown in Model 2 (daily data) in Table 5-2 (Chapter 5). Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors (the Driscoll-Kraay estimator) in parentheses for the logged number of hours shown on the "most-emailed" list (i.e., social influence indicator; lagged). For continuous moderators: L = M – 1SD; H = M + 1SD. *** $p < .001$.

Table E-3. Social Influence Effects on Social Media Retransmissions Moderated by Message Features (Hourly Data)

		Emotional Responses - Positivity (L)				Emotional Responses - Positivity (H)			
		Positivity - Words (L)		Positivity - Words (H)		Positivity - Words (L)		Positivity - Words (H)	
		Useful (L)	Useful (H)	Useful (L)	Useful (H)	Useful (L)	Useful (H)	Useful (L)	Useful (H)
Efficacy Information Absent	Exemplars Absent	.32*** (.02)	.35*** (.01)	.38*** (.02)	.41*** (.01)	.39*** (.01)	.42*** (.01)	.45*** (.02)	.48*** (.01)
	Exemplars Present	.35*** (.01)	.37*** (.01)	.41*** (.02)	.43*** (.01)	.42*** (.02)	.44*** (.02)	.48*** (.02)	.50*** (.02)
Efficacy Information Present	Exemplars Absent	.34*** (.01)	.37*** (.01)	.40*** (.02)	.43*** (.01)	.41*** (.01)	.43*** (.01)	.47*** (.02)	.50*** (.01)
	Exemplars Present	.37*** (.01)	.39*** (.01)	.43*** (.02)	.45*** (.01)	.44*** (.01)	.46*** (.02)	.50*** (.02)	.52*** (.02)

Note. Results are based on the fixed effects regression model shown in Model 2 (hourly data) in Table 5-3 (Chapter 5). Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors (the Driscoll-Kraay estimator) in parentheses for the logged number of hours shown on the "most-emailed" list (i.e., social influence indicator; lagged). For continuous moderators: L = M – 1SD; H = M + 1SD. *** $p < .001$.

Table E-4. Social Influence Effects on Social Media Retransmissions Moderated by Message Features (Daily Data)

		Emotional Responses - Positivity (L)				Emotional Responses - Positivity (H)			
		Positivity - Words (L)		Positivity - Words (H)		Positivity - Words (L)		Positivity - Words (H)	
		Useful (L)	Useful (H)	Useful (L)	Useful (H)	Useful (L)	Useful (H)	Useful (L)	Useful (H)
Efficacy Information Absent	Exemplars Absent	.18*** (.03)	.24*** (.02)	.26*** (.04)	.32*** (.03)	.29*** (.03)	.35*** (.03)	.37*** (.03)	.43*** (.03)
	Exemplars Present	.23*** (.03)	.29*** (.03)	.31*** (.04)	.37*** (.03)	.34*** (.03)	.39*** (.03)	.42*** (.03)	.47*** (.03)
Efficacy Information Present	Exemplars Absent	.22*** (.02)	.27*** (.02)	.30*** (.03)	.35*** (.02)	.32*** (.02)	.38*** (.03)	.40*** (.03)	.46*** (.03)
	Exemplars Present	.26*** (.03)	.32*** (.03)	.34*** (.03)	.40*** (.02)	.37*** (.03)	.43*** (.04)	.45*** (.03)	.51*** (.03)

Note. Results are based on the fixed effects regression model shown in Model 2 (daily data) in Table 5-3 (Chapter 5). Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors (the Driscoll-Kraay estimator) in parentheses for the logged number of hours shown on the "most-emailed" list (i.e., social influence indicator; lagged). For continuous moderators: L = M – 1SD; H = M + 1SD. *** $p < .001$.

Appendix F. Temporal Dynamics Models of Attractability and Virality: an Alternative Autocorrelation Specification (Chapter 5)

Table F-1. Fixed Effects Models of Attractability: Results Based on an Alternative Autocorrelation Specification

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Viewed List [a, b]	2.35*** (.21)	1.12*** (.11)	1.10*** (.11)	1.94*** (.17)	.85*** (.19)	.83*** (.19)
MV List × Efficacy Information [c]			.07*** (.01)			.07* (.03)
Hours Shown in Prominent Locations [a, b]	2.93*** (.26)	.46*** (.07)	.46*** (.07)	2.28*** (.21)	.32*** (.09)	.32*** (.09)
Time Since Online Publication [a]	−2.38*** (.21)	−1.62*** (.19)	−1.62*** (.19)	−2.74*** (.32)	−1.81*** (.38)	−1.81*** (.38)
Within R^2		.66***	.66***		.61***	.61***
N		541,095	541,095		21,995	21,995

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to 293 lags for hourly data and 11 lags for daily data. All variance inflation factors (VIFs) for Model 2 < 2.11 for hourly data (< 2.21 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). * $p < .05$, *** $p < .001$.

Table F-2. Fixed Effects Models of Virality (Email Propagations): Results Based on an Alternative Autocorrelation Specification

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Emailed List [a, b]	1.31*** (.11)	.64*** (.04)	.60*** (.04)	1.00*** (.09)	.46*** (.05)	.41*** (.05)
ME List × Efficacy Information [c]			.06*** (.01)			.08*** (.02)
ME List × Usefulness [c]			.15*** (.03)			.22*** (.04)
ME List × Positivity (Responses) [c]			.05*** (.01)			.09*** (.02)
Selection Count [a]	.39*** (.04)	.16*** (.01)	.16*** (.01)	.39*** (.04)	.16*** (.01)	.16*** (.01)
Hours Shown in Prominent Locations [a, b]	1.57*** (.12)	.28*** (.02)	.30*** (.02)	1.18*** (.09)	.24*** (.03)	.26*** (.03)
Time Since Online Publication [a]	−1.12*** (.11)	−.34*** (.05)	−.35*** (.05)	−1.27*** (.18)	−.29*** (.06)	−.29*** (.06)
Within R^2		.82***	.82***		.74***	.74***
N		540,390	539,694		21,995	21,966

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to 268 lags for hourly data and 10 lags for daily data. All variance inflation factors (VIFs) for Model 2 < 2.45 for hourly data (< 2.47 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). *** p < .001.

Table F-3. Fixed Effects Models of Virality (Social Media Propagations): Results Based on an Alternative Autocorrelation Specification

	Hourly Data			Daily Data		
	Bivariate FE Regression	Multiple FE Regression		Bivariate FE Regression	Multiple FE Regression	
		Model 1	Model 2		Model 1	Model 2
Hours Shown on the Most-Emailed List [a, b]	1.07*** (.10)	.43*** (.03)	.40*** (.03)	.82*** (.09)	.35*** (.03)	.30*** (.03)
ME List × Efficacy Information [c]			.02* (.01)			.03* (.01)
ME List × Usefulness [c]			.04+ (.02)			.09** (.03)
ME List × Positivity (Responses) [c]			.08*** (.01)			.13*** (.01)
ME List × Positivity (Words) [c]			.02*** (.003)			.02*** (.004)
ME List × Exemplification [c]			.03*** (.01)			.05*** (.01)
Selection Count [a]	.33*** (.04)	.15*** (.01)	.15*** (.01)	.34*** (.04)	.14*** (.01)	.14*** (.01)
Hours Shown in Prominent Locations [a, b]	1.43*** (.11)	.39*** (.04)	.40*** (.04)	1.05*** (.08)	.31*** (.02)	.33*** (.02)
Time Since Online Publication [a]	−1.00*** (.11)	−.31*** (.05)	−.31*** (.05)	−1.09*** (.17)	−.20*** (.05)	−.20*** (.05)
Within R^2		.78***	.78***		.70***	.70***
N		539,217	538,521		21,995	21,966

Note. Dependent variables were log-transformed. Cell entries are unstandardized fixed effects (within) regression coefficients with robust standard errors in parentheses. The Driscoll-Kraay estimator was used to obtain standard errors that are robust to autocorrelated, cross-sectionally dependent, and heteroskedastic model residuals. Residuals were allowed to be serially correlated up to 292 lags for hourly data and 10 lags for daily data. All variance inflation factors (VIFs) for Model 2 < 3.02 for hourly data (< 3.04 for daily data). [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix G. Message Effects on the First-Time Appearing on the "Most-Viewed" List: Full Results (Chapter 5)

	Bivariate Cox Regression	Multiple Cox Regression
	b (se)	*b (se)*
Efficacy Information Present	.03 (.10)	.06 (.13)
Usefulness	−.05 (.08)	−.03 (.10)
Emotional Positivity (Responses)	−.07 (.11)	−.06 (.15)
Expressed Positivity (Words)	−.002 (.02)	−.04 (.03)
Controversiality	.19** (.07)	.18* (.09)
Emotional Arousal (Responses)	.27 (.17)	.10 (.19)
Expressed Emotionality (Words) [a]	.11 (.07)	.06 (.08)
Novelty	−.18+ (.10)	−.04 (.12)
Diseases / Bad Health Conditions Mentioned	−.06 (.08)	−.22* (.11)
Professional Sources Mentioned	−.18* (.09)	−.18+ (.10)
Death-Related Words Present	.13 (.16)	.12 (.18)
Health Words [a]	.24*** (.07)	.001 (.08)
Social-Processes Words [a]	.18** (.06)	.12+ (.07)
Word Count	.03*** (.01)	.002 (.01)
Writing Complexity (Words > 6 Letters)	.03** (.01)	.01 (.01)
Article Category (Reference = "Well")		
The New Old Age	−.81*** (.13)	−.83*** (.18)
Other	.03 (.09)	−.30** (.11)
Shown in Prominent Locations [b]	1.11*** (.09)	1.04*** (.10)
Publication Month (Reference = July 2012)		
August 2012	−.22 (.18)	−.15 (.19)
September 2012	−.39* (.17)	−.58** (.19)
October 2012	.17 (.18)	.24 (.18)
November 2012	−.44* (.18)	−.39* (.19)
December 2012	−.18 (.19)	−.09 (.19)
January 2013	−.21 (.18)	−.29 (.19)
February 2013	−.08 (.18)	−.03 (.19)
Publication Day of the Week (Reference = Monday)		
Tuesday to Friday	.40*** (.08)	.66*** (.10)
Saturday & Sunday	.81*** (.23)	.84*** (.25)
Generalized (Cox-Snell) R^2		.32

Note. N = 109,652 for the multiple Cox regression model. Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. [a] Log-transformed. [b] Lagged. + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix H. Message Effects on the First-Time Appearing on the "Most-Emailed" List: Full Results (Chapter 5)

	Bivariate Cox Regression	Multiple Cox Regression
	b (se)	*b (se)*
Efficacy Information Present	.21* (.10)	.24* (.11)
Usefulness	.72*** (.13)	.75*** (.15)
Emotional Positivity (Responses)	.19^{+} (.10)	.21 (.13)
Expressed Positivity (Words)	.04^{+} (.02)	−.02 (.03)
Controversiality	.18* (.07)	−.13 (.11)
Emotional Arousal (Responses)	.66*** (.18)	−.17 (.19)
Expressed Emotionality (Words)	.06* (.03)	.05^{+} (.03)
Novelty	.16 (.11)	.19 (.13)
Exemplification	.38*** (.09)	−.01 (.11)
Credibility Statements		
1	−.55*** (.14)	.03 (.19)
2+ with no opposing statements	.38** (.12)	.08 (.19)
2+ with opposing statements	.42** (.16)	−.02 (.23)
Topic (Reference = Health Policy)		
Disease / Health Conditions	−.02 (.11)	−.03 (.14)
Other	−.27^{+} (.16)	−.21 (.19)
Writing Style – 1st Person Point of View	.20* (.09)	.21^{+} (.13)
Death-Related Words Present	.13 (.08)	−.10 (.09)
Health Words [a]	.08 (.09)	−.13 (.11)
Social-Processes Words [a]	.42*** (.10)	.09 (.15)
Word Count × 10^{-2}	.15*** (.01)	.11*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	−.17^{+} (.09)	.26^{+} (.15)
(Writing Complexity)^{2}		−.37^{+} (.20)
Images Present	.14 (.09)	−.22 (.15)
Number of Hyperlinks [a]	.53*** (.07)	.20** (.07)
Article Category (Reference = "Well")		
The New Old Age	−.49*** (.13)	−.10 (.24)
Other	.10 (.09)	−.20 (.15)
Shown in Prominent Locations [b]	1.07*** (.09)	.49*** (.10)
Shown on the "Most-Viewed" List	2.26*** (.10)	2.11*** (.12)
Publication Month (Reference = July 2012)		
August 2012	−.11 (.20)	−.09 (.20)
September 2012	−.22 (.19)	−.12 (.20)
October 2012	.22 (.18)	−.04 (.20)
November 2012	−.30 (.19)	−.16 (.19)
December 2012	−.12 (.20)	.001 (.19)
January 2013	−.10 (.19)	.10 (.18)
February 2013	−.04 (.20)	.20 (.20)
Publication Day of the Week (Reference = Monday)		
Tuesday to Friday	.31*** (.09)	.09 (.10)
Saturday & Sunday	.90*** (.23)	.26 (.24)
Generalized (Cox-Snell) R^2		.67

Note. N = 144,967 for the multiple Cox regression model. Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. Writing Complexity was mean-centered. [a] Log-transformed. [b] Lagged. $^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix I. The Impact of Social Influence and Message Features on the Persistence of News Attractability: Full Results (Chapter 6)

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.15 (.10)	.07 (.11)	.07 (.11)
Usefulness	−.02 (.08)	−.001 (.10)	.002 (.10)
Emotional Positivity (Responses)	−.36*** (.10)	−.32* (.13)	−.32* (.13)
Expressed Positivity (Words)	−.02 (.02)	−.002 (.02)	−.002 (.02)
Controversiality	.09 (.06)	−.09 (.08)	−.08 (.08)
Emotional Arousal (Responses)	−.29+ (.15)	−.22 (.17)	−.22 (.17)
Expressed Emotionality (Words) [a]	−.25*** (.06)	−.17* (.07)	−.25** (.09)
Novelty	.18* (.09)	.14 (.11)	.14 (.11)
Diseases / Bad Health Conditions Mentioned	.19** (.07)	−.005 (.10)	−.0004 (.10)
Professional Sources Mentioned	.14+ (.08)	.03 (.08)	.03 (.08)
Death-Related Words Present	.11 (.11)	−.07 (.12)	−.07 (.12)
Health Words [a]	.09 (.06)	.10 (.07)	.10 (.07)
Social-Processes Words [a]	−.11+ (.05)	−.02 (.06)	−.02 (.06)
Word Count	−.01 (.004)	−.0002 (.01)	−.0004 (.01)
Writing Complexity (Words > 6 Letters)	−.01 (.01)	−.01 (.01)	−.01 (.01)
Article Category (Reference = "Well")			
The New Old Age	.20* (.08)	.11 (.10)	.12 (.10)
Other	.47*** (.09)	.41*** (.09)	.42*** (.09)
Hours Shown in Prominent Locations [a, b]	−.24** (.09)	−.14 (.09)	−.15 (.09)
Hours Shown on the Most-Viewed (MV) List [a, b]	−.30*** (.04)	−.26*** (.04)	−.27*** (.04)
MV List × Expressed Emotionality (Words) [c]			−.11* (.05)
Publication Month (Reference = July 2012)			
August 2012	−.14 (.12)	−.08 (.11)	−.08 (.11)
September 2012	−.07 (.18)	−.04 (.17)	−.04 (.17)
October 2012	.90*** (.18)	1.02*** (.18)	1.02*** (.18)
November 2012	.45*** (.11)	.34** (.11)	.35** (.11)
December 2012	.16 (.11)	.18 (.11)	.18 (.11)
January 2013	.11 (.11)	.06 (.11)	.06 (.11)
February 2013	.10 (.12)	.09 (.12)	.09 (.12)
Publication Day of the Week (Reference = Monday)			
Tuesday to Friday	−.15* (.08)	−.13 (.09)	−.13 (.09)
Saturday & Sunday	−.18 (.21)	−.40+ (.23)	−.41+ (.23)
Generalized (Cox-Snell) R^2		.22	.22

Note. N = 5,998 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix J. The Impact of Social Influence and Message Features on the Persistence of News Virality (Email): Full Results (Chapter 6)

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.36*** (.08)	−.21* (.09)	−.21* (.09)
Usefulness	−.70*** (.11)	−.29* (.13)	−.29* (.13)
Emotional Positivity (Responses)	−.25** (.09)	−.11 (.11)	−.11 (.11)
Expressed Positivity (Words)	−.02 (.02)	.0002 (.02)	.003 (.02)
Controversiality	.01 (.06)	−.10 (.08)	−.10 (.08)
Emotional Arousal (Responses)	−.38* (.15)	−.08 (.18)	−.09 (.18)
Expressed Emotionality (Words)	−.14*** (.03)	−.07** (.02)	−.20*** (.05)
Novelty	−.29** (.09)	−.17 (.11)	−.17 (.11)
Exemplification	−.23** (.08)	.05 (.09)	.06 (.09)
Credibility Statements			
1	−.11 (.12)	−.03 (.16)	−.03 (.16)
2+ with no opposing statements	−.47*** (.11)	−.09 (.16)	−.09 (.16)
2+ with opposing statements	−.17 (.16)	.20 (.20)	.20 (.20)
Topic (Reference = Health Policy)			
Disease / Health Conditions	−.32** (.11)	−.27* (.12)	−.28* (.12)
Other	−.28+ (.15)	−.35* (.15)	−.36* (.15)
Writing Style – 1st Person Point of View	−.13 (.09)	.03 (.11)	.04 (.11)
Death-Related Words Present	.07 (.07)	.16* (.08)	.17* (.08)
Health Words [a]	.02 (.08)	−.003 (.09)	.002 (.09)
Social-Processes Words [a]	−.41*** (.09)	−.04 (.10)	−.03 (.10)
Word Count × 10^{-2}	−.08*** (.01)	−.08*** (.01)	−.08*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.07 (.09)	.05 (.12)	.06 (.12)
Images Present	−.27*** (.08)	−.13 (.13)	−.13 (.12)
Number of Hyperlinks [a]	−.18*** (.05)	−.15** (.06)	−.15** (.06)
Article Category (Reference = "Well")			
The New Old Age	−.001 (.09)	−.31+ (.16)	−.32* (.16)
Other	.34*** (.09)	.38*** (.11)	.38*** (.11)
Hours Shown in Prominent Locations [a, b]	−.41** (.15)	−.23+ (.14)	−.24+ (.14)
Hours Shown on the Most-Emailed (ME) List [a, b]	−.73*** (.06)	−.60*** (.06)	−.65*** (.07)
ME List × Expressed Emotionality (Words) [c]			−.14** (.05)
Publication Month (Reference = July 2012)			
August 2012	−.08 (.14)	.05 (.15)	.06 (.15)
September 2012	−.29* (.14)	−.24+ (.14)	−.24+ (.14)
October 2012	.02 (.14)	.04 (.14)	.04 (.14)
November 2012	.47** (.15)	.48** (.15)	.47** (.15)
December 2012	−.08 (.15)	.17 (.14)	.17 (.14)
January 2013	−.06 (.14)	−.09 (.15)	−.09 (.15)
February 2013	−.23 (.16)	−.26 (.17)	−.26 (.16)
Publication Day of the Week (Reference = Monday)			
Tuesday to Friday	.09 (.07)	.13 (.08)	.13 (.08)
Saturday & Sunday	−.03 (.19)	.04 (.23)	.02 (.23)
Generalized (Cox-Snell) R^2		.39	.40

Note. N = 6,290 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Appendix K. The Impact of Social Influence and Message Features on the Persistence of News Virality (Social Media): Full Results (Chapter 6)

	Bivariate Cox Regression	Multiple Cox Regression	
		Model 1	Model 2
	b (se)	*b (se)*	*b (se)*
Efficacy Information Present	−.26** (.08)	−.01 (.09)	−.01 (.09)
Usefulness	−.41*** (.11)	−.14 (.12)	−.14 (.12)
Emotional Positivity (Responses)	−.42*** (.09)	−.25* (.11)	−.26* (.11)
Expressed Positivity (Words)	−.08*** (.02)	−.05* (.02)	−.05* (.02)
Controversiality	.04 (.06)	−.20* (.09)	−.19* (.09)
Emotional Arousal (Responses)	−.58*** (.16)	−.24 (.18)	−.24 (.18)
Expressed Emotionality (Words)	−.13*** (.03)	−.09*** (.02)	−.09*** (.02)
Novelty	−.18+ (.10)	−.04 (.12)	−.03 (.12)
Exemplification	−.23** (.07)	−.02 (.09)	−.32+ (.19)
Credibility Statements			
1	.29* (.13)	.20 (.17)	.20 (.17)
2+ with no opposing statements	−.07 (.12)	.05 (.17)	.04 (.17)
2+ with opposing statements	.15 (.17)	.29 (.21)	.27 (.21)
Topic (Reference = Health Policy)			
Disease / Health Conditions	−.23* (.11)	−.25* (.12)	−.24+ (.12)
Other	−.19 (.14)	−.23 (.15)	−.22 (.15)
Writing Style – 1st Person Point of View	−.14+ (.08)	.08 (.11)	.08 (.11)
Death-Related Words Present	.13+ (.07)	.17* (.08)	.42** (.14)
Health Words [a]	.11 (.08)	−.02 (.09)	−.03 (.09)
Social-Processes Words [a]	−.44*** (.09)	−.05 (.10)	−.06 (.10)
Word Count × 10^{-2}	−.08*** (.01)	−.05*** (.01)	−.05*** (.01)
Writing Complexity ([% words > 6 letters] × 10^{-1})	.26** (.10)	.32** (.12)	.33** (.12)
Images Present	−.42*** (.08)	−.35** (.12)	−.34** (.12)
Number of Hyperlinks [a]	−.13** (.05)	−.16* (.06)	−.16* (.06)
Article Category (Reference = "Well")			
The New Old Age	.29** (.10)	−.05 (.16)	−.04 (.16)
Other	.34*** (.09)	.36** (.11)	.37** (.11)
Hours Shown in Prominent Locations [a, b]	−.74*** (.21)	−.53** (.19)	−.55** (.19)
Hours Shown on the Most-Emailed (ME) List [a, b]	−.64*** (.06)	−.50*** (.06)	−.59*** (.09)
ME List × Exemplification [c]			−.34* (.17)
ME List × Death-Related Words Present [c]			.28* (.11)
Publication Month (Reference = July 2012)			
August 2012	−.01 (.13)	.07 (.14)	.07 (.14)
September 2012	−.26+ (.14)	−.26+ (.14)	−.26+ (.14)
October 2012	−.06 (.13)	−.001 (.14)	.004 (.13)
November 2012	.25+ (.14)	.19 (.14)	.18 (.14)
December 2012	.01 (.14)	.18 (.13)	.19 (.13)
January 2013	.02 (.13)	.004 (.13)	.004 (.13)
February 2013	−.12 (.14)	−.17 (.15)	−.18 (.15)
Publication Day of the Week (Reference = Monday)			
Tuesday to Friday	−.01 (.07)	−.04 (.09)	−.03 (.09)
Saturday & Sunday	−.30+ (.18)	−.43* (.21)	−.44* (.21)
Generalized (Cox-Snell) R^2		.37	.38

Note. N = 5,855 for the multiple Cox regression models (Model 1 & 2). Cell entries are unstandardized Cox regression coefficients (*b*) with robust standard errors (*se*) in parentheses. [a] Log-transformed. [b] Lagged. [c] Continuous variables were mean-centered before entry (Model 2). + $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

REFERENCES

Alhabash, S., McAlister, A. R., Hagerstrom, A., Quilliam, E. T., Rifon, N. J., & Richards, J. I. (2013). Between likes and shares: Effects of emotional appeal and virality on the persuasiveness of anticyberbullying nessages on Facebook. *Cyberpsychology, Behavior, and Social Networking, 16*(3), 175-182. doi: 10.1089/cyber.2012.0265

Alliance for Audited Media. (2012). Research and data: Top 25 U.S. newspapers for September 2012. Retrieved from http://www.auditedmedia.com/news/research-and-data/top-25-us-newspapers-for-september-2012.aspx

Allison, P. D. (1978). Measures of inequality. *American Sociological Review, 43*(6), 865-880. doi: 10.2307/2094626

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: Sage.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed.). Cary, NC: SAS Institute Inc.

Allison, P. D. (2014). *Event history and survival analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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. doi: 10.1287/mnsc.1110.1421

Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011). Trends in social media: Persistence and decay. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 434-437. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2781

Atkin, C. K. (1973). Instrumental utilities and information seeking. In P. Clarke (Ed.), *New models for mass communication research* (pp. 205-242). Beverly Hills, CA: Sage.

Atkin, C. K. (1985). Informational utility and selective exposure to entertainment media. In D. Zillmann & J. Bryant (Eds.), *Selective exposure to communication* (pp. 63-91). Hillsdale, NJ: Lawrence Erlbaum.

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 65-74. doi: 10.1145/1935826.1935845

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, 519-528. doi: 10.1145/2187836.2187907

Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 26-33. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4646

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*(1), 1-26. doi: doi:10.1146/annurev.psych.52.1.1

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education and Behavior, 31*(2), 143-164. doi: 10.1177/1090198104263660

Bandura, A. (2009). Social cognitive theory of mass communication. In J. Bryant & M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (3rd ed., pp. 94-124). New York, NY: Routledge.

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment, 21*(1), 79-88. doi: 10.1037/a0014643

Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size impacts what people share. *Journal of Marketing Research*. Advance online publication. doi: 10.1509/jmr.13.0238

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science, 15*(5), 215-227. doi: 10.1287/mnsc.15.5.215

Baum, C. F., & Schaffer, M. E. (2013). *A general approach to testing for autocorrelation*. Paper presented at the Stata Conference, New Orleans, LA.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370. doi: 10.1037/1089-2680.5.4.323

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavioral Research Methods, 43*(3), 800-813. doi: 10.3758/s13428-011-0081-0

Bennett, W. L., & Iyengar, S. (2008). A new era of minimal effects? The changing foundations of political communication. *Journal of Communication, 58*(4), 707-731. doi: 10.1111/j.1460-2466.2008.00410.x

Bennett, W. L., & Iyengar, S. (2010). The shifting foundations of political communication: Responding to a defense of the media effects paradigm. *Journal of Communication, 60*(1), 35-39. doi: 10.1111/j.1460-2466.2009.01471.x

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science, 22*(7), 891-893. doi: 10.1177/0956797611413294

Berger, J. (2013). *Contagious: Why things catch on*. New York, NY: Simon & Schuster.

Berger, J., & Heath, C. (2005). Idea habitats: How the prevalence of environmental cues influences the success of ideas. *Cognitive Science, 29*(2), 195-221. doi: 10.1207/s15516709cog0000_10

Berger, J., & Iyengar, R. (2013). Communication channels and word of mouth: How the medium shapes the message. *Journal of Consumer Research, 40*(3), 567-579. doi: 10.1086/671345

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205. doi: 10.1509/jmr.10.0353

Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *Journal of Marketing Research, 48*(5), 869-880. doi: 10.1509/jmkr.48.5.869

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368. doi: 10.1093/pan/mpr057

Bigsby, E., Cappella, J. N., & Seitz, H. H. (2013). Efficiently and effectively evaluating public service announcements: Additional evidence for the utility of perceived

effectiveness. *Communication Monographs, 80*(1), 1-23. doi: 10.1080/03637751.2012.739706

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100*(5), 992-1026. doi: 10.2307/2138632

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives, 12*(3), 151-170. doi: 10.1257/jep.12.3.151

Blackmore, S. (2000). *The meme machine*. New York, NY: Oxford University Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295-298. doi: 10.1038/nature11421

Bordia, P., & DiFonzo, N. (2005). Psychological motivations in rumor spread. In G. A. Fine, V. Campion-Vincent & C. Heath (Eds.), *Rumor mills: The social impact of rumor and legend* (pp. 87-101). New Brunswick, NJ: Aldine Transaction.

Boster, F. J., Carpenter, C. J., Andrews, K. R., & Mongeau, P. A. (2012). Employing interpersonal influence to promote multivitamin use. *Health Communication, 27*(4), 399-407. doi: 10.1080/10410236.2011.595771

Boster, F. J., Kotowski, M. R., Andrews, K. R., & Serota, K. (2011). Identifying influence: Development and validation of the connectivity, persuasiveness, and maven scales. *Journal of Communication, 61*(1), 178-196. doi: 10.1111/j.1460-2466.2010.01531.x

Brosius, H.-B. (1999). The influence of exemplars on recipients' judgements: The part played by similarity between exemplar and recipient. *European Journal of Communication, 14*(2), 213-224. doi: 10.1177/0267323199014002004

Brosius, H.-B., & Bathelt, A. (1994). The utility of exemplars in persuasive communications. *Communication Research, 21*(1), 48-78. doi: 10.1177/009365094021001004

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. doi: 10.1177/1745691610393980

Cai, H., Chen, Y., & Fang, H. (2009). Observational Learning: Evidence from a Randomized Natural Field Experiment. *American Economic Review, 99*(3), 864-882. doi: 10.1257/aer.99.3.864

Campo, S., Askelson, N. M., Spies, E. L., Boxer, C., Scharp, K. M., & Losch, M. E. (2013). "Wow, that was funny": The value of exposure and humor in fostering campaign message sharing. *Social Marketing Quarterly, 19*(2), 84-96. doi: 10.1177/1524500413483456

Cappella, J. N. (2002). Cynicism and social trust in the new media environment. *Journal of Communication, 52*(1), 229-241. doi: 10.1111/j.1460-2466.2002.tb02541.x

Cappella, J. N. (2006). Integrating message effects and behavior change theories: Organizing comments and unanswered questions. *Journal of Communication, 56*(s1), S265-S278. doi: 10.1111/j.1460-2466.2006.00293.x

Cappella, J. N., & Jamieson, K. H. (1997). *Spiral of cynicism: The press and the public good*. New York, NY: Oxford University Press.

Cappella, J. N., Mittermaier, D. J., Weiner, J., Humphryes, L., & Falcone, T. (2007). *Framing genetic risk in print and broadcast news: A content analysis*. Paper presented at the annual meeting of the National Communication Association, Chicago, IL.

Carter, O. B., Donovan, R., & Jalleh, G. (2011). Using viral e-mails to distribute tobacco control advertisements: An experimental investigation. *Journal of Health Communication, 16*(7), 698-707. doi: 10.1080/10810730.2011.551998

Cha, M., Benevenuto, F., Haddadi, H., & Gummadi, K. (2012). The world of connections and information flow in Twitter. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 42*(4), 991-998. doi: 10.1109/tsmca.2012.2183359

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson & C. P. Hermann (Eds.), *Social influence: The Ontario symposium* (Vol. 5, pp. 3-39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chen, Y., Wang, Q., & Xie, J. (2011). Online social Interactions: A natural experiment on word of mouth versus observational learning. *Journal of Marketing Research, 48*(2), 238-254. doi: 10.1509/jmkr.48.2.238

Chen, Z., & Berger, J. (2013). When, why, and how controversy causes conversation. *Journal of Consumer Research, 40*(3), 580-593. doi: 10.1086/671465

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Boston, MA: Little, Brown.

Christophe, V., & Rimé, B. (1997). Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology, 27*(1), 37-54. doi: 10.1002/(sici)1099-0992(199701)27:1<37::aid-ejsp806>3.0.co;2-1

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science, 12*(4), 105-109. doi: 10.1111/1467-8721.01242

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591-621. doi: 10.1146/annurev.psych.55.090902.142015

comScore. (2012). Most read online newspapers in the world: Mail Online, New York Times and The Guardian. Retrieved from http://www.comscoredatamine.com/2012/12/most-read-online-newspapers-in-the-world-mail-online-new-york-times-and-the-guardian/

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220. doi: 10.2307/2985181

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). New York, NY: Chapman and Hall.

Cumby, R. E., & Huizinga, J. (1992). Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. *Econometrica, 60*(1), 185-195. doi: 10.2307/2951684

Dang-Xuan, L., Stieglitz, S., Wladarsch, J., & Neuberger, C. (2013). An investigation of influentials and the role of sentiment in political communication on Twitter

during election periods. *Information, Communication & Society, 16*(5), 795-825. doi: 10.1080/1369118X.2013.783608

Dawkins, R. (2006). *The selfish gene* (30th anniversary ed.). New York, NY: Oxford University Press.

De Angelis, M., Bonezzi, A., Peluso, A. M., Rucker, D. D., & Costabile, M. (2012). On braggarts and gossips: A self-enhancement account of word-of-mouth generation and transmission. *Journal of Marketing Research, 49*(4), 551-563. doi: 10.1509/jmr.11.0136

DiFonzo, N., & Bordia, P. (2007). Rumors influence: Toward a dynamic social impact theory of rumor. In A. R. Pratkanis (Ed.), *The science of social influence: Advances and future progress* (pp. 271-295). New York: Psychology Press.

Dillard, J. P., & Nabi, R. L. (2006). The persuasive influence of emotion in cancer prevention and detection messages. *Journal of Communication, 56*(s1), S123-S139. doi: 10.1111/j.1460-2466.2006.00286.x

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology, 32*(1), 271-297. doi: doi:10.1146/annurev.soc.32.061604.123127

Donsbach, W. (1991). Exposure to political content in newspapers: The impact of cognitive dissonance on readers' selectivity. *European Journal of Communication, 6*(2), 155-186. doi: 10.1177/0267323191006002003

Driscoll, J. C., & Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *The Review of Economics and Statistics, 80*(4), 549-560. doi: 10.2307/2646837

Durkin, S. J., Biener, L., & Wakefield, M. A. (2009). Effects of different types of antismoking ads on reducing disparities in smoking cessation among socioeconomic subgroups. *American Journal of Public Health, 99*(12), 2217-2223. doi: 10.2105/AJPH.2009.161638

Eastin, M. S. (2001). Credibility assessments of online health information: The effects of source expertise and knowledge of content. *Journal of Computer-Mediated Communication, 6*(4). doi: 10.1111/j.1083-6101.2001.tb00126.x

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association, 72*(359), 557-565. doi: 10.1080/01621459.1977.10480613

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K., & Lieberman, M. D. (2013). Creating buzz: The neural correlates of effective message propagation. *Psychological Science, 24*(7), 1234-1242. doi: 10.1177/0956797612474670

Falk, E. B., O'Donnell, M. B., & Lieberman, M. D. (2012). Getting the word out: Neural correlates of enthusiastic message propagation. *Frontiers in Human Neuroscience, 6*. doi: 10.3389/fnhum.2012.00313

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press.

Fisher, W. R. (1999). Narration as a human communication paradigm: The case of public moral argument. In J. L. Lucaites, C. M. Condit & S. Caudill (Eds.), *Contemporary rhetorical theory: A reader* (pp. 265-287). New York, NY: The Guilford Press.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. doi: 10.1037/h0057532

Freedman, J. L., & Sears, D. O. (1965). Selective exposure. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 57-97). New York, NY: Academic Press.

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. doi: 10.1002/asi.21641

Garrett, R. K. (2011). Troubling consequences of online political rumoring. *Human Communication Research, 37*(2), 255-274. doi: 10.1111/j.1468-2958.2010.01401.x

Gilbert, N. (2007). Computational social science: Agent-based social simulation. In D. Phan & F. Amblard (Eds.), *Agent-based modeling and simulation* (pp. 115-133). Oxford, UK: Blackwell.

Gleick, J. (2011). *The information: A history, a theory, a flood*. New York, NY: Vintage.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., . . . Verlegh, P. (2005). The firm's management of social interactions. *Marketing Letters, 16*(3), 415-428. doi: 10.1007/s11002-005-5902-4

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). The structure of online diffusion networks. *Proceedings of the 13th ACM Conference on Electronic Commerce*, 623-638. doi: 10.1145/2229012.2229058

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878-1881. doi: 10.1126/science.1202775

Goldstein, N. J., & Cialdini, R. B. (2007). Using social norms as a lever of social influence. In A. R. Pratkanis (Ed.), *The science of social influence: Advances and future progress* (pp. 167-192). New York, NY: Psycholollgy Press.

Gottschall, J. (2013). *The storytelling animal: How stories make us human*. New York, NY: Mariner Books.

Graber, D. A. (1988). *Processing the news: How people tame the information tide*. New York, NY: Longman.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420-1443. doi: 10.1086/226707

Harber, K. D., & Cohen, D. J. (2005). The emotional broadcaster theory of social sharing. *Journal of Language and Social Psychology, 24*(4), 382-400. doi: 10.1177/0261927x05281426

Harcup, T., & O'Neill, D. (2001). What is news? Galtung and Ruge revisited. *Journalism Studies, 2*(2), 261-280. doi: 10.1080/14616700118449

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555-588. doi: 10.1037/a0015701

Hartmann, T. (Ed.). (2009). *Media choice: A theoretical and empirical review*. New York, NY: Routledge.

Harvey, C. G., Stewart, D. B., & Ewing, M. T. (2011). Forward or delete: What drives peer-to-peer message propagation across social networks? *Journal of Consumer Behaviour, 10*(6), 365-372. doi: 10.1002/cb.383

Hastall, M. R. (2009). Informational utility as determinant of media choices. In T. Hartmann (Ed.), *Media choice: A theoretical and empirical review* (pp. 149-166). New York, NY: Routledge.

Hastall, M. R., & Knobloch-Westerwick, S. (2013). Severity, efficacy, and evidence type as determinants of health message exposure. *Health Communication, 28*(4), 378-388. doi: 10.1080/10410236.2012.690175

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251-1271. doi: 10.2307/1913827

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408-420. doi: 10.1080/03637750903310360

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89. doi: 10.1080/19312450709336664

Heath, C. (1996). Do people prefer to pass along good or bad news? Valence and relevance of news as predictors of transmission propensity. *Organizational Behavior and Human Decision Processes, 68*(2), 79-94. doi: 10.1006/obhd.1996.0091

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology, 81*(6), 1028-1041. doi: 10.1037/0022-3514.81.6.1028

Heath, C., & Heath, D. (2007). *Made to stick: Why some ideas survive and others die*. New York, NY: Random House.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to

articulate themselves on the Internet? *Journal of Interactive Marketing, 18*(1), 38-52. doi: 10.1002/dir.10073

Hermida, A., Fletcher, F., Korell, D., & Logan, D. (2012). Share, like, recommend: Decoding the social media news consumer. *Journalism Studies, 13*(5-6), 815-824. doi: 10.1080/1461670x.2012.664430

Ho, J. Y. C., & Dempsey, M. (2010). Viral marketing: Motivations to forward online content. *Journal of Business Research, 63*(9-10), 1000-1006. doi: 10.1016/j.jbusres.2008.08.010

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal, 7*(3), 281-312. Retrieved from http://www.stata-journal.com/article.html?article=st0128

Holbert, R. L., Garrett, R. K., & Gleason, L. S. (2010). A new era of minimal effects? A response to Bennett and Iyengar. *Journal of Communication, 60*(1), 15-34. doi: 10.1111/j.1460-2466.2009.01470.x

Hornik, R. C. (2002). Exposure: Theory and evidence about all the ways it matters. *Social Marketing Quarterly, 8*(3), 31-37. doi: 10.1080/15245000214135

Hornik, R. C., & Yanovitzky, I. (2003). Using theory to design evaluations of communication campaigns: The case of the National Youth Anti-Drug Media Campaign. *Communication Theory, 13*(2), 204-224. doi: 10.1111/j.1468-2885.2003.tb00289.x

Hu, Y., & Sundar, S. S. (2010). Effects of online health sources on credibility and behavioral intentions. *Communication Research, 37*(1), 105-132. doi: 10.1177/0093650209351512

Huang, C.-C., Lin, T.-C., & Lin, K.-J. (2009). Factors affecting pass-along email intentions (PAEIs): Integrating the social capital and social cognition theories. *Electronic Commerce Research and Applications, 8*(3), 160-169. doi: 10.1016/j.elerap.2008.11.001

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication, 59*(1), 19-39. doi: 10.1111/j.1460-2466.2008.01402.x

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the convervative media establishment*. New York, NY: Oxford University Press.

Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable media: Creating value and meaning in a networked culture*. New York, NY: New York University Press.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 154-161. doi: 10.1145/1076034.1076063

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems, 25*(2), 1-27. doi: 10.1145/1229179.1229181

Kaiser Family Foundation, & Pew Research Center. (2009). Health news coverage in the U.S. media: January 2007 – June 2008. *Pew Research Center's Journalism Project*. Retrieved from http://www.journalism.org/2008/11/24/health-news-coverage-in-the-u-s-media/

Kalyanaraman, S., & Sundar, S. S. (2006). The psychological appeal of personalized content in web portals: Does customization affect attitudes and behavior? *Journal of Communication, 56*(1), 110-132. doi: 10.1111/j.1460-2466.2006.00006.x

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. (1961). The social itinerary of technical change: Two studies on the diffusion of innovation. *Human Organization, 20*(2), 70-82.

Katz, E. (1968). On reopening the question of selectivity in exposure to mass communications. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 788-796). Chicago, IL: Rand McNally.

Katz, E. (1976). On the use of the concept of compatibility in research on the diffusion of innovation. *Proceedings of the Israel Academy of Sciences and Humanities, 5*, 126-145.

Katz, E. (1999). Theorizing diffusion: Tarde and Sorokin revisited. *ANNALS of the American Academy of Political and Social Science, 566*(1), 144-155. doi: 10.1177/000271629956600112

Katz, E. (2006). Rediscovering Gabriel Tarde. *Political Communication, 23*(3), 263-270. doi: 10.1080/10584600600808711

Katz, E., & Lazarsfeld, P. F. (2006). *Personal influence: The part played by people in the flow of mass communications*. New Brunswick, N.J.: Transaction Publishers (Original work published 1955).

Katz, E., Levin, M. L., & Hamilton, H. (1963). Traditions of research on the diffusion of innovation. *American Sociological Review, 28*(2), 237-252. doi:10.2307/2090611

Keller, E. B., & Fay, B. (2012). *The face-to-face book: Why real relationships rule in a digital marketplace*. New York, NY: Free Press.

Kim, H. S., Bigman, C. A., Leader, A. E., Lerman, C., & Cappella, J. N. (2012). Narrative health communication and behavior change: The influence of exemplars in the news on intention to quit smoking. *Journal of Communication, 62*(3), 473-492. doi: 10.1111/j.1460-2466.2012.01644.x

Kim, H. S., Lee, S., Cappella, J. N., Vera, L., & Emery, S. (2013). Content characteristics driving the diffusion of antismoking messages: Implications for cancer prevention in the emerging public communication environment. *Journal of the National Cancer Institute Monographs, 2013*(47), 182-187. doi: 10.1093/jncimonographs/lgt018

Klapper, J. T. (1960). *The effects of mass communication*. Glencoe, IL: Free Press.

Kline, R. B. (2010). *Principles and practices of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Knobloch-Westerwick, S. (2008). Informational utility. In W. Donsbach (Ed.), *International encyclopedia of communication* (pp. 2273-2276). Malden, MA: Blackwell.

Knobloch-Westerwick, S., Carpentier, F. D., & Blumhoff, A. (2005). Selective exposure effects for positive and negative news: Testing the robustness of the informational utility model. *Journalism & Mass Communication Quarterly, 82*(1), 181-195. doi: 10.1177/107769900508200112

Knobloch-Westerwick, S., Johnson, B. K., & Westerwick, A. (2013). To your health: Self-regulation of health behavior through selective exposure to online health messages. *Journal of Communication, 63*(5), 807-829. doi: 10.1111/jcom.12055

Knobloch-Westerwick, S., & Kleinman, S. B. (2012). Preelection selective exposure: Confirmation bias versus informational utility. *Communication Research, 39*(2), 170-193. doi: 10.1177/0093650211400597

Knobloch-Westerwick, S., & Sarge, M. A. (2013). Impacts of exemplification and efficacy as characteristics of an online weight-loss message on selective exposure and subsequent weight-loss behavior. *Communication Research*. Advance online publication. doi: 10.1177/0093650213478440

Knobloch-Westerwick, S., Sharma, N., Hansen, D. L., & Alter, S. (2005). Impact of popularity indications on readers' selective exposure to online news. *Journal of Broadcasting & Electronic Media, 49*(3), 296-313. doi: 10.1207/s15506878jobem4903_3

Knobloch, S. (2003). Mood adjustment via mass communication. *Journal of Communication, 53*(2), 233-250. doi: 10.1111/j.1460-2466.2003.tb02588.x

Knobloch, S., Carpentier, F. D., & Zillmann, D. (2003). Effects of salience dimensions of information utility on selective exposure to online news. *Journalism & Mass Communication Quarterly, 80*(1), 91-108. doi: 10.1177/107769900308000107

Knobloch, S., Hastall, M. R., Zillmann, D., & Callison, C. (2003). Imagery effects on the selective reading of Internet newsmagazines. *Communication Research, 30*(1), 3-29. doi: 10.1177/0093650202239023

Knobloch, S., & Zillmann, D. (2002). Mood management via the digital jukebox. *Journal of Communication, 52*(2), 351-366. doi: 10.1111/j.1460-2466.2002.tb02549.x

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Lazarsfeld, P. F., Berelson, B. R., & Gaudet, H. (1968). *The people's choice: How the voter makes up his mind in a presidential campaign* (3rd ed.). New York, NY: Columbia University Press.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York, NY: Oxford University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721-723. doi: 10.1126/science.1167742

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior, 28*(2), 331-339. doi: 10.1016/j.chb.2011.10.002

Lee, C. S., Ma, L., & Goh, D. H.-L. (2011). Why do people share news in social media? In N. Zhong, V. Callaghan, A. Ghorbani & B. Hu (Eds.), *Active media technology: Lecture notes in computer science* (Vol. 6890, pp. 129-140). Berlin, Germany: Springer-Verlag.

Lee, J. H. (2008). Effects of news deviance and personal involvement on audience story selection: A web-tracking analysis. *Journalism & Mass Communication Quarterly, 85*(1), 41-60. doi: 10.1177/107769900808500104

Lee, J. H. (2009). News values, media coverage, and audience attention: An analysis of direct and mediated causal relationships. *Journalism & Mass Communication Quarterly, 86*(1), 175-190. doi: 10.1177/107769900908600111

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 90-97. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1509

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497-506. doi: 10.1145/1557019.1557077

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. doi: 10.1177/1529100612451018

Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association, 84*(408), 1074-1078. doi: 10.1080/01621459.1989.10478874

Loewenstein, J., & Heath, C. (2009). The repetition-break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive Science, 33*(1), 1-19. doi: 10.1111/j.1551-6709.2008.01001.x

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

Macy, M. W., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology, 28*(1), 143-166. doi: doi:10.1146/annurev.soc.28.110601.141117

Magee, L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician, 44*(3), 250-253. doi: 10.1080/00031305.1990.10475731

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. doi: 10.1037/1082-989X.1.1.30

Meffert, M. F., Chung, S., Joiner, A. J., Waks, L., & Garst, J. (2006). The effects of negativity and motivated information processing during a political campaign. *Journal of Communication, 56*(1), 27-51. doi: 10.1111/j.1460-2466.2006.00003.x

Messing, S., & Westwood, S. J. (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research.* Advance online publication. doi: 10.1177/0093650212466406

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton, NJ: Princeton University Press.

Moldovan, S., Goldenberg, J., & Chattopadhyay, A. (2011). The different roles of product originality and usefulness in generating word-of-mouth. *International Journal of Research in Marketing, 28*(2), 109-119. doi: 10.1016/j.ijresmar.2010.11.003

Moriarty, C. M., & Stryker, J. E. (2008). Prevention and screening efficacy messages in newspaper accounts of cancer. *Health Education Research, 23*(3), 487-498. doi: 10.1093/her/cyl163

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science, 341*(6146), 647-651. doi: 10.1126/science.1240466

Mutz, D. C. (1998). *Impersonal influence: How perceptions of mass collectives affect political attitudes*. New York, NY: Cambridge University Press.

Myers, S., Zhu, C., & Leskovec, J. (2012). Information diffusion and external influence in networks. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 33-41. doi: 10.1145/2339530.2339540

Napoli, P. M. (2011). *Audience evolution: New technologies and the transformation of media audiences*. New York, NY: Columbia University Press.

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies, 61*(4), 631-653. doi: 10.2307/2297912

Norenzayan, A., & Atran, S. (2004). Cognitive and emotional processes in the cultural transmission of natural and nonnatural beliefs. In M. Schaller & C. S. Crandall (Eds.), *The psychological foundations of culture* (pp. 149-170). Mahwah, NJ: Lawrence Erlbaum Associates.

Norenzayan, A., Atran, S., Faulkner, J., & Schaller, M. (2006). Memory and mystery: The cultural selection of minimally counterintuitive narratives. *Cognitive Science, 30*(3), 531-553. doi: 10.1207/s15516709cog0000_68

O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory, 13*(3), 251-274. doi: 10.1111/j.1468-2885.2003.tb00292.x

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*(3), 801-823. doi: 10.1111/j.1083-6101.2007.00351.x

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419. Retrieved from http://journal.sjdm.org/

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication, 64*(2), 355-360. doi: 10.1111/jcom.12090

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC. Austin, TX: LIWC.net.

Pennebaker, J. W., Chung, C. K., Ireland, M. E., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. [LIWC manual]. Austin, TX: LIWC.net.

Pentland, A. (2014). *Social physics: How good ideas spread – the lessons from a new science*. New York, NY: The Penguin Press.

Pesaran, M. H. (2004). *General diagnostic tests for cross section dependence in panels*. Working Paper. University of Cambridge, Cambridge, United Kingdom. Retrieved from http://ssrn.com/abstract=572504

Peters, K., & Kashima, Y. (2007). From social talk to social action: Shaping the social triad with emotion sharing. *Journal of Personality and Social Psychology, 93*(5), 780-797. doi: 10.1037/0022-3514.93.5.780

Peters, K., Kashima, Y., & Clark, A. (2009). Talking about others: Emotionality and the dissemination of social information. *European Journal of Social Psychology, 39*(2), 207-222. doi: 10.1002/ejsp.523

Pew Research Center. (2010). Understanding the participatory news consumer: How internet and cell phone users have turned news into a social experience. *Pew Research Center's Internet & American Life Project*. Retrieved from http://www.pewinternet.org/Reports/2010/Online-News.aspx

Pew Research Center. (2012). In changing news landscape, even television is vulnerable - Trends in news consumption: 1991-2012. *Pew Research Center for the People & the Press*. Retrieved from http://www.people-press.org/2012/09/27/in-changing-news-landscape-even-television-is-vulnerable/

Pew Research Center. (2013a). Amid criticism, support for media's 'watchdog' role stands out. *Pew Research Center for the People & the Press*. Retrieved from

http://www.people-press.org/2013/08/08/amid-criticism-support-for-medias-watchdog-role-stands-out/

Pew Research Center. (2013b). The state of the news media 2013: An annual report on American journalism. *Pew Research Center's Project for Excellence in Journalism*. Retrieved from http://stateofthemedia.org/2013/overview-5/

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research, 44*(4), 333-348. doi: doi:10.1017/S0021849904040371

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

Rimal, R. N. (2008). Modeling the relationship between descriptive norms and behaviors: A test and extension of the theory of normative social behavior (TNSB). *Health Communication, 23*(2), 103-116. doi: 10.1080/10410230801967791

Rimé, B. (2009). Emotion elicits the social sharing of emotion: Theory and empirical review. *Emotion Review, 1*(1), 60-85. doi: 10.1177/1754073908097189

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Rosen, E. (2009). *The anatomy of buzz revisited: Real-life lessons in word-of-mouth marketing*. New York, NY: Doubleday.

Rosen, S. (1981). The economics of superstars. *American Economic Review, 71*(5), 845-858.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320. doi: 10.1207/s15327957pspr0504_2

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. doi: 10.1126/science.1121066

Salganik, M. J., & Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly, 71*(4), 338-355. doi: 10.1177/019027250807100404

Salganik, M. J., & Watts, D. J. (2009a). Social influence: The puzzling nature of success in cultural markets. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology* (pp. 315-341). New York, NY: Oxford University Press.

Salganik, M. J., & Watts, D. J. (2009b). Web-based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science, 1*(3), 439-468. doi: 10.1111/j.1756-8765.2009.01030.x

Sarafidis, V., & Wansbeek, T. (2011). Cross-sectional dependence in panel data analysis. *Econometric Reviews, 31*(5), 483-531. doi: 10.1080/07474938.2011.611458

Schaller, M., & Crandall, C. S. (Eds.). (2004). *The psychological foundations of culture*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schank, R. C., & Abelson, R. P. (1995). Knowledge and memory: The real story. In R. S. Wyer (Ed.), *Knowledge and memory: The real story. Advances in social cognition* (Vol. 8, pp. 1-85). Hillsdale, NJ: Laurence Erlbaum.

Schelling, T. C. (2006). *Micromotives and macrobehavior*. New York, NY: Norton (Original work published 1978).

Schudson, M. (1989). How culture works: Perspectives from media studies on the efficacy of symbols. *Theory and Society, 18*(2), 153-180. doi: 10.1007/BF00160753

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061-1086. doi: 10.1037/0022-3514.52.6.1061

Shibutani, T. (1966). *Improvised news: A sociological study of rumor*. Indianapolis: Bobbs-Merrill.

Shifman, L. (2012). An anatomy of a YouTube meme. *New Media & Society, 14*(2), 187-203. doi: 10.1177/1461444811412160

Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations*. New York, NY: Penguin Press.

Shoemaker, P. J. (1996). Hardwired for news: Using biological and cultural evolution to explain the surveillance function. *Journal of Communication, 46*(3), 32-47. doi: 10.1111/j.1460-2466.1996.tb01487.x

Shoemaker, P. J., Chang, T., & Brendlinger, N. (1987). Deviance as a predictor of newsworthiness: Coverage of international events in the U.S. media. In M. L. McLaughlin (Ed.), *Communication yearbook* (Vol. 10, pp. 348-365). Beverly Hills, CA: Sage.

Shoemaker, P. J., & Cohen, A. A. (2006). *News around the world: Content, practitioners, and the public*. New York, NY: Routledge.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi: 10.1037/0033-2909.86.2.420

Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion, 5*(1), 89-102. doi: 10.1037/1528-3542.5.1.89

Silvia, P. J. (2008). Interest − The curious emotion. *Current Directions in Psychological Science, 17*(1), 57-60. doi: 10.1111/j.1467-8721.2008.00548.x

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory, 17*(3), 281-303. doi: 10.1111/j.1468-2885.2007.00296.x

Smith, K. C., Niederdeppe, J., Blake, K. D., & Cappella, J. N. (2013). Advancing cancer control research in an emerging news media environment. *Journal of the National Cancer Institute Monographs, 2013*(47), 175-181. doi: 10.1093/jncimonographs/lgt023

Smith, P., Bansal-Travers, M., O'Connor, R., Brown, A., Banthin, C., Guardino-Colket, S., & Cummings, K. M. (2011). Correcting over 50 years of tobacco industry misinformation. *American Journal of Preventive Medicine, 40*(6), 690-698. doi: 10.1016/j.amepre.2011.01.020

Southwell, B. G. (2013). *Social networks and popular understanding of science and health: Sharing disparities*. Baltimore, MD: Johns Hopkins University Press.

Southwell, B. G., & Yzer, M. C. (2007). The roles of interpersonal communication in mass media campaigns. In C. S. Beck (Ed.), *Communication yearbook* (Vol. 31, pp. 420-462). New York, NY: Lawrence Erlbaum Associates.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Malden, MA: Blackwell.

Stephens, M. (2007). *A history of news* (3rd ed.). New York, NY: Oxford University Press.

Strizhakova, Y., & Krcmar, M. (2007). Mood management and video rental choices. *Media Psychology, 10*(1), 91-112. doi: 10.1080/15213260701301152

Stroud, N. J. (2011). *Niche news: The politics of news choice*. New York, NY: Oxford University Press.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73-100). Cambridge, MA: The MIT Press.

Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication, 51*(1), 52-72. doi: 10.1111/j.1460-2466.2001.tb02872.x

Sundaram, D. S., Mitra, K., & Webster, C. (1998). Word-of-mouth communications: A motivational analysis. *Advances in Consumer Research, 25*(1), 527-531.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Sunstein, C. R. (2009). *On rumors: How falsehoods spread, why we believe them, what can be done*. New York, NY: Farrar, Straus and Giroux.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM, 53*(8), 80-88. doi: 10.1145/1787234.1787254

Tarde, G. (1903). *The laws of imitation* (E. C. Parsons, Trans.). New York: Henry Holt.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. doi: 10.1177/0261927x09351676

Tewksbury, D., & Rittenberg, J. (2012). *News on the Internet: Information and citizenship in the 21st century*. New York, NY: Oxford University Press.

Thorson, E. A. (2008). Changing patterns of news consumption and participation: News recommendation engines. *Information, Communication & Society, 11*(4), 473 - 489. doi: 10.1080/13691180801999027

Thorson, E. A. (2013). *Belief echoes: The persistent effects of corrected misinformation* (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia, PA.

Thurman, N., & Schifferes, S. (2012). The future of personalization at news websites. *Journalism Studies, 13*(5-6), 775-790. doi: 10.1080/1461670x.2012.664341

Turner-McGrievy, G., Kalyanaraman, S., & Campbell, M. K. (2013). Delivering health information via podcast or web: Media effects on psychosocial and physiological responses. *Health Communication, 28*(2), 101-109. doi: 10.1080/10410236.2011.651709

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458. doi: 10.1126/science.7455683

van den Hooff, B., Schouten, A. P., & Simonovski, S. (2012). What one feels and what one knows: The influence of emotions on attitudes and intentions towards knowledge sharing. *Journal of Knowledge Management, 16*(1), 148-158. doi: 10.1108/13673271211198990

van Dijk, T. A. (1988). *News as discourse*. Hillsdale, NJ: Lawrence Erlbaum.

van Noort, G., Antheunis, M. L., & van Reijmersdal, E. A. (2012). Social connections and the persuasiveness of viral campaigns in social network sites: Persuasive intent as the underlying mechanism. *Journal of Marketing Communications, 18*(1), 39-53. doi: 10.1080/13527266.2011.620764

Van den Bulte, C., & Lilien, Gary L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology, 106*(5), 1409-1435. doi: 10.1086/320819

Vicsek, T. (2002). Complexity: The bigger picture. *Nature, 418*(6894), 131. doi: 10.1038/418131a

Walther, J. B., & Jang, J.-w. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication, 18*(1), 2-15. doi: 10.1111/j.1083-6101.2012.01592.x

Wang, C., & Huberman, B. A. (2012). Long trend dynamics in social media. *EPJ Data Science, 1*(1), 1-8. doi: 10.1140/epjds2

Watts, D. J. (2007). The collective dynamics of beliefs. In V. Nee & R. Swedberg (Eds.), *On capitalism* (pp. 241-272). Stanford, CA: Stanford University Press.

Watts, D. J. (2011). *Everything is obvious once you know the answer*. New York, NY: Crown Business.

Weimann, G. (1994). *The influentials: People who influence people*. Albany, NY: Albany, NY.

Westerwick, A., Kleinman, S. B., & Knobloch-Westerwick, S. (2013). Turn a blind eye if you care: Impacts of attitude consistency, importance, and credibility on seeking of political information and implications for attitudes. *Journal of Communication, 63*(3), 432-453. doi: 10.1111/jcom.12028

Williams, B. A., & Delli Carpini, M. X. (2011). *After broadcast news: Media regimes, democracy, and the new information environment*. New York, NY: Cambridge University Press.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York, NY: McGraw-Hill.

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education and Behavior, 27*(5), 591-615. doi: 10.1177/109019810002700506

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Mason, OH: South-Western Cengage Learning.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences, 104*(45), 17599-17601. doi: 10.1073/pnas.0704916104

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*, 705-714. doi: 10.1145/1963405.1963504

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 177-186. doi: 10.1145/1935826.1935863

Zhang, J. (2010). The sound of silence: Observational learning in the U.S. kidney market. *Marketing Science, 29*(2), 315-335. doi: 10.1287/mksc.1090.0500

Zillmann, D. (1988). Mood management through communication choices. *American Behavioral Scientist, 31*(3), 327-340. doi: 10.1177/000276488031003005

Zillmann, D. (2000). Mood management in the context of selective exposure theory. In M. E. Roloff (Ed.), *Communication yearbook* (Vol. 23, pp. 103-123). Thousand Oaks , CA: Sage.

Zillmann, D. (2006). Exemplification effects in the promotion of safety and health. *Journal of Communication, 56*(s1), S221-S237. doi: 10.1111/j.1460-2466.2006.00291.x

Zillmann, D., & Brosius, H.-B. (2000). *Exemplification in communication: The influence of case reports on the perception of issues*. Mahwah, NJ: Lawrence Erlbaum.

Zillmann, D., & Bryant, J. (1985). Affect, mood, and emotion as determinants of selective exposure. In D. Zillmann & J. Bryant (Eds.), *Selective exposure to communication* (pp. 157-190). Hillsdale, NJ: Lawrence Erlbaum.

Zillmann, D., Chen, L., Knobloch, S., & Callison, C. (2004). Effects of lead framing on selective exposure to Internet news reports. *Communication Research, 31*(1), 58-81. doi: 10.1177/0093650203260201

Zillmann, D., Knobloch, S., & Yu, H.-S. (2001). Effects of photographs on the selective reading of news reports. *Media Psychology, 3*(4), 301-324. doi: 10.1207/S1532785XMEP0304_01