8-2020

Diffusion of Falsehoods on Social Media

Kelvin Kizito King
The University of Texas Rio Grande Valley

DIFFUSION OF FALSEHOODS ON SOCIAL MEDIA

A Dissertation

by

Kelvin Kizito King

Submitted to the Graduate College of
The University of Texas Rio Grande Valley
In partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2020

Major Subject: Information Systems

DIFFUSION OF FALSEHOODS ON SOCIAL MEDIA

A Dissertation
by
KELVIN KIZITO KING

COMMITTEE MEMBERS

Dr. Bin Wang
Chair of Committee

Dr. Diego Escobari
Committee Member

Dr. Jun Sun
Committee Member

Dr. Tamer Oraby
Committee Member

August 2020

ABSTRACT

King, Kelvin K., The Diffusion of Falsehoods on Social Media. Doctor of Philosophy(Ph.D.), August 2020, 132 pp., 17 tables, 13 figures, references, 119 titles.

Misinformation has captured the interest of academia in recent years with several studies looking at the topic broadly. However, these studies mostly focused on rumors which are social in nature and can be either classified as false or real. In this research, we attempt to bridge the gap in the literature by examining the impacts of user characteristics and feature contents on the diffusion of (mis)information using verified true and false information. We apply a topic allocation model augmented by both supervised and unsupervised machine learning algorithms to identify tweets on novel topics. We find that retweet count is higher for fake news, novel tweets, and tweets with negative sentiment and lower lexical structure. In addition, our results show that the impacts of sentiment are opposite for fake news versus real news. We also find that tweets on the environment have a lower retweet count than the baseline religious news and real social news tweets are shared more often than fake social news. Furthermore, our studies show the counter intuitive nature of current correction endeavors by FEMA and other fact checking organizations in combating falsehoods. Specifically, we show that even though fake news causes an increase in correction messages, they influenced the propagation of falsehoods. Finally our empirical results reveal that correction messages, positive tweets and emotionally charged tweets morph faster. Furthermore we show that tweets with positive sentiment or those emotionally charged morph faster over time. Word count and past morphing history also positively affect morphing behavior.

.

DEDICATION

The completion of my doctoral studies would not have been possible without the love and support of my family. My wife Lucy and my kids, Dora and Emmanuelle, wholeheartedly inspired, motivated and supported me to accomplish this degree. Thank you for your love and patience. To God, “the author and finisher of our faith”. ~ Hebrews 12:2

AKNOWLEDGEMENTS

I will always be grateful to Dr. Bin Wang, my advisor and chair of my dissertation committee for her continuous support throughout my Ph.D. Studies. Furthermore, her patience, motivation, and immense knowledge and support enabled me to build up a comprehension within the subject. Her guidance was instrumental in the crafting, writing and in the successful completion of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. research.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Diego Escobari, for his creative and overflowing stream of ideas, Dr. Jun Sun for his mentorship throughout my graduate program, and Dr. Tamer Oraby for his patience in guiding me through complex possibilities. I thank all of them for their insightful comments, encouragement and feedback.

I will also like to thank the INFS department chair, Dr. Jerald Hughes for his enthusiasm and encouragement for this project, and my 2017 INFS cohort, especially James Wairimu for being a wonderful colleague.

Finally, many thanks to Nora Ramirez and Shay Nelson of the INFS department, where would we be without you!

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The number of online social media users have skyrocketed in recent times and with it comes the reliance on social media platforms by a large majority of the population for both news and information (Reuters 2017). Unlike other media platforms, there is no easy way to filter information on social media due to the high variety, velocity and volume of the data that flows through the networks (Bello-Orgaz et al. 2016; Tear and Southall 2019; Tsou 2015). Veracity, a major characteristic of big data has most likely become one of the most challenging issues with respect to filtering contents in online social networks. Therefore it comes as no surprise that false information has gained root and can now reach the masses quite ostensibly since the Brexit campaign and the US elections, both in 2016 (Barthel et al. 2016).

Borrowing from previous literature, we define fake news as “news articles that are intentionally and verifiably false, and could mislead readers” (Allcott and Gentzkow 2017). The motivations for the creation of falsehoods range from monetary to ideological (Allcott and Gentzkow 2017). Irrespective of the motivation, it is agreed that fake news has proven to be a challenge well beyond the capacity and scope of one field of research. With the growing influence of fake news and the consequences that it brings to society, it is imperative that we combat this phenomenon.

Several studies in social science (Vosoughi et al. 2018), computer science (Pérez-Rosas et al. 2017) and information systems (IS) (Osatuyi and Hughes 2018) have attempted to look at this phenomenon through several lenses. Some have looked at the it from a proactive standpoint using detections (Karimi et al. 2018; King and Sun 2019; Long et al. 2016), while others have looked at it from a strictly behavioral standpoint (Kim and Dennis 2017). Despite the abundance of studies on the topic, an efficient approach still eludes researchers. For example, proactive methodologies such as fake news detection have produced abysmal results, with several techniques producing subpar performances from their classifiers (Karimi et al. 2018; Wang 2017). Another approach is to tackle the problem from the behavioral perspective. More specifically, to prevent the dissemination of falsehoods through slowing the spread. One of several ways of addressing this concept is to find out why people share false contents. Studies have shown that false news gains virality by reaching a wider audience when it is shared from person to person on social media (Allcott and Gentzkow 2017; Friggeri et al. 2014; Vosoughi et al. 2018). Prior studies on the sharing behavior have looked at it from the point of rumors (Dunn and Allen 2005; Friggeri et al. 2014; Oh et al. 2010; Vosoughi et al. 2018). However, no study to date has looked at the concept using verified fake and real news. Rumors, however, differ from fake news in several ways. Most notably, rumors are social in nature and according to previous studies are "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that functions to help people make sense and manage risk" (DiFonzo and Bordia 2007). This implies that rumors can be either true or false and is most prevalent during crisis events. Moreover, most of the studies have been inconsistent and contradictory. For example, a recent study showed that malicious bots were very instrumental in the sharing behavior of rumors (Shao et al. 2018), while another study contradicted that notion by showing that the presence of bots had no influence in the general spread of rumors (Vosoughi et al. 2018). Prior studies have also laid claim to the fact that positive tweets were more likely to be shared as compared to negative ones (Berger and Milkman 2012). However, another study contradicted the claim and showed that negative tweets influenced sharing of information more (Bene 2017). Therefore, given the growth of misinformation and the conflicting theoretical viewpoints concerning the features affecting its dissemination, investigating what features were more important in the sharing of fake versus real news has never been more pertinent. This study not only bridges the gap in the literature and provides implications for practice, but also reconciles difference results from previous studies on the topic.

Several studies have also investigated the concept of false news using reactive measures such as source credibility (Oh et al. 2010), news presentation and source rating (Kim and Dennis 2017). One of such proposed methods is the use of corrections and rebuttals to fight misinformation. This method has been used quite successfully since the early ages to suppress insurgencies and as propaganda tools by repressive regimes (Bauer and Gleicher 1953). In recent years however, government agencies and policy makers have embraced such tools in the fight to dispel rumors and misinformation. For example, quite recently the Department of Homeland Security (DHS) established a 22-member coalition - "The Social Media Working Group for Emergency Services and Disaster Management (SMWGESDM)" - to help provide recommendations for curbing the spread of misinformation (*SMWG* 2018). The agency's primary tool for combating falsehood is targeting such contents with rebuttals and corrections. However, according to the SMWG's 2018 report, the use of rebuttals and corrections on falsehoods under certain conditions may not be effective in combating the spread of misinformation. Given the importance of this topic and the abundance of research on rebuttals (Huang 2017; Lewandowsky et al. 2012; Ozturk et al. 2015), it is quite surprising that no study has looked at the effectiveness of correction messages on falsehoods and of its causal inference. It is therefore imperative that we investigate the efficacy of correction messages on falsehoods and develop a method to measure the causal relationships between them. It is also equally important to understand how such a relationship between both falsehoods and correction plays out in the grand scheme of things in order to better improve our mitigation strategies. This understanding will bridge the gap in the literature and provide a framework for policy makers and government agencies like FEMA and DHS in the redesigning and updating their mitigation strategies. The knowledge gained may also guide administrators in the development of effective warning systems during crisis events.

Considering that the importance of the efficacy of correction messages on falsehoods are of utmost importance and that both are dependent on the messages being shared in the network, it is imperative to understand the sharing behavior of both falsehoods and correction messages at a more granular level. However, studies on the topic, though general, have revealed that people are sometimes influenced by their emotional inclinations during information sharing (Berger and Milkman 2010; Harris and Paradice 2007; Jonah Berger 2011). Studies have also noted the sharing of news and information on social networks usually includes the modification of the textual content of the original messages in order to personalize the message (Boyd et al. 2010). Though studies have in retrospect looked at how memes may evolve and come back several times (Vosoughi et al. 2018) and how a message may evolve as whole (Shin et al. 2018), no study has investigated how the textual contents of both falsehoods and correction messages change during the diffusion process. More importantly, no study has also looked at the role user sentiments play in the sharing process, specifically the morphing of the textual contents in the Twittersphere at the granular level.

To investigate these questions and bridge the gap in literature, we propose three studies that combine knowledge from the fields such as mathematics, econometrics, computer science, social science and information systems. We leverage on the use of unique Twitter datasets extracted during multiple shock events for our analysis as studies have showed that falsehoods are prevalent and thrive the most during the three “Cs”: conflict events, crisis scenarios and catastrophes (Koenig 1985).

Specifically, this study focuses on the following research questions:

- What factors affect the retweetability of false versus real news?
- Are the impacts of sentiments, novelty and lexical density on retweetability different for false versus real news?
- How does the retweet count of tweets in different news categories differ for false versus real news?
- Is there a bidirectional relationship between the diffusion of falsehoods and correction messages on Twitter during shock events?
- What are the effects of correction messages on falsehoods and vice versa?
- How do false, real and correction messages morph on social media?
- What are the effects of sentiments on the morphing of false news and correction messages?

Abstracts of the three essays

Essay#1: Diffusion of False versus Real news on Social Media

Misinformation has captured the interest of academia in recent years with several studies looking at the topic broadly. However, these studies mostly focused on rumors which are social in nature and can be either classified as false or real. In this research, we attempt to bridge the gap in the literature by examining the impacts of user characteristics and feature contents on the diffusion of (mis)information using verified true and false information. We apply a topic allocation model augmented by both supervised and unsupervised machine learning algorithms to identify tweets on novel topics. We find that retweet count is higher for fake news, novel tweets, and tweets with negative sentiment and lower lexical structure. In addition, our results show that the impacts of sentiment are opposite for fake news versus real news. We also find that tweets on the environment have a lower retweet count than the baseline religious news and real social news tweets are shared more often than fake social news.

Essay#2: Dynamic Effects of Correcting Falsehood on Social Media

Online social networks make it incredibly easy for the propagation of misinformation, due to a large majority of the public relying on microblogging sites such as Twitter for news and information. To mitigate the diffusion of falsehoods on social media, government agencies and fact-checking websites have issued corrective messages. However, there has been no research on the impacts of correction messages and falsehoods on each other’s diffusion and whether the correction messages are effective in combating falsehood. Drawing on the competitive exclusion principle, we first develop a theoretical model on the diffusion of both falsehoods and correction messages analogous to two species competing for the limited resources in the ecosystem. Our

results show that increasing the replication rates for correction messages suppresses and causes falsehoods to eventually die out while enabling correction messages to thrive. Next, leveraging on panel vector autoregressive models and machine learning techniques, we employ unique panel datasets containing information on 279,597 social media interactions during shock events to empirically investigate the dynamic effects of falsehoods and correction messages sent from the United States' Federal Emergency Management Agency (FEMA) and several fact-checking organizations on each other. Contrary to popular perceptions, we find that correction messages cause an increase in the propagation of falsehoods on social media. The results also show that falsehoods would eventually die out without an introduction of correction messages but increase when there is an increase in correction messages. This study has important implications for both theory and practice.

Essay#3: Effects of Sentiment on the Morphing Behavior of Falsehoods and Correction Messages.

Fake news has become a thorn in the sides of researchers and industry practitioners alike. Studies show that one of the major factors that drive its diffusion on social media is the emotional context of the message. However, studies have largely missed a major component in the diffusion process, which is the morphing of the textual contents themselves during that process. Leveraging on cosine similarity and econometric modeling, we investigate the role sentiment plays in this morphing process. We find that positive sentiments, emotionally charged messages and correction messages positively affect the morphing of messages during the diffusion process. We also find out that as time goes by emotionally charged messages and sentiments influence the morphing of the messages. Our study shows us that while falsehoods morphs aggressively at the initial stages, correction messages were more aggressive in the long run. Finally, falsehoods lasted longer than correction messages.

CHAPTER II

DIFFUSION OF FALSE VERSUS REAL NEWS ON SOCIAL MEDIA

Introduction

Online social networks (OSN) have become an efficient way for information dissemination (Bowler et al. 2009). For example, statistics show that in 2017, 67% of US adults depended on online social network platforms such as Snapchat, Facebook and Twitter for news as compared with 62% in 2016 (Reuters 2017). Twitter has close to 9 out of 10 of its users using the social media outlet primarily for news, and of those that do so, 74% use the platform daily (Rosenstiel et al. 2015). As a result, Twitter has come to replace mainstream media as the number one choice for news especially among the younger generation, and its reach is expected to grow even further (Stieglitz and Dang-Xuan 2013). Despite its popularity, OSN as an information diffusion channel has an inherent disadvantage. Because users of these platforms act as gatekeepers and are prone to their own individual biases, it is difficult to assess the veracity of news items being propagated. This issue has been exasperated even more by the sheer amount of data that flows through online social networks. This has led to both true and false news items on social media, reaching a larger audience than news from major mainstream news outlets (Allcott and Gentzkow 2017). Hence, it is not surprising that headlines of falsehoods deceive American adults about seventy-five percent of the time and the most popular false news stories usually garnered a lot more shares than real news (Silverman and Singer-Vine 2016). Extant research has identified salient features that may influence the retweet or diffusion of information on Twitter (Boyd et al. 2010; Lee et al. 2014) These features may be categorized into three broad categories: a) user-based, which are features directly linked to the behavior and characteristics of the users b) time-based, which is based on the time the tweet is generated and posted, and c) content-based, which is based on the content of the text embedded in the message (Hoang and Mothe 2018). Various studies have also examined the diffusion of related phenomena such as rumors, misinformation and disinformation. For example, some studies focused on the diffusion of rumors, which are unverified news items that are inherently social and can be either true or false (Astapova 2017; Cheng et al. 2013; Ma et al. 2016). Other studies have looked at this phenomenon based on misinformation, which is news that contains inaccurate information designed to deliberately deceive but shared unknowingly (Allcott et al. 2018; Chen 2016). This differs from disinformation, which is news that is inaccurate information with the intention to mislead. Propaganda is a prime example. In the current research, we focus on the diffusion of verifiable false (fake) and real news.

Several studies rooted in social science, computer science and business have utilized various methodologies including experiments (e.g., Osatuyi and Hughes 2018), cascades (e.g., Friggeri et al. 2014; Vosoughi et al. 2018) and network analysis (e.g., Agrawal et al. 2013; Kim et al. 2018) to determine the characteristics, factors and features that affect the diffusion of misinformation. Though behavioral characteristics can be inferred from textual contents, few studies have attempted to use a combination of techniques to verify the factors responsible for the diffusion. Given the complexity of analyzing Twitter data and the unavailability of complete datasets due to Twitter’s rate limits (Twitter 2019), large scale studies using both econometric and algorithmic techniques have been rare. Furthermore, results from a handful of analyses have been inconsistent. For example, a study using textual contents from Twitter showed that both positively and negatively charged tweets were retweeted more often and quicker than neutral ones (Stieglitz and Dang-Xuan 2013). The researchers concluded that sentiments inferred from social media contents might be positively associated with information diffusion. Other studies have related the propagation of falsehoods to automated entities and claimed that these entities actively spread falsehoods. They show that in the earlier phases, the automated entities target mostly influential users on social media that eventually lead to more diffusion of falsehoods (Shao et al. 2018). This finding was contradicted by another study which revealed that contrary to conventional wisdom, malicious entities were not any more responsible for the propagation of falsehoods than humans were (Vosoughi et al. 2018). Rather, the study claimed that human behavior contributed more to the differential propagation of falsehoods than automated entities did.

The current study employs a unique dataset of tweets across a 5-week period during Hurricane Harvey. We empirically investigate the retweet counts based on user- and content-based features to determine what factors affect the retweetability of false versus real news. Our empirical results show that the retweet count was higher for false news, novel tweets, and tweets with negative sentiments and lower lexical density. We show that the impacts of sentiment were different for fake news than real news. In addition, tweets on the environment had lower retweet counts compared with the baseline religious tweets, and the retweet count of social tweets were higher for real news than false news. Despite the burgeoning literature on fake news diffusion, this is the first time a study addresses the lexical component of a retweet. Our results show that when a tweet contains more lexical words, there is a decrease in sharing. These findings have implications for research and practice and provide guidelines for administrators in online social networks.

Background literature

Our study draws upon two major streams of research and encompasses areas in behavioral and linguistic analysis. They are false news and text mining and language use. The following sections in our study explore each stream as they are applied to our research.

False News

Online social media outlets lack the regular news media's editorial standards and procedures for ensuring the veracity of information (Lazer et al. 2018). This has given rise to the increase in false news stories circulating in our cyberspace. However, the extant literature has been mostly focused on rumors (Agrawal et al. 2013; Cheng et al. 2013; Friggeri et al. 2014; Ma et al. 2016; Oh et al. 2010; Shin et al. 2018). For example, two studies (Dunn and Allen 2005; Friggeri et al. 2014) have revealed that rumors run deeper in social networks and cause the spread of misinformation in the absence of verifiable information. Other researchers have studied the phenomenon from a purely empirical perspective with contrasting findings (Chen 2016; Oh et al. 2010) with one study revealing the action of sharing, rather than the perceived accuracy of the message and the characteristics of the information being shared, affect diffusion (Chen 2016) and another finding that the credibility of the sources was what mattered to most users, suppressed anxiety from the community and led to rumor control (Oh et al. 2010). In addition, information without notable sources was the largest culprit in rumor propagation (Agrawal et al. 2013). Other studies have been mainly interested in textual content-related features that may cause the spread of misinformation like novelty and sentiments. One study revealed that negative sentiments were often favored by false news to attract sharing (Osatuyi and Hughes 2018), while another showed that false news with both negative and positive sentiments were shared equally (Stieglitz and Dang-Xuan 2013). Using cascades, Vosoughi et al. (2018) explained that false news may be more novel, and this may be the reason it is more likely to be shared. However, the authors stopped short of conducting any analysis as to the impact of novelty on false news sharing, We summarize the relevant literature in Table 1.

Table 1. Summary of the false news literature.

Topic	**Theory**	**Data & Methods**	**Results**
Fake vs. real news (Osatuyi and Hughes 2018)	Elaboration likelihood model	Semantic analysis, secondary data and Internet platform using Fortune 500 business websites	Fake news provides less information than real news.
	Epistemology of testimony	Empirical analysis of survey using social network sites.	Social tie variety and cognitive homogeneity are essential predictors of truth in analyzing and creating fake news awareness.
News and Rumors (Jin et al. 2013)	Epidemiological model	SEIZ model using Twitter	It involved capturing, characterizing and distinguishing the diffusion of rumor

			topics from those that are news.
Extreme events (rumors) (Oh et al. 2010)	Rumor theory	Network analysis, position, semantics, emotional words (not on propagation, managing anxiety) using Twitter	Information that depicted credible sources reduced anxiety and encouraged the sharing of credible information.
Extreme events and rumors (Agrawal et al. 2013)	Rumor theory	Analyses of social ties on Twitter during the Mumbai terrorist attack, the Seattle cafe shooting incident and Toyota recalls in 2008, 2012 and 2010 respectively using logistic regression and content analysis.	Information without definitive sources was the most important rumor causing factor on Twitter with respect to social crises, followed by personal involvement next and anxiety the least important.
Extreme events (Kim et al. 2018)	Social capital	Social network analysis (community structure and	Certain types of Twitter users who were news & weather

		centrality) and text analysis using Twitter data during Storm Cindy.	agencies were more pronounced as information sources and diffusers.
Information diffusion (Stieglitz and Dang-Xuan 2013)	Sentiment analysis	Sentiment analysis on tweets and information sharing.	Emotionally charged political tweets were shared quicker and more often than neutral ones.
Fake news (Allcott and Gentzkow 2017)	The economics of fake and media markets	Mixed approach using both survey and secondary data on Facebook.	The average adult in the U.S. reads and remembers several falsehoods that were shared during election periods. Pro-Trump articles had a higher exposure rate compared to pro-Clinton ones.
Rumors (Ma et al. 2016)	Neural networks	Detection and debunking of rumor on microblogs using Twitter data.	Recurrent neural network (RNN)-based models perform significantly

			better than state-of-the-art technology in detecting rumors.
Rumor diffusion (Cheng et al. 2013)	Epidemic model	Secondary data and social networks using social blog sites and catalogs.	Selecting weak ties cannot improve the dissemination of rumors but the propagation is improved after deleting some of them. The strength of tie influences the propagation of rumors on social networks.
Rumor and cascades (Friggeri et al. 2014)	Cascades	Secondary data using network analysis on Facebook’s information propagation.	Information can be easily and quickly transmitted via social ties, even when it is not verifiable. True rumors were

			most viral and elicited the largest cascades.
Knowledge and leadership in online communities (Faraj et al. 2015)	Social capital and social networks	Survey and content analysis on Usenet newsgroup online communities.	Sociable behavior does not correlate with being identified as a leader in online communities. The likelihood of a central participant being identified as a leader is higher if they exhibit knowledge contribution and sociable behaviors.
News stories (Kim and Dennis 2017)	Literature on news stories, storyteller, source primacy and information processing	Primary data and survey on Facebook users using multilevel mixed-effects linear regression.	Presenting OSN stories in a story format with source ratings causes users to evaluate their truthfulness more critically and

			affects its believability.
Information propagation (Hoang and Mothe 2018)	Information diffusion	Predictive modeling on tweets during Hurricane Sandy.	Examined user-based, time-based and content-based variables and found that the time a user posts is strongly correlated with retweetability and thus virality.
Rumor and misinformation (Shin et al. 2018)	Information diffusion	Text analysis of temporal pattern, content mutation and misinformation on Twitter.	Unlike real news, misinformation tends to come back multiple times after the initial publication. True rumors however do not.

Next, we briefly review recent studies on language use and different methods for mining textual contents on social media.

Text Mining and Language Use

Several text mining techniques have in recent years been used to analyze large volumes of unstructured or semi-structured data with the sole purpose of discovering useful patterns necessary for decision making (Abbasi et al. 2018; Huei Chou et al. 2010). For example, clustering analysis (Aldenderfer and Blashfield 1984) is a widely used unsupervised technique for exploring and categorizing both structured and unstructured data. It has been applied in fields such as marketing (Punj and Steward 1983), information technology (Wei et al. 2009) and management (Ketchen Jr. and Shook 1996). However, clustering also has the unintended consequence of generating clusters even when the congenital structure of the data is questionable (Balijepally et al. 2011).

Another technique used in text mining is Latent Dirichlet allocation (LDA), which is a statistical model for topic modeling (Blei 2003; Shu et al. 2009). This generative technique allows researchers to identify underlying topics in text documents and their similarity or differences within or between documents. Unlike cluster analysis, LDA does not measure distance nor use clustering. LDA is essentially a Bayesian model with three levels of hierarchy, and every element in a corpus can be modeled as a limited combination on a primordial set of topics (Blei 2003). The probability of each topic is a definite delineation of the corpora. Simply put, the documents in a corpus constitute an arbitrary synthesis over latent topics, and each of those topics are represented using a distribution over words (Abramowitz and Stegun 1964). When used in tandem with other supervised and unsupervised learning techniques such as cosine similarity and cluster analysis, LDA can be very accurate and fast for the processing of large, irregular sized or unstructured data (Sun et al. 2008). An example is when using the cosine and term frequency–inverse document frequency solution as an initialization point for LDA.

The extant literature on the analysis of textual contents from microblogging sites typically rely on clustering to sort out records using date and time attributes, which often works well on more structured and smaller datasets. Clustering techniques are not very efficient with larger datasets (Balijepally et al. 2011). Hierarchical methods are also not amenable for large datasets and are more vulnerable to outliers (Hair et al. 2016). A solution would be to use random subsamples which may not be very useful in our study. In this research, we employ LDA algorithms to determine the novelty of a tweet relative to other tweets on the same topics. This approach differs from prior manual coding approaches.

Of all text analytics methods and techniques, lexical density is one of the least used in academia. Lexical density originated from a branch of computational linguistics (statistical modeling) and refers to the ratio of lexical over functional words in a text corpus (Halliday 1989; Ure 1971). This relates to the generality of "lexico-grammar" attributed to the hierarchy of words used in a language (Halliday 1989). One of the uses of lexical density is to measure the readability and complexity of language as they pertain to literacy (To et al. 2013). Ure (1971) proposed that lexical words should be measured as all content words in the corpus divided by the total number of all words. The measurement was later refined by Halliday (1989), the originator of Systemic Functional Linguistics, to be the number of lexical items as a "proportion of the number of clauses." To measure lexical density, it is imperative to first distinguish functional words from their lexical counterparts and their differences therein. Lexical items comprise of four word classes which are traditionally made up of nouns, verbs, adjectives and adverbs. Function words comprise of determiners, some of which include pronouns, most prepositions, various auxiliary verbs, conjunctions, some sets of adverbs (Halliday 1989). Lexical density has been used in spoken words and compared with textual contents. Results comparing both over the years show that, overall, textual contents are much denser lexically compared to speech (Nesi 2001). A recent study showed that there is a relationship between the lexical density of a text and it’s readability (To et al. 2013). Given the character limit on each tweet, lexical density may be an important variable that affects readability and the desire to share tweets.

Research Hypotheses

As previously discussed, rumors have social connotation and can either be true of false. However, their impact is usually felt during uncertain times and in some cases during the deliberate dissemination of false information (DiFonzo and Bordia 2007). Such misinformation or disinformation can then be carried by rumors at an alarming rate within the network (Dunn and Allen 2005). Studies show that when users are exposed to an additional TV campaign ad, there is a change in the number of shared votes by 0.02 percentage points (Spenkuch and Toniatti 2015). This implies that exposure to false news as a form of persuasive TV campaign ad may influence individuals’ perceptions on its truthfulness and may cause sharing. Several studies on rumors show that false news are shared more than real news (Friggeri et al. 2014; Vosoughi et al. 2018). Therefore, we hypothesize:

H1: *An original tweet’s veracity is negatively associated with its retweet count.*

Several definitions of novelty exist. The most common one for this multidimensional construct is the condition of an item being new or unusual (Lee and Crompton 1992). Novelty is the exact opposite of familiarity and of which attention is fueled. Studies using information theory and Bayesian decision theory have showed that novelty attracts human attention and encourages information sharing (Berger and Milkman 2012; Itti and Baldi 2009). This might be because in a group, people with novel news are usually seen as important, therefore the sharing of such news increases their importance within the group. Hence, people who have information that they believe may be novel to others may seek to share that information not for altruistic purposes but to improve their worth in the minds of peers. An important aspect of novel news as stated earlier is that it updates humans on their environment and increases knowledge which in turn improves their ability to make informed decision. This confers a social status on the bearer of such novel information which makes them more inclined to share. Novelty through the stimulus of cues of a message in some cases may be able to trigger surprises. Using Bayesian theory, the idea of novelty had been measured using the differences between people's priors experiences in relation to their posterior's, and studies on human behavior show how information effects of an observer's beliefs may yield increasing terms of surprise and awe if the information is seen as novel or rare (Itti and Baldi 2009). This may be because of the tendencies of novel news to be more valuable and awe inspiring, and as such it inspires attention and ultimately sharing. The sharing of novel information may give a sense of importance to the diffuser such that people may approach or be inclined to want information that fewer people are privy to, thus putting them in an advantageous position (Aral and Dhillon 2016). The rush for such information will eventually lead to an increase in the desire to share. Therefore, we hypothesize:

H2: *An original tweet's novelty is positively associated with its retweet count.*

Positive news and contents are known to trigger emotions and affective components in individuals. Studies on mood regulation suggests that people's decisions are consistently geared towards preserving positivity in their everyday lives (Di Muro and Murray 2012). In addition, emotionally charged tweets, positive or negative, have been known to be disseminated more and in a quicker manner than neutral ones (Stieglitz and Dang-Xuan 2013). News stories do not have to be negative to have an effect, especially during extreme weather conditions where positive news may spread far (Bene 2017; Berger and Milkman 2013). Hence, even though both positive and negative sentiments may be prominent, we expect positive tweets to hold more sway as compared to negative tweets. This in turn may encourage the retweeting of positive news as compared to both negative and neutral tweets. We therefore hypothesize:

H3: *An original tweet's sentiment is positively associated with its retweet count.*

Microblogs such as Twitter designed originally for mobile use was not designed to support conversations and can be highly incoherent with regards to syntax and character limitations (Honeycutt and Herring 2009). Furthermore, a surprising degree of coherence is facilitated especially using taglines such as "@," which may act as a marker of an adjective and brings to bare the sheer limitations of the lexical nature of Twitter's design. Nevertheless, the use of those tags and markers including the use of urban slangs and shortened words makes Twitter a communication tool for the youth. Their use though may be coherent to the readers or users of the platform even though they may seem grammatically incoherent and score low on lexical density. This may increase sharing action by the younger generation who make up the bulk of twitter's users. One of several measures for the readability of texts is the lexical density. This measurement has been adapted and used to measure the textual contents and readability of textbooks used in our school system. It has been shown that the higher the lexical density score the more difficult it is to read and vice versa (Nesia and Ginting 2014). For example, in measuring the lexical density of the reading texts for grade twelve, researchers found out that the highest lexical density in their texts could be found in the explanation texts and accounted for a lexical denser score of 58.42% which was slightly higher than that of the narrative texts which was 43.97%. It was then suggested that further textbook writers may focus on lowering the lexical density of reading texts so that it could be better understood by both students and educators (Nesia and Ginting 2014; To et al. 2013). Similarly, if the textual contents in the tweets are more readable, we expect that the lexical density of the texts would be lower and the text more understood and shared by a larger population. We therefore hypothesize:

H4: *An original tweet's lexical density is negatively associated with its retweet count*.

Novelty of news has not been extensively examined in the context of false news except for Vosoughi et al. (2018), who attempted to measure differential effects of rumors and real news using cascades. The researchers further found that false rumors contained more novelty than true news and that people were more likely to share information that is deemed novel. The authors noted that it is possible that the novelty of rumors helps to propel sharing to such an extent that they become contagious. Vosoughi et al. (2018) found that falsehood reached far more people than those that contained the truth on Twitter, which means that the more novel a news story is for false news the faster they are shared at the initial stages and thus they may have a higher likelihood of being retweeted. Another research noted that, although false political news during elections travel fast on social media, it was not just the act of false news being retweeted that drives the sharing behavior but other components of the news story (Ehrenberg 2012). For example, if information not privy to the public on a political opponent is shared on social media, receivers of the news may feel more inclined to share based on the newness or awe factor of the news story which may also be further propelled by the individual's political biases. New or risky information is what often increases the diffusion of false news but not real news in a network. Since false news has the added edge of being novel (Vosoughi et al. 2018), false news will be tweeted more. Hence, we hypothesize:

H5: *The positive association between novelty and retweet count is stronger for fake news than real news.*

A recent study showed that trends between real and false news were quite distinguishable with factors such as readability and positivity and that those factors coupled with the credibility of the news stories were hallmarks for real news, while aesthetics novelty and negative sentiments were found to be synonymous with false news stories (Osatuyi and Hughes 2018). Furthermore, as stated by Stieglitz and Dang-Xuan (2013), affective or emotional messages or information tend to be more viral. However, with regards to the interactive effect of false news and its sentiment, a recent study is of the view that false news stories will have more negative valence and as such may have a higher retweetability (Osatuyi and Hughes 2018). However, those studies were conducted in controlled environments. On a relatively normal day, there is the likelihood that negative news may be more out of the norm and receive a larger amount of sharing, while real news will be synonymous with the current state of affair and not be "new" and hence receive a lower probability of sharing. In contrast, during extreme events when there may be more negative news stories, any news that is positive may be received more fervently and may induce more sharing. Though in general, false news may be retweeted more, the negative or positive valence of the message may increase the sharing depending on the valence of the current climate. We therefore hypothesize:

H6: *The positive association between sentiment and retweet count is stronger for fake news than real news.*

The presentation and format of a news item has been used to influence users and visitors of a medium to subscribe to and remain engaged to stories (Kim and Dennis 2017). However, in recent times the format of news stories has been used as a tool for change, example being the change of users' current beliefs and ideas. A recent study showed that the format of a news piece affects its believability without recourse to the credibility of the source (Kim and Dennis 2017).

Osatuyi and Hughes (2018) examined false versus real news using the elaboration likelihood model and explained that this might be because false news contained less information implying that such news stories may have relatively lower lexical density. The authors argued that false news with less information will usually appeal more to our affective states as compared to real news which appeals to our cognitive states. The introduction of false news to a receiver may affect the recipients' affective state and may likely induce a reaction or action which in some cases is translated to the act of sharing the news story. On the other hand, real news, which we posit will usually have a higher lexical density due to the inclusion of more content words as compared to false news may only result in minimal action due to its appeal on an individual's cognition but not their affective state. Affective engagement cause more knee jerk reactions while cognitive processes influence more somber decision making processes (King and Sun 2018). Moreover, as stated in prior research, false news contains less information due to their dependence on attracting the use of an individual's peripheral route of information processing instead of their central route of decision making. We therefore hypothesize:

H7: *The negative association between lexical density and retweet count is stronger for fake news than real news.*

Data and Methodology

Sample and Variables

In this research, we examine the spread of real and false news on the Twitter network. We collected 42,638,147 unique tweets through Twitter's API from August 18 to September 22, 2017 during the Hurricane Harvey disaster. We filtered out all retweets from the dataset and retained only original tweets whose veracity (real or false) were consistently verified using three independent fact checking websites Snopes, Factcheck.org and BS Detector as well as information from the rumor control page of the Federal Emergency Management Agency (FEMA), a United States' government agency set up to respond to emergencies. This was accomplished by extracting the keywords of both false and real news stories from the last part of their respective websites' URLs. For example, "harvey-relief-donation-rumors" was extracted from "https://www.snopes.com/harvey-relief-donation-rumors." Next we consolidated our data by removing tweets that were redundant. We did this to ensure that only verifiable false and real news were used in our analysis. The total number of verified original real and false news topics within that period was 3,589 tweets that were shared 31,623 times. For each of these verified tweets, we collected their total retweet count throughout the period of the Hurricane. Table 2 summarizes our variables and their definitions.

We measured novelty based on the number of days that had elapsed between the first tweet on the same topic and the day the subsequent tweets were introduced into the Twittersphere. In order to calculate novelty, we first identified each tweet's topics using LDA (Albuquerque et al. 2019; Blei 2003). We began with an initial allocation over 50 topics. This model further arranges the distribution based on topics in their order of arrival relative to the first time the original tweet was posted using standardized metrics of one day between each topic. That means each tweet fell within a spectrum of Day 0 (the first day the original tweet was posted) to Day n (the nth day after the original tweet occurred). We defined the baseline novel tweet as the first news item on a topic. We measured novelty based on days because it is more manageable and easier to identify novelty based on days than seconds or minutes. Furthermore, Hurricane Harvey occurs as a "day" event . As part of our preprocessing and data preparation, we extracted links and hashtags, images and videos from each tweet and cross referenced them with data crawled from the fact checking websites. We performed this analysis because there were tweets that did not have any textual contents but rather contained links to images and videos. Furthermore, those links must be traced and if the tweet contained only images, they were cross referenced to verify their veracity.

Table 2. Variables and Definitions.

Variable	**Definition**
Dependent Variable	
Retweet count	The number of times an original tweet was retweeted during the first 24 hours it came out.
Independent Variables	
Veracity	1 if the original tweet is true and 0 if false.
Sentiment	Positive (1), neutral (0) or negative (-1) sentiment
Novelty	The number of days that elapsed between the first tweet on the same topic and the day the current tweet came out (reverse-coded).
Lexical density	The proportion of lexical words to the total number of words.
Control Variables	
URL	1 if there is at least one URL in the tweet, 0 otherwise.
City	1 if a city is specified in the original tweeter profile, 0 otherwise.
ln(Follower_count + 1)	The natural logarithm of one plus the original tweeter's number of followers.
ln(User_list_count +1)	The natural logarithm of one plus the number of public lists that the original tweeter is a member of.

ln(User_status_count +1)	The natural logarithm of one plus the number of tweets (including retweets) posted by the original tweeter.
ln(User_favorite_count +1)	The natural logarithm of one plus the number of public tweets that the original tweeter has liked.
ln(User_friend_count + 1)	The natural logarithm of one plus the number of people the original tweeter is following.
Social	1 if the original tweet is a news item with social and political connotations, 0 otherwise.
Environment	1 if the original tweet is a natural disaster, weather, or environment-related news item, 0 otherwise.

We initially specified a Dirichlet Model using 20 passes with an initial set of 50 topics. We selected 50 topics due to the probability nature of LDA and then narrowed it down to seven major false news topics and seven real news topics. This created a tweet distribution probability. Then we reclassified the resulting output based on their tweet IDs in chronological order.

Sentiment and lexical density of the tweets were analyzed using both supervised and unsupervised learning algorithms as behaviors and emotions can be deduced form reactivity contents (Liang et al. 2016). We used SpaCY and the NLP algorithms in Python to calculate the sentiment and lexical density of the tweets. Since SpaCY does not come with a pre-created sentiment analysis model, we implemented a text classifier for sentiment analysis using its NLP library. We performed tokenization, clean up, POS tagging, dependency parsing and converted our data into feature vectors to be fed into our model. The filter method was chosen in this context over the wrapper method because of its efficiency with large datasets (Huei Chou et al.

2010). We extracted and calculated the use of first- and second-person pronouns, the word count, lexical structure, and corresponding density of the tweet, pronouns, nouns, sentiment, hashtags and emotions in each tweet. SpaCY was selected due to its accuracy, speed and ability to handle larger datasets, and its capability to select the most accurate and state of the art algorithm for the task (Malhotra 2018). The sentiment analysis in SpaCY was initialized using a neural sentiment classification model that was trained using Keras in SpaCY. SpaCY splits the corpus into sentences and not words such that each sentence is classified. The polarity and subjective scores for each sentence are then saved. The polarity score is the raw sentiment orientation of the textual content, which ranges from 1 to 9 for positive sentiment, 0 for neutral and -1 to -9 for negative sentiment. Due to the size of our data and number of variables, the use of the raw score may negatively affect our classifier. Therefore, we recoded our raw polarity scores to an ordinal variable of positive (1), neutral (0) or negative (-1) sentiment.

We included the following control variables: user follower count, user friend count, user status count and user favorite count. Recent studies have shown that these variables using bivariate, univariate and multivariate distributions are very important predictors of retweet behavior on Twitter (Hoang and Mothe 2018; King and Sun 2019).

Table 3 summarizes the sample descriptive statistics.

Table 3. Sample Descriptive Statistics (N=3,589)

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Dependent variable				
Retweet count	8.811	227.499	0	10,919
Independent variables				
Veracity	0.495	0.500	0	1
Sentiment	-0.329	0.660	-1	1
Novelty	3.677	3.586	0	27
Lexical density	19.488	13.846	0	99.25
URL	0.904	0.294	0	1
City	0.782	0.413	0	1

Follower count	50,008.42	1,021,346	0	41,700,000
User list count	429.171	4,477.541	0	196,072
User Status count	90,719.84	327,577	3	9,076,030
User favorite count	13,951.6	37,425.71	0	652,220
ln(Follower_count + 1)	6.826	2.408	0	17.546
ln(User_list_count +1)	3.460	2.137	0	12.186
ln(User_status_count +1)	9.967	1.836	1.386	16.021
ln(User_favorite_count +1)	6.568	3.335	0	13.388
ln(User_friend_count + 1)	6.461	2.121	0	13.053
Social	1.727	1.483	0	3
Environment	0.358	0.479	0	1
*ln*Sentiment*Veracity	-0.202	0.454	-1	1
*ln*Lexical*Veracity	9.513	13.183	0	95.722
*ln*Novelty*Veracity	2.074	2.923	0	18
*ln*Social*Veracity	0.625	1.219	0	3
*ln*Environment*Veracity	0.254	0.435	0	1

Empirical Models

Our dependent variable – the retweet count – is a count with excessive zeros. We ran tests for over dispersion and confirmed the variance of the retweet count was much higher than the mean. Hence, we used the zero-inflated negative binomial regression to analyze the data (Rodrıguez 2013). The density of the retweet count follows the distribution of a mixture of two states - a zero state where we observe no retweet and a state where we observe a positive number of retweets.

$$\text{f}\,(Retweet\_Count_{ij}) = \begin{cases} \varphi_i + (1-\varphi_i)NB(0)\ if\ Retweet\_Count_{ij} = 0 \\ (1-\varphi_i)NB(Retweet\_Count_{ij})\ \ if\ Retweet\_Count_{ij} > 0\text{’} \end{cases}$$

where *Retweet_Count*$_{ij}$ is the number of times an original tweet *j* by user *i* was retweeted, NB(.) is a negative binomial distribution with mean μ_{ij} and variance $\mu_{ij}(1 + \alpha\ \mu_{ij})$. μ_{ij} is specified through the following function:

$$\mu_{ij} = exp(\beta_0 + \beta_1 \cdot Veracity_{ij} + \beta_2 \cdot Novelty_{ij} + \beta_3 \cdot Lexical\_density_{ij} + \beta_4 \cdot Sentiment_{ij} + \beta_5 \cdot Social_{ij}$$

$$+ \beta_6 \cdot Religion_{ij} + \beta_7 \cdot Environment_{ij} + \beta_8 \cdot Sentimen_{ij}t \cdot Veracity_{ij} + \beta_9 \cdot Novelty_{ij} \cdot Veracity_{ij}$$

$$+ \beta_{10} \cdot Lexical\_density_{ij} \cdot Veracity_{ij} + \beta_{11} \cdot Social_{ij} \cdot Veracity_{ij} + \beta_{12} \cdot Religion_{ij} \cdot Veracity_{ij}$$

$$+ \beta_{13} \cdot Environment_{ij} \cdot Veracity_{ij} + \beta_{14} \cdot \ln(User\_list\_count + 1)_i + \beta_{15} \cdot \ln(follower\_count + 1)_i$$

$$+ \beta_{16} \cdot \ln(User\_status\_count + 1)_i + \beta_{17} \cdot URLs_{ij} + \beta_{18} \cdot User\_URLs_i + \beta_{19} \cdot City\_Present_i \ )$$

(1)

Table 4 summarizes the results of our empirical analyses. All models had variance inflation factors (VIFs) less than 5. Hence, multicollinearity is not an issue in our empirical analysis. Our first model depicts the retweet count as a function of the control variables. This is our baseline model. User follower, friend, status and favorite counts, are the zero-inflated variables as they may influence the probability of users on the network not responding or retweeting the message. We find that all the control variables were significant at the 0.01 level for both our count and in the estimation of the zero inflation. However, the intercept was not significant in the baseline model. All inflated portions were consistently significant for all models. The Likelihood-ratio test's alpha was significant at 0.01 for all models and showed that there are differences between the negative binomial and the zero inflated Poisson models and that the negative binomial is the best method to use.

Table 4. Results of Negative Inflated Binomial Regression models (N=3,589).

	Model 1	**Model 2**	**Model 3**	**Model 4**
Intercept	-0.251 (0.661)	1.191** (0.561)	2.378*** (0.624)	2.880*** (0.717)
Veracity			-0.307** (0.138)	-1.500** (0.600)
Novelty			-0.039*** (0.017)	-0.062** (0.025)
Sentiment			-0.231*** (0.017)	-0.462*** (0.024)
Lexical_density			-0.024***	-0.027***

			(0.005)	(0.006)
Environment		-1.644*** (0.262)	-1.686*** (0.268)	-1.728*** (0.463)
Social		0.200 (0.086)	0.034 (0.090)	-0.204 (0.134)
Novelty*Veracity				0.006 (0.038)
Sentiment*Veracity				0.801*** (0.035)
Lexical*Veracity				0.018 (0.011)
Social*Veracity				0.518*** (0.178)
Environment*Veracity				0.539 (0.572)
URL	-1.466*** (0.205)	-1.461*** (0.191)	-1.672*** (0.197)	-1.449*** (0.206)
User location	0.982*** (0.188)	0.895*** (0.177)	0.739*** (0.181)	0.599*** (0.186)
ln(User_status_count +1)	0.336*** (0.056)	0.193*** (0.047)	0.198*** (0.047)	0.193*** (0.048)
Inflation				
Intercept	1.053*** (0.508)	1.410*** (0.476)	1.402*** (0.476)	2.880*** (0.717)
ln(Follower_count + 1)	-1.002*** (0.074)	-1.014*** (0.073)	-1.016*** (0.073)	-1.023*** (0.073)
ln(Friends_count + 1)	0.204*** (0.071)	0.220*** (0.070)	0.224*** (0.070)	0.228*** (0.070)
ln(User_fav_count +1)	-0.288*** (0.031)	-0.285*** (0.030)	-0.288*** (0.030)	-0.286*** (0.030)
ln(User_status_count +1)	0.689*** (0.062)	0.658*** (0.059)	0.660*** (0.059)	0.658*** (0.059)

Notes: ** p<0.05; ***p<0.01.

In Model 2, we added two dummy variables (i.e., environment, social) representing the category of the tweet with religion being the base category. The results show that the retweet counts of environment-based tweets were significantly lower than religious tweets at the 0.01 level. However, the retweet counts of social tweets were not statistically different from those of baseline religious tweets.

In Model 3, we added our independent variables including veracity, novelty, lexical density and sentiment. The coefficients for novelty, lexical density and sentiment were all negative and significant at the 0.01 level, while veracity was negative and significant at the 0.05 level. Our results show that retweet count is higher for fake news than real news. This is consistent with both H1 and the previous literature where false news were more likely to spread and were shared faster than real news (Vosoughi et al. 2018). The coefficient estimate for novelty was negative and significant in Model 3 at the 0.01 level. Because the novelty variable was reverse-coded, this result suggests a positive relationship between novelty and retweet count. Hence, H2 was supported. This is consistent with previous studies (Berger and Milkman 2012) which showed that novel news items were more likely to go viral due to their ability to attract human attention. The coefficient estimate for sentiment is negative and significant at the 0.01 level, indicating that sentiment was negatively associated with retweet count. Hence, H3 was not supported and tweets with negative sentiment diffused faster than neutral or positive ones. Though we expected that positive tweets or messages may be more influential during extreme events, we find that negative tweets were shared more than positive and neutral messages. This is supported by the previous literature that showed that emotionally charged messages had a higher propagation chance than neutral messages (Stieglitz and Dang-Xuan 2013).

The coefficient estimate for lexical density was negative and significant at the 0.01 level. Hence, the lower the lexical density, the higher the retweet count, supporting H4. This could be a result of the higher readability of the tweets with lower lexical density and the use of the medium by the younger generation.

In Model 4, we included the interaction terms between veracity and the independent and category dummy variables. The coefficient estimates and their significance levels were consistent with those from Model 3. Among the newly added interaction terms, the coefficient estimate for the interaction term between sentiment and veracity was positive and significant at the 0.01 level. The coefficient estimate for sentiment was negative for false news (β=-.462) but positive for real news (β=-.462+.801=.339). This result suggests that the relationship between sentiment and retweet count is different for real vs. fake news. That is, the retweet counts for negatively charged fake news tweets were higher than those of positively charged fake news tweets, *ceribus paribus*. In contrast, the opposite was true for real news where a positive sentiment stimulates diffusion. Hence, H6 was not supported. The interaction term between the social tweets dummy variable and veracity is positive and significant at the 0.01 level, indicating that real social tweets were retweeted more often than fake social tweets. The interaction terms between other variables and veracity were not significant. Hence, H5 and H7 were not supported. This may be attributed to the desire for more social-related news that may seem true. Such news may give the audience a break from the norm of receiving weather-related news. Such true news also has the added effect of being less dramatic and thus appealing to the cognitive processes.

Discussion

Using data collected during Hurricane Harvey, this research examines the factors that affect the diffusion of real and false news on Twitter using the retweet counts. Our research has the following theoretical contributions and practical implications.

Theoretical Contributions

This study makes several contributions to the literature on fake news and information diffusion on social media. First, we develop a framework for predicting the aggregate retweet behavior of social media users on Twitter by combining textual and feature characteristics from the network. We identify not only the important predictors of retweet counts but also reveal the importance of the interactions of both the main effects and category variables in our study. The inclusion of the interaction terms show that the effects of the variables were not simply additive. To the best of our knowledge, this is the first time such interactions and variables have been included and applied in the context of verifiable fake versus real news studies.

Second, we identify factors such as veracity, novelty, sentiment, lexical density, and news category that affect the diffusion of information on the Twitter network. Our results show that people are more likely to retweet falsehood as compared to real news. This is consistent with the previous literature (e.g., Friggeri et al. 2014; Vosoughi et al. 2018). One explanation is that false news generally contains less information, has an element of surprise and as such can be more easily adopted by the masses (Osatuyi and Hughes 2018). The authors used the elaboration likelihood model (ELM) to explain this phenomenon and argued that such news articles appeal to our affective emotion that rely on the peripheral route of information processing rather than the central route that focuses on the information itself. In this context, the behavior of an individual would be accompanied by individuals not verifying the news stories but readily spreading them because they may rely on their peripheral and not the central route for information processing. We also find that the diffusion of tweets gradually abates when the news stories no longer retain their novelty. This may be because of the tendency of novel news to be more valuable and awe inspiring such that it inspires attention and ultimately sharing (Itti and Baldi 2009). This is consistent with previous studies that show contents that evoke high emotions like awe is generally more viral (Berger and Milkman 2012). However, as time goes by during the diffusion process when more people are likely to see, hear or read about the news stories from other channels, the less novel it may eventually become. At that point the news may no longer retain its novelty and potential to be viral. Our results further showed that tweets with negative sentiments had a higher retweet count than neutral and positive news. Recent studies had shown that virality may be facilitated by emotion and negative contents on social media (Bene 2017; Berger and Milkman 2012). This might be because the salience of negative emotion-filled contents such as sadness and anger may affect receivers of the news stories in a network and induce reactivity. We also find that the lower the lexical structure the higher the retweet count. This result was expected as micro bloggers predominantly adopted Twitter due to their use of emojis, slang words, and short words that at most times have no real grammatical meaning. As such the lexical components would be lost, and the acceptance for that language may be high. Studies have shown that microblogs such as Twitter were not originally designed to support conversations and can be highly incoherent with regards to syntax bidirectional responses (Honeycutt and Herring 2009). In addition, the lower lexical density means that the message is more readable to a much larger audience and is usually devoid of more grammatically complex words. This translates to the text being more readable for a larger number of people on the network. Our analysis of all news categories showed that news stories about the environment are retweeted less often whilst social news items were not significantly different from the baseline religious messages. News about the environment negatively affected retweet count which contradicts a recent study that showed that with regards to extreme weather, information that is derived from credible sources during such shock events usually helps suppress and control some forms of rumors on social media (Oh et al. 2010). One explanation may be that environmental news may be a reminder of their current predicament and as such individuals may be less inclined to be associated with such news stories. The environmental news story may be depressing during shock events such as hurricanes or earthquakes and most of the news may be negative. The combination of these factors may cause them to garner less retweets or shares than other news stories. Furthermore, with the abundance of environmental news stories during extreme events, users may not want to be associated with the spreading of more negative non-novel news but may be more inclined to share other types of news stories. Users in the network may then be inclined to share news that are more positive and contain more useful information about their situations and of the environment, e.g., the receding of the floods or the availability of aid. A possible explanation for the insignificant result for social news may be because there may be some similarity between social news and religious news. Hence, social news stories may not be viewed any more differently from the baseline religious news stories. Another explanation may be because they are inherently considered as both social type news and as such may not be shared any more differently from each other.

Third, we further compare the impacts of the above-mentioned factors on the retweet count for fake versus real news. Our results show that even though negative sentiment propels diffusion on social media, this relationship only holds for fake news. That is, negative sentiment promotes the diffusion of fake news. This may be caused by people being more inclined to retweet false negative news as compared to false positive or neutral news. During shock events, the awe of messages coupled with the negativity and novelty of false news may create a combined effect that makes it very easy to attract attention. As per other studies, false news has some novelty about it and negative sentiment has also been known to cause sharing. In contrast, positive sentiment promotes the diffusion of real news. The results further showed that people may be more inclined to retweet real news that is inherently more positive than those that were negative or neutral. This in an interesting finding, and a possible explanation is that real news does not usually come with as much awe as false news stories. As such real news that is positive will be more accepted. This is consistent with previous literature that shows real news appeals more to one’s central route of information processing and decision-making and does not rely on awe inspiring headlines or words (Osatuyi and Hughes 2018). Besides, during shock events, when there is already an overwhelming abundance of negative awe-inspiring news stories, a real news that is positive may hold some novelty.

In addition, social real news stories diffuse more than social fake news. This is an interesting result which contradicts previous studies as politics is inherently a social topic and its falsehood has been known to go viral quite quickly (Botha 2014). Misinformation using political discourse has been a worry to academia for quite some time (Ehrenberg 2012). The impacts for novelty and lexical density are not statistically different for real versus fake news. This means that the impact of novelty of a news story on diffusion is not propelled by the veracity and most people may pay more attention to the uniqueness of the message or how new the message is and less on the veracity of the news stories. The uniqueness of the messages may be the motivational entity in the messages that allows for those who interact with the messages to share irrespective of the veracity of the news stories. Therefore, whether true or false, the user may not be motivated enough to verify the information and be baited solely on its novelty. This is consistent with previous literature that showed that surprises when measured using the human brain did not require semantic understanding of the data (Itti and Baldi 2009). Lexical density is the readability of information, and logic dictates that the easier a news story is to read, the higher the total number of people who understand the message. The lexical component that are synonymous with microblogging sites such as Twitter may compel users who interact with the message to use their peripheral route of information processing rather than their central route. And users who use the peripheral route of information processing have been known to exhibit negative security behaviors without discourse to consequences (King and Sun 2018). Even though textual contents are usually associated with extending cognitive efforts, this may not be the case of Twitter as microblogs are not originally designed to support conversations and can be highly incoherent with regards to syntax and character limitations (Honeycutt and Herring 2009). The use of taglines and slangs may cause users to follow whatever trend there is without the actual verification of news or information. The trending of news stories may take more precedence over the story itself and the lexical effects may be the same for both real and false news. This may account for why both lexically dense false and real news are shared equally. In addition, environmental real and fake news do not differ in their diffusion. News stories on the environment, real or false, may be viewed the same due to the constant information of the environment being interacted with. The users in the network may not feel the need to differentiate the false versus the real environmental news stories due to the need of being the first to share the news. Environmental news being in abundance and having already a negative impact may mean any new information irrespective of it being false or real may be seen as valuable and accorded with the same level of importance (Aral and Dhillon 2016). This effect means that the veracity of the news story may not really be seen any differently for real or false and as such may not be shared any differently.

Fourth, we show the importance of additional control variables such as user location and user status count on news diffusion in social media. Our control variables including if there is a URL present in the message, the user location and the user status count are all significant predictors of the retweet count. A text containing URLs normally would mean that it contains an external link which could be perceived by the receiver as evidence of authenticity of what the message was purported to imply. Even though not in the context of falsehoods, the presence of URLs in a message has been known to be a general predictor of an individual's sharing of the message (Lee et al. 2014). Tweets from users whose city is known are retweeted more often. This may be attributed to trust. Followers may be more inclined to retweet from a user whose location or city is known because they trust that the user is human and not a malicious automated entity or a troll. It may signify that they have nothing to hide and as such trust may be reciprocated. Also, when an original tweeter is in the immediate area or proximity of a shock event, there is the tendency for others to believe his tweets as firsthand and as such people may retweet them more often. Also, a user's status count signifies the number of tweets posted by the user. It is possible that users who post more tweets are more active users who have accumulated a large network of followers. Hence, their tweets may be retweeted more often by their followers and the network.

Fifth, our research reveals important factors such as the original tweeter's follower count, friend count, favorite count, and status count as important predictors of the inflated zero counts in the data. That is, these factors are predictors of tweets having an excessive zero retweet count. An individual's follower count is negatively related to having a zero retweet count. When a user has a large network of followers, the chance of at least one follower retweeting the message is higher, and the chance of having a zero retweet count is reduced. An original tweeter's friend count is positively related to having an inflated zero retweet count. A tweeter with many friends or following many users may mainly use Twitter to follow others and obtain news and information rather than sharing information with others. As a result, when they post an original tweet, the retweet count is more likely to be zero. An individual's favorite count is the number of tweets that the user has liked and is negatively related to the chance of the user's original tweet having a zero retweet count. Users with high favorite counts are more active Twitter users who are more familiar with the norms of writing tweets. Hence their original tweets may be better received by others and have a higher chance of being retweeted at least once. Users with higher status counts tweet or retweet more often. The large number of tweets they send out may be the reason their original tweets are less likely to be retweeted. That is, they inundate their followers with large numbers of tweets on a regular basis, which causes the followers to less likely to retweet them out.

Practical Implications

This study has several practical implications. First, the models used in this study may help system administrators and policy makers predict the retweet counts of false and real news. System administrations and policy makers may be able to use the factors and features that are significant predictors of retweet count to develop systems to predict if a message may be retweeted or not. The use of the features responsible for individuals' retweet behavior on social media can be used to gauge the potential damaging effects of false news stories and if possible, nip it in the bud before it is disseminated more widely. This knowledge gathered from the prediction may better help them prepare in the event of fall outs from false news. They may be able to use this to project future diffusion events.

Second, policy makers such as FEMA trying to mitigate the spread of false news may be better equipped to reduce the flow of falsehoods during extreme events. Equipped with the ability to predict retweet behaviors, policy makers may be able to prevent the spread of misinformation on social media by deleting falsehoods that is predicted to reach a wider audience before it does.

Third, this study has the potential of assisting policy makers and social media platforms in developing new policies, regulations, and strategies that facilitate the diffusion of real information while preventing the propagation of false information. These results and models may also be used in the development of emergency warning systems that may be able to reach a larger audience especially during extreme weather events. Using the major features that are good predictors of information sharing, policy makers and agencies such as FEMA may be able to use this model outside of the false news nomology and into the general sphere of society by being able to predict and develop their own effective systems. They would be able to create and send out messages and information to the public during extreme events that may be able to reach a wider audience in a shorter period. This may be the difference between life and death in cases of extreme events like earthquakes, forest fires and hurricanes. The system can used to determine the potential of a message not being shared at all and to better redesign and improve such messages to reach a wider audience. Doing this would mean that users may now receive messages quicker during extreme events, and it might help reduce cost and material damages. The effective dissemination of information during such periods may save countless lives and improve people’s lives.

Conclusions and Limitations

We examine various factors that influence the retweet count of news items on Twitter and how they differ for real versus fake news. Using a combination of text mining and machine learning techniques, we show that fake, novel, and negatively-tuned news as well as those with lower lexical density diffuse more on social media. In addition, environment tweets diffuse less compared to other types of tweets. Our research also reveals the differential impacts of these factors on real versus fake news. Specifically, negative sentiment promotes the diffusion of fake news while positive sentiment stimulates the sharing of real news. Social real news is also retweeted more often than social fake news. Also, our research shows that the effects of sentiment are different for false news versus real news.

This study has several limitations. First and most importantly, our sample was collected during Hurricane Harvey and was further reduced to include only verified real and fake news. This may limit the generalizability of our findings. Future research can collect a more comprehensive dataset to cross-validate our results. Second, we analyzed the aggregate retweet counts. Future research can instead analyze the retweeting behavior at the individual level. Finally, due to limitations of the Twitter data that exhibited signs of deleted tweets, we were unable to collect the retweet count after Day 1. Future research can systematically collect completed data in order to accurately estimate the propagation of tweets.

CHAPTER III

DYNAMIC EFFECTS OF CORRECTING FALSEHOODS ON SOCIAL MEDIA

Introduction

Studies have shown that a vast majority of Americans are easily deceived by false news, with 50% of the American public willingly endorsing at least a conspiracy theory (Oliver and Wood 2014) and up to 75% believing false news headlines (Silverman and Singer-Vine 2016). Online social networks make it incredibly easy for the propagation of misinformation (Kazienko and Chawla 2015), not just due to the dynamic nature of the data but that the vast majority of the public relies on microblogging sites such as Twitter for news and information (Reuters 2017).

Over the years, information systems (IS) researchers have also investigated the concept of misinformation (Agrawal et al. 2013; Jindal et al. 2010; Oh et al. 2010), with several studies suggesting ways of combating this phenomena (Friggeri et al. 2014; Torres et al. 2018, 2018). One such proposed method is the use of correction messages that are meant to mitigate potentially disruptive rumors that can drive a cycle of extreme repercussions. However, such proposed methods may be ineffective if they end up causing blowbacks, which is counterproductive (Lewandowsky et al. 2012). Quite recently the Department of Homeland Security (DHS) in recognition for the need to address this issue established a 22-member agency Social Media

Working Group for Emergency Services and Disaster Management (SMWGESDM), to which FEMA is a member, to help provide recommendations to curb the menace. The agency published a set of guidelines for dispelling rumors in its 2018 white paper report (*SMWG* 2018). However, results from these studies have not only been fragmented, inconsistent, and inconclusive but also lack the power of inference. For example, some studies show that when corrections are used there is a high reduction in reported misinformation (Ecker et al. 2011), while others show insignificant results (Jolley and Douglas 2014). Most studies relied on surveys mainly due to the unavailability of adequate data (Cameron et al. 2013; Lewandowsky et al. 2012; Oh et al. 2010; Schwarz et al. 2007).

In this study, we investigate the relationship between falsehoods and correction messages both theoretically based on the competitive exclusion principle and empirically using a panel data set on the diffusion of falsehoods and correction messages. Specifically, we examine the following questions:

1) Is there a bidirectional relationship between the diffusion of falsehoods and correction messages on Twitter during shock events?
2) What are the effects of correction messages on falsehoods and vice versa?

To address these questions and advance our knowledge of this very crucial yet intricate relationship between falsehoods and correction messages, we first develop a theoretical model on the diffusion of both falsehoods and correction messages based on the competitive exclusion principle. We model the two types of messages as two species competing for limited resources in the environment and show the conditions under which they both die off, both survive, or correction messages survive but falsehoods die off. Next, we verify our theoretical results empirically using a unique panel dataset containing 279,597 social media interactions across five weeks during Hurricane Harvey in 2017 and Hurricane Florence in 2018, respectively, to investigate the bidirectional relationships between the diffusion of falsehoods and correction messages. We employ the panel vector autoregression (PVAR) methodology to investigate the dynamic feedback effects of falsehoods on social media and correction messages from FEMA and fact-checking organizations on each other's diffusion.

Our study has the following contributions to the literature on information diffusion, misinformation and rumors. First, our research is the first to holistically examine the diffusion of falsehoods and correction messages together and their dynamic effects on each other. As a result, we have a better understanding of how falsehoods and correction messages co-diffuse rather than just examining the diffusion of each type of messages on their own. Second, the combined use of a theoretical model and empirical validation allows us to first identify scenarios under which each type of messages survive or die off and then empirically validate the theoretical results using real Twitter data during Hurricanes Harvey and Florence. Third, we introduce the competitive exclusion principle to IS research and apply it to the study on the diffusion of falsehoods and correction messages on social media using a deterministic model of system equations that depicts the competition between two species. This is a new theory to the IS discipline and our theoretical model results show that increasing the replication or per capita rates of correction messages causes falsehoods to eventually die out and correction messages to thrive. We derive four possible outcomes where in the first scenario, the number of tweets and retweets for both falsehoods and correction messages both die out eventually. In the second and third scenarios, where the per capita growth rate of falsehoods is less than the decay for correction messages and vice versa, one species (either falsehoods or correction) dies out and the other survives. The second scenario where falsehoods die off and correction messages survive is the ideal scenario. And finally, we illustrate a rare but possible scenario where both falsehoods and correction messages survive. In this scenario, one of the species survives in a different niche. In addition, our empirical results provide validations of the theoretical predictions based on the principle and demonstrate its applicability in IS research. Fourth, our empirical research shows counterintuitive results that falsehoods cause an increase in correction messages and rebuttals and the current state of correction messages is ineffective in reducing falsehoods. This solution has the potential to inform policy makers, social media administrators and developers of emergency warning systems on how to effectively combat the propagation of falsehoods while improving the diffusion of correction messages.

Related Literature

In this section, we first discuss the literature on proposed solutions for reducing falsehoods on social media. Next, we introduce the competitive exclusion principle.

Recent Proposed Solutions

There have been several major proposals for curbing misinformation. For example, researchers have recommended the use of detection (King and Sun 2019; Ma et al. 2016) to target automated entities on social media and prevent their spreading of falsehoods. While this might seem like an effective solution, several studies using large datasets have shown that falsehoods are not primarily shared by automated entities but rather by humans (Friggeri et al. 2014; Vosoughi et al. 2018). Others have also proposed the use of source credibility as a solution to this phenomenon (Agrawal et al. 2013). However, the results of these studies have proven to be quite inconsistent. For example, Oh et al. (2010) showed that credible sources may be able to lower anxiety and be more successful in suppressing falsehoods, while a more recent study (Lewandowsky et al. 2012) stated that appealing to coherence was more successful in reducing falsehoods than relying on source credibility. Finally, a recent research (Kim et al. 2019) proposed the use of source rating in flagging and suppressing falsehoods while penalizing the disseminators of falsehoods. Using a survey instrument, the authors found that presenting news stories in a story format with source ratings may cause users to evaluate the news contents' veracity more critically. Such a solution though is mostly used in the retail domain where users rate credible buyers and sellers. However, such rating systems also falls prey to manipulations and can be extremely hard to detect (Kumar et al. 2018).

Information Correction

IS researchers have examined information correction using predictions (Hovland 1959) and inoculation theories (McGuire 1964). In recent times, government agencies created rumor control mechanisms and aim to identify, investigate and mitigate potentially disruptive rumors especially during extreme events such as during the aftermath of the great East Japan Earthquake on March 11, 2011 (Takayasu et al. 2015). While results on their effectiveness have been mixed and inconclusive, questions arise as to the ideal ways in which to create and deploy effective correction mechanisms to combat falsehoods. For example, some studies show that including both the facts as well as the falsehood in the same message causes engagement and leads to an overall increase in knowledge about the falsehoods (Cameron et al. 2013). However, using qualitative studies and reviews from previous studies. Schwarz et al. (2007) showed that repeating falsehoods could be detrimental to the overall correction efforts because receivers of the messages may not be able to differentiate facts from falsehood and may end up misremembering the message. Quite recently, using a meta-analysis of several studies, Walter and Murphy (2018) showed that corrections may reduce misinformation across diverse audiences and may be more successful in informing the receivers. Their study was partially supported by findings from Huang (2017), whose results showed that corrections reduce people's beliefs in specific rumor contents but the senders of the messages were often unable to recover the trust that was lost due to the falsehood. Other studies have introduced factors such as source credibility and coherence in tackling against misinformation correction. For example, Lewandowsky et al. (2012) showed that, when sending a correction message, appealing to coherence was more successful at minimizing influence of misinformation than using fact checking and source credibility. This is in direct contrast to Oh et al. (2010), who showed that using reliable information with credible sources can lower anxiety and be much more successful in suppressing falsehoods. Another method currently used by rumor control mills is the pairing of correction messages with warnings which were supported by Ozturk et al. (2015). The results showed that, when paired with warning messages or counter messages, rumors tend to reduce their spread. However, this is the first study of its kind to use Twitter data prior to, during and after shock events and captures in its entirety the interactions between a government agency, fact checking organizations and their responses to falsehood.

In summary, our review of prior research reveals the following gaps in the extant literature:

1. Due to the fragmented and inconsistent findings, our understanding of the intricate relationship between correction messages and falsehoods is limited. For example, some prior research (Ecker et al. 2011; Lewandowsky et al. 2012) stated that correction is effective in reducing misinformation, while Jolley and Douglas (2014) presented contradictory findings that such methods might not be very effective.
2. Several prior studies on correction have been based on rumors. For example, Ozturk et al. (2015) showed that, when correction messages are paired with warning messages, rumors

tend to reduce in their dissemination. Rumors are social in nature and can be either true or false. It is therefore necessary to investigate verifiable false and correction messages.

3. Findings from prior research have favored the use of small and individual survey samples and may lack reliability (e.g., Cameron et al. 2013; McPeek 2014). Cameron et al. (2013) used a sample of 125 and showed that messages that included facts, myths, and evidence could be used to effectively counteract myths. However, results from such small sample surveys may not be generalizable to other contexts.
4. Existing studies did not treat the diffusion of both falsehoods and correction messages as endogenous that have feedback loops. Furthermore, none has attempted to investigate the potential for causal inference. To the best of our knowledge there has not been any study that treats falsehoods and correction messages as bidirectional loops. As a result, the endogeneity and possible causal relationship between both variables needs to be further investigated.
5. None of the studies has presented a solution supported by theory to mitigate the flow of falsehoods while improving the efficacy of correction messages. Our research examining the co-diffusion of falsehoods and correction will help us better understand how the two affect each other in the diffusion process and allow us to design better mechanisms to counter the spread of falsehoods.

Our goal is to address these research gaps using a theoretical study through the lens of the competitive exclusion framework supplemented by an empirical verification of the derived results using two Twitter datasets. We next introduce the competitive exclusion principle that serves as the theoretical backdrop of our research.

Competitive Exclusion

IS researchers over the years have relied on the use of exploratory and predictive models in understanding the diffusion process (Mei Li et al. 2017). Of the former, epidemic models have been widely used in exploring the intricate relationship between people and the diffusion and interaction of information on social networks (Cheng et al. 2013; Jin et al. 2013; Jindal et al. 2010; Mei Li et al. 2017). Epidemiology models can be applied in research on falsehoods by treating them as a virus which spreads from infected users to susceptible ones (Pastor-Satorras and Vespignani 2001). However, such methods do not allow us to consider the bidirectional relationship between falsehoods and correction messages. To address this limitation, we use the competitive exclusion principle.

Competitive exclusion principle, also known in Ecology as Gause’s law, states that two species competing with one another for the same limited resources in an ecosystem cannot indefinitely exist together with their population values remaining constant (Gause 1932). It further states that, as both jointly utilize a vital resource that is in short supply either in abundance or availability to the species, one species will eventually eliminate the other from the ecosystem (Jaeger 1974). According to the principle, when one of the species has even the slightest advantage over the other, the one with the advantage will dominate in the long term, which ultimately leads either to the extinction of the weaker competitor or to an evolutionary or behavioral shift of the weaker species toward a different ecological niche. What these means is that sometimes the weaker species may end up settling in a different niche or mutating in order to survive while the stronger species survives in the original ecosystem.

Several studies in ecology (Brown 1971; Rácz and Karsai 2006) and biology (Brose 2008; Jaeger 1974) have used the principle to explain behaviors of species in competition in their natural habitat. For example, Brown (1971) studied two species of Chipmunks in a habitat and showed that the competitive advantage can also be determined by the habitat, where the more social chipmunks for example had a competitive edge over the more aggressive one. The author was convinced that habitats can give a competitive edge over aggression. In this scenario, the more social chipmunks were able to reproduce more, and the more aggressive chipmunks ended up wasting energy on fruitless chases. This study showed that the more dominant species does not always come out on top, and there may be other factors such as habits that may influence the competitive advantage. Another study also found out that the aggregation of a weaker species and the length of the competition can have an positive effect and prevent the extinction of that specie (Rácz and Karsai 2006). The researchers stated that aggregation can aid in the slowing of the extinction of the weaker competitor and give it surplus time to evolve and survive.

For the purpose of this study we model false news and correction messages as two different species in competition with each other in a natural environment which is the Twittersphere. We model the sharing of both false and correction news as the replication rates of each and the effect each has on the other as a bidirectional dynamic loop. The domination is often influenced by not just the traits such as the competition rate but by the increase in replication rates, which in our case refer to the spread of information in the network or community. We therefore expect that for correction messages to eliminate falsehoods on Twitter in the long term, they will have to gain an ecological advantage by having a higher competition rate and a higher replication rate than falsehoods.

Influence of Falsehoods on Correction Messages and Vice Versa

As suggested by previous studies, falsehood tends to be resilient (Friggeri et al. 2014) and the format and presentation of the message may have an impact on the believability of the message (Kim and Dennis 2017). While most studies have centered on correction messages as a general topic, none of the studies has investigated this phenomenon using real datasets nor has any attempted to infer causality. We observe that the diffusion of correction messages and falsehoods may show unique dynamic feedback behavior. That is, their effects on each other may be bidirectional and reciprocated where both influence each other. While several studies have shown that correction messages can be very effective in addressing misinformation on social media by reducing the credibility of the refuted content and as such are shared more (Chua et al. 2017; Huang 2017), other studies have also pointed out that even after the rebuttals, falsehoods continually influences memory and reasoning even if the retraction is recalled (Ecker et al. 2010). The study further noted that even after the retraction of false information and specific warnings were combined with an explanation for the misinformation, users still remembered and were influenced significantly by falsehoods. One study further noted that in some cases even though corrections reduce beliefs in the misinformation, the trust may be lost (Huang 2017).

Ozkurk et al. (2015) argued that when combating falsehood, presenting warning messages may lower its propagation and thus improve the quality of information that is being shared on online social networks. This analogy does not state that the news ultimately dies but that it counterbalances the information flow. We therefore believe that once the correction message is introduced and people see the message, some may refrain from resharing the false messages while others may rather engage the correction message. This will slow the rate of the diffusion of falsehoods but increase the correction messages over time.

Theoretical Framework

Using the competitive exclusion principle (Gause 1932) from ecology, we develop a deterministic model consisting of a system of ordinary differential equations and implemented it using numerical simulations. The model depicts the competition of two species - falsehoods and correction messages - with natural death and reflects the competitive exclusion of the species. Several studies have described the behavior of information and its diffusion using epidemiology models in the understanding of falsehoods (Cheng et al. 2013; Jin et al. 2013). However, these studies assume that rumors and falsehoods are diseases and are mutually exclusive. In contrast, we model the behavior and interactions between two different species, competing for the same limited resources in an environment or niche, in this case Twitter.

Using μ_i as the rate of decay (mortality) of message (species) i ($i = 1{,}2$), K as the number of users, $N_i(t)$ as the number of tweets and retweets at time t with fraction $n_i(t) = \frac{N_i(t)}{K}$, r_i as the per-capita growth rate of message (species) i ($i = 1{,}2$), and $\beta_{i,j}$ as the competition rate of specious j to i ($i\ , j = 1{,}2;\ i\ \neq j$), we have:

1)

$$\frac{d\,N_1}{dt} = r_1 N_1 \left[(1 - \frac{N_1}{K}) - \beta_{12} \frac{N_2}{K}\ \ \right] - \mu_1 N_1$$

$$\frac{d\,N_2}{dt} = r_2 N_2 \left[(1 - \frac{N_2}{K}) - \beta_{21} \frac{N_1}{K}\ \ \right] - \mu_2 N_2$$

Thus,

$$\frac{d\,n_1(t)}{dt} = r_1 n_1(t) \left[(1 - \frac{\mu_1}{r_1}) - n_1(t) - \beta_{12}\, n_2(t) \right] \tag{1}$$

$$\frac{d\,n_2(t)}{dt} = r_2 n_2(t) \left[(1 - \frac{\mu_2}{r_2}) - n_2(t) - \beta_{21}\, n_1(t) \right] \tag{2}$$

Next, we will perform a stability analysis of the system of equations (1-2).

Stability Analysis

At stability of the fraction of the two species $\frac{d\, n_i(t)}{dt}$ =0. Thus, that ensuing system of algebraic equations has the following four possible solutions (that are called equilibria):

1) n_1 = 0, and n_2 = 0;

2) n_1 = 0, and n_2 = $(1-\frac{\mu_2}{r_2}) = \ R_2$, which exists only if $\mu_2 < r_2$;

3) n_1 = $(1-\frac{\mu_1}{r_1}) = \ R_1$, and n_2 = 0, which exists only if μ_1 < r_1; or

4) $n_1 = \frac{\left(1-\frac{\mu_1}{r_1}\right)-\beta_{12}\left(1-\frac{\mu_2}{r_2}\right)}{1-\beta_{12}\,\beta_{21}} = \frac{R_1-\beta_{12}R_2}{1-\beta_{12}\,\beta_{21}}$, and $n_2 = \frac{\left(1-\frac{\mu_2}{r_2}\right)-\beta_{21}\left(1-\frac{\mu_1}{r_1}\right)}{1-\beta_{21}\,\beta_{12}} = \frac{R_2-\beta_{21}R_1}{1-\beta_{21}\,\beta_{12}}$, which exists only when $0\ \le\ n_1, n_2\ \le\ 1$.

To study the local stability of the equilibria of the system, let

$$f_1(r_1, R_1, \beta_{12}) = r_1 n_1[R_1 - n_1 - \beta_{12}\, n_2] \tag{3}$$

$$f_2(r_2, R_2, \beta_{21}) = r_2 n_2[R_2 - n_2 - \beta_{21}\, n_1] \tag{4}$$

Then the Jacobian matrix of the system is given by:

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial n_1} & \frac{\partial f_1}{\partial n_2} \\ \frac{\partial f_2}{\partial n_1} & \frac{\partial f_2}{\partial n_2} \end{bmatrix} = \begin{bmatrix} r_1[R_1 - 2n_1 - \beta_{12}\, n_2] & -r_1\beta_{12}\, n_1 \\ -r_2\beta_{21}\, n_2 & r_2[R_2 - 2n_2 - \beta_{21}\, n_1] \end{bmatrix}.$$

The first equilibrium n_1 = 0, and n_2 = 0 is always not stable since the Jacobian

$$J_1 = \begin{bmatrix} r_1R_1 & 0 \\ 0 & r_2R_2 \end{bmatrix},$$

which is only stable if the eigenvalues of J_1 are negative.

That is, when $\mu_1 > r_1$ and $\mu_2 > r_2$ or when the two messages die faster than they grow.

For the second equilibrium, n_1 = 0, and $n_2 = \ R_2 > 0$

$$J_2 = \begin{bmatrix} r_1[R_1 - \beta_{12}\, R_2] & 0 \\ -r_2\beta_{21}\, R_2 & -r_2R_2 \end{bmatrix},$$

which is stable only if $\frac{R_1}{R_2} < \beta_{12}$ and is when the rebuttal message is aggressive enough to put an end to the falsehood. The third equilibrium, $n_1 = R_1$, and $n_2 = 0$ with a Jacobian J_3 is similarly stable when $\frac{R_2}{R_1} < \beta_{21}$ as the falsehood is stronger than that threshold of $\frac{R_2}{R_1}$. These two equilibria are the competitive exclusion cases in which one of the two message puts an end to the other.

The last equilibrium is the co-existence of the two messages when $n_1 = \frac{R_1 - \beta_{12} R_2}{1 - \beta_{12}\,\beta_{21}}$, and $n_2 = \frac{R_2 - \beta_{21} R_1}{1 - \beta_{21}\,\beta_{12}}$. It is stable when

$$J_4 = \begin{bmatrix} r_1\left[R_1 - 2\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}} - \beta_{12}\,\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}}\right] & -r_1\beta_{12}\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}} \\ -r_2\beta_{21}\,\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}} & r_2\left[R_2 - 2\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}} - \beta_{21}\,\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}}\right] \end{bmatrix}$$

In this model, we find a scenario where both falsehoods and correction messages either eventually die off or survive after competition.

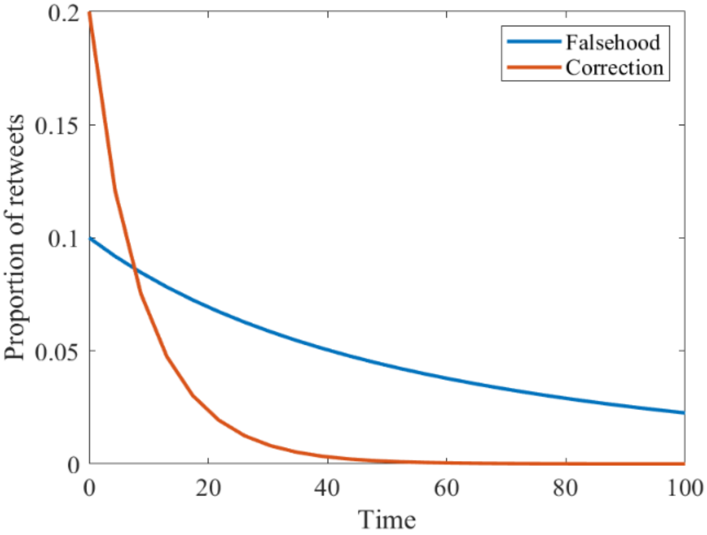


Figure 1. Both falsehood and correction messages die off.

Their Initial values are n_1 (0) =.1, and n_2 (0) = .2 r1 =.1, r2=.1, μ1=.11, μ2=.2.

In Figure 1 we simulated the effects of both falsehoods and real news as two species. Using Equation 1 we show that eventually both falsehoods and correction messages die off. This usually happens when n_1 = 0, and n_2 = 0; A typical example would be that the weaker specie dies first and the much stronger dies a bit later. However, eventually the number of tweets and retweets for both get to and remain at zero.

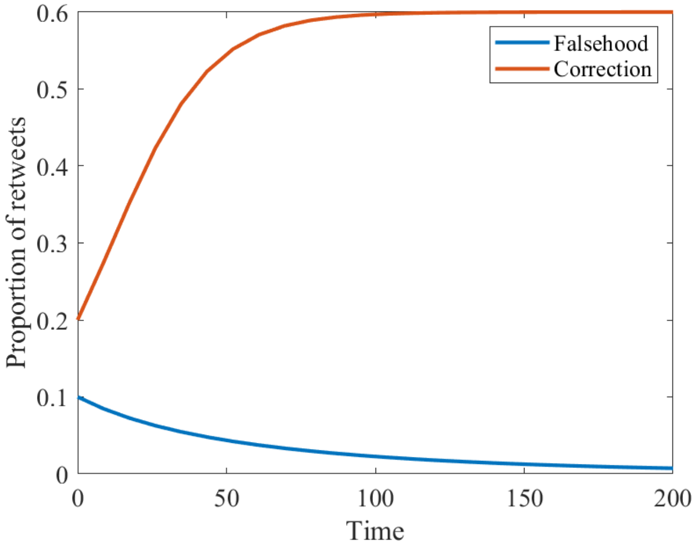


Figure 2. Either falsehoods or correction messages survive .

Their initial values are n_1 (0) = .1, and n_2 (0) = .2, r1 =.15, r2=.2, µ1=.08, µ2=.06). Fig 2 shows the effects when $n_1 = 0$, $n_2 = R_2 > 0$. At this point, tweets containing falsehood die off while correction tweets increase by making $R_2 > R_{1/}\beta_2$). This scenario (Equation 2) is our ideal scenario in the fight against misinformation. We can reduce and eliminate falsehoods while ensuring that correction and real messages survive. In this simulation we show that correction messages eliminate falsehoods while increasing in diffusion. This model can be interchanged to ensure that the opposite happens; falsehoods survive by eliminating correction messages. This can be done by flipping the r_1 and r_2 values as in Equation 3.

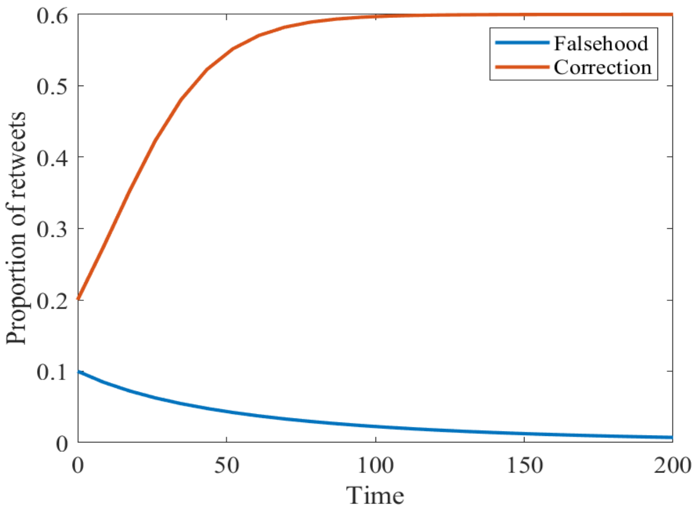


Figure 3. Both falsehood and correction messages survive.

Their Initial values are n_1 (0) =.1 and n_2 (0) = .2, r1 =.15, r2 = .2, µ1=.08, µ2 = .06)**.** In our simulation of Equation 4, in figure two, we show that both correction messages can be made to survive as two species vying for similar resources which are messages in the ecosystem (Twitter). This occurs where $0 \leq n_1, n_2 \leq 1$. An evolutional shift sometimes occurs where the weaker message might for instance mutate and opt for a different niche in the presence of the much stronger species.

Empirical Evidence

Empirical Method

Based on our theoretical simulation results, we next verify which of them hold in the real world. Specifically, we use time series and panel vector autoregression (VAR) (Abrigo and Love 2015; Lütkepohl 2005) models to characterize the dynamic relationships between falsehoods and correction messages on Twitter at the hourly level including the differences in individual- and aggregate-level messages during two shock events Hurricanes Harvey and Florence. VAR is a stochastic process model commonly used in economics to capture the linear relationships among multiple time series variables (Abrigo and Love 2015; Killins et al. 2017). It has a more robust framework that can be easily verified and replicated and addresses biases such as autocorrelations, endogeneity and causal inferences (Luo et al. 2013). In IS research, it has been applied to examine the relationships between sentiments from microblogs and stock returns (Deng et al. 2018) and the existence of several patterns of supply-side technology relationships in the context of wireless networking (Adomavicius et al. 2012). In our research, VAR allows us to model falsehood and correction message tweets as two time series and examine how the diffusion of falsehoods and correction messages unfold as a result of falsehood and correction message tweets in the past. As a result, we can treat both falsehoods and correction message tweets as endogenous and predict them using their lagged values. This enables us to capture the feedback loops among falsehoods and correction messages (Abrigo and Love 2015). For example, falsehoods in the current period may influence the number of correction messages in the next period, which may cause a change in the next period's falsehoods.

Data

Our data consists of verified false tweets and their correction tweets during Hurricane Harvey in 2017 and Hurricane Florence in 2018. We selected these two extreme events because studies have shown that misinformation are propagated the most "during events that have importance in the lives of individuals and when the news received about them is either lacking or subjectively ambiguous" (Allport and Postman 1946). Hurricanes match these descriptions because they create public safety concerns and are quite unpredictable. Furthermore, hurricanes allow researchers to accurately record exactly when the rumor associated with the shock event begins and ends. As a result, we can measure the beginning of the shock event (hurricanes) and its ending, which cannot be said for other crisis scenarios. For example, it is not easy to predict when an earthquake or other natural disasters occur unlike hurricanes. Hence, when we capture data after a shock event occurs, it may lead to not only data with missing information but also endogeneity.

Our data collection steps were as follows. First, we identified and collected all tweets during Hurricane Harvey from August to September 2017 and Hurricane Florence from September to October 2018 using the Twitter API. This resulted in 12,357,530 tweets and retweets during the two extreme events. Next, we identified fake news stories from both hurricanes from FEMA's rumor control page and their Twitter timelines within those periods. We retained only verified tweets, false and correction messages whose veracity could be verified using information from FEMA's website and Twitter handle and matched them to five independent fact checking websites, Snopes, Factcheck.org, PolitiFact, Truth or Fiction and BS Detector. This was done by crawling and extracting the keywords of both false and corrected news stories from the last section websites' URL. For example, we extracted "Harvey relief donation rumors" from https://www.snopes.com/harvey-relief-donation-rumors**.** Due to the difficulties in analyzing Twitter textual contents, we utilized Latent Dirichlet allocation (LDA) to filter the data based on topics and used term stemming to break each tweet word down to its root and converted them into document term matrices that had tweets as rows and words as columns. This generative modeling technique exposes and identifies underlying topics in text documents and their similarities between them (Blei 2003). Finally, we removed sparse terms that occurred less than 5% of the time. As a result, we obtained 279,597 tweets and retweets of verified falsehoods and their correction messages during the two hurricanes based on 21 false and rebutted topics.

In order to obtain the diffusion count data, we aggregated the tweets and retweets by hour for each falsehood topic and each correction topic over the duration of the data collection period and obtained a minimum of 22 and a maximum of 648 hours of tweets for every falsehood or correction topic. This ultimately resulted in 6,343 hourly observations for 21 topics.

Table 5 summarizes the descriptive statistics of the counts of the tweets and retweets by veracity and topic. The correlation between the hourly tweet and retweet counts of falsehoods and correction messages is 0.41.

Table 5. Descriptive statistics of the diffusion of falsehoods and correction messages.

Topic	**Count of False Tweets**			**Count of Correction Tweets**			**No. of Hours**
	Obs.	**Mean**	**Std. Dev.**	**Obs.**	**Mean**	**Std. Dev.**	
1	10	0.08	0.4	9,651	72.56	140.9	133
2	10,130	25.33	59.41	500	1.25	6.2	400
3	793	8.62	13.61	45	0.49	2.89	92
4	6	0.02	0.15	755	2.71	14.57	279
5	8,433	32.31	68.57	5	0.02	0.14	261
6	2,787	6.88	7.9	1,562	3.86	12.74	405
7	42,197	126.34	231.72	40,182	120.31	439.22	334
8	10	0.04	0.22	357	1.47	4.24	243
9	33	0.41	0.92	141	1.74	3.38	81
10	13,559	39.53	55.32	694	2.02	4.72	343
11	12,043	35.84	128.32	45	0.13	0.5	336
12	4,824	8.82	45.79	191	0.35	4.44	547
13	106,748	207.68	113.26	65	0.13	0.72	514
14	31	0.1	0.47	10	0.03	0.24	314
15	3	0.14	0.47	25	1.14	4.89	22
16	716	3.2	9.8	248	1.11	3.55	224
17	87	0.34	0.92	72	0.28	1.15	254
18	195	1.08	2.52	5	0.03	0.2	181
19	14	0.04	0.24	127	0.34	1.22	369
20	8,461	23.31	126.18	13,755	37.89	132.98	363
21	80	0.12	0.85	2	0.003	0.06	648
Total	**211,160**	**33.29**	**99.56**	**68,437**	**10.79**	**111.44**	**6,343**

One example of falsehood circulated on Twitter during both hurricanes was the perceived discovery of sharks swimming on the streets. Figure 4 below shows the hourly tweet and retweet counts of both falsehoods and correction messages for these shark stories during each hurricane. As we can see from both diagrams, falsehoods first started, then the number of correction messages increased. Correction messages seem to be more pronounced during the first 100 hours, possibly in reaction to falsehoods. As time went by, the tweet and retweet count of both types of messages decreased.

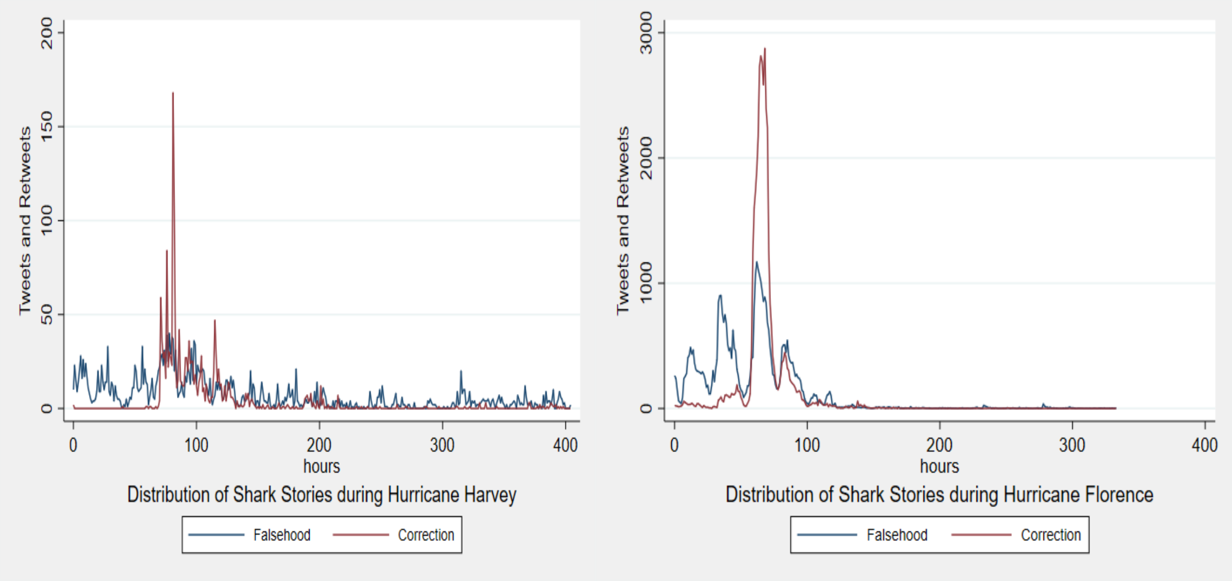


Figure 4. Distributions of falsehoods and correction messages of shark stories during both Hurricanes (Harvey in 2017 and Hurricane Florence in 2018).

Model Specification

Building on a structural VAR method, we analyzed the dynamic effects of the feedback loop between false tweets and their correction messages during Hurricanes Harvey and Florence. We estimated the diffusion of both types of messages based on their past diffusion histories. Due to the non-stationarity nature of our sample and excessive zeros, we followed previous approach of (Adomavicius et al. 2012) and transformed our data by taking natural log plus 0.5 of the variables. We performed both the Dickey Fuller and the Fisher type unit root tests for nonstrongly balanced datasets to test for unit root. Furthermore, we controlled for the year in our models when necessary and the hour of day to eliminate the impact of timing on the hourly counts.

Based on (Abrigo and Love 2016), we modeled the diffusion of our two time series of falsehoods and correction messages $\mathbf{Y}_{it}$ as a (2 x 1) vector of dependent variables including the number of falsehoods diffused and the number of correction messages on topic $i \in \{1, 2, ...,21\}$ during hour $t \in \{1, 2, ..., Ti\}$. $\boldsymbol{\gamma}_{it}$ is a (2 x 1) vector with dummy variables representing the year or hurricane and $\boldsymbol{h}_{it}$ is a (2 x 23) matrix of dummy variates representing the hour of the day, $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_{it}$ are (2 x 1) vectors of dependent variable-specific panel fixed effect and idiosyncratic errors, respectively. The (2x2) matrices $\mathbf{A}_1$, $\mathbf{A}_2$, . . . , $\mathbf{A}_{P-1}$, $\mathbf{A}_P$ are the parameters to be estimated as recommended by (Abrigo and Love 2015), where P is the lag order (e.g. 1, 2, …) to be estimated empirically. The assumptions are that the innovations may be denoted by: $\mathbf{E}(\mathbf{e}_{it}) = \mathbf{0}$, $\mathbf{E}(\mathbf{\acute{e}}_{it}\mathbf{e}_{it}) = \boldsymbol{\Sigma}$, and $\mathbf{E}(\mathbf{\acute{e}}_{it}\mathbf{e}_{is}) = \mathbf{0}$ for all $t > s$, which is a white noise multivariate process of our two variables. Following studies by (Holtz-Eakin et al. 1988), we assume that the cross-sectional units share the same underlying data generating process, with the reduced-form parameters $\mathbf{A}_1$, $\mathbf{A}_2$, . . . , $\mathbf{A}_{p-1}$, $\mathbf{A}_p$ common among them. The structural representation can be given as:

$$\mathbf{Y}_{it} = \mathbf{Y}_{it-1}\mathbf{A}_1 + \mathbf{Y}_{it-2}\mathbf{A}_2 + \ldots + \mathbf{Y}_{it-P+1}\mathbf{A}_{P-1} + \mathbf{Y}_{it-P}\mathbf{A}_P + \boldsymbol{\gamma}_{it}\boldsymbol{\beta}_1 + \boldsymbol{h}_{it}\boldsymbol{\beta}_2 + \mathbf{u}_i + \boldsymbol{\varepsilon}_{it}. \quad (5)$$

We test the relationships between the number of falsehood tweets and correction tweets at the hourly level. We rewrite Equation 5 into Equation 6 where Falsehood and Correction denote the hourly count of falsehood tweets and the hourly count of correction tweets on a topic, respectively. C_{1i} and C_{2i} are the intercepts, $A_{i,j}^{t-p}$ is the effect of endogenous variable j on endogenous variable i, and $\gamma_{1\mathrm{it}}$ and $\gamma_{21\mathrm{t}}$ represent the impacts of Year 2018 or Hurricane Florence on the falsehood and correction tweet counts, respectively. $h_{1,1it}$ to $h_{1,23it}$ represent the impacts of the hour of the day on the falsehood tweet count, and $h_{2,1it}$ to $h_{2,23it}$ represent the impacts of the hour of the day on the correction tweet count. For any given hourly count of either falsehoods or correction tweets, at the most only one value among $h_{1,1it}$ to $h_{1,23it}$ and one value among $h_{2,1it}$ to $h_{2,23it}$ is non-zero, representing the impacts of the time of the day on that particular hour's tweet counts. ε_{1it} and ε_{2it} denote the error terms for false tweets and correction tweets for topic *i* during hour *t*, which means they are serially uncorrelated of zero and finite variance. The time period is from 0 hours to 313 hours. We present the derivation of the VAR model based on the theoretical model in Appendix A. Our bivariate model is specified in Equation 6 below.

$$\begin{bmatrix} Falsehood_{it} \\ Correction_{it} \end{bmatrix} = \begin{bmatrix} C_{1i} \\ C_{2i} \end{bmatrix} + \sum_{p=1}^{P} \begin{bmatrix} A_{1,1}^{t-p} & A_{1,2}^{t-p} \\ A_{2,1}^{t-p} & A_{2,2}^{t-p} \end{bmatrix} \begin{bmatrix} Falsehood_{i,t-p} \\ Correction_{i,t-p} \end{bmatrix} + \begin{bmatrix} \gamma_{1it} \\ \gamma_{2it} \end{bmatrix} + \sum_{q=1}^{23} \begin{bmatrix} h_{1,qit} \\ h_{2,qit} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1it} \\ \varepsilon_{2it} \end{bmatrix} \quad (6)$$

Results

Summary of panel VAR model results

The panel VAR models were estimated using the generalized method of moments (GMM) estimators (Abrigo and Love 2015). We used the natural log of falsehood tweet count plus 0.5 and natural log of correction tweet count plus 0.5 as the dependent variables instead of the raw counts. We tested five different models: Model 1 with the full sample including both Hurricanes Harvey and Florence data; Model 2 with only Hurricane Harvey data; Model 3 with only Hurricane Florence data; Model 4 with only tweets and retweets on environmental news; and Model 5 with only tweets and retweets on social news. The first three models allow us to compare and contrast the results between the two hurricanes, and the last two models allow us to examine if the results are consistent for tweets in different categories.

We selected the optimal lag for each model and the moment condition based on the Bayesian information criteria (BIC) as recommended by (Andrews and Lu 2001). The Bayesian information criteria (BIC), Akaike information criteria (AIC) and the Hannan–Quinn information criteria (MQIC) are amoung the most commly used maximum likelihood selection criterias for models (Abrigo and Love 2015; Love and Zicchino 2006). Table 6 reports the results of the model lag selection. All lags were positive and significant and consistent with the Granger causality test results (p< 0.01). For Models 1, 2 and 3, a lag of three had the best performance, whereas for Model 4 and 5 a lag of two had the best performance. To improve the efficiency of our model, we included longer sets of lags as instruments for all models as recommended by (Abrigo and Love 2016). However, this approach has the unattractive property of reducing observations especially with unbalanced panels and those with missing data because past realizations are not included. A proposed solution recommended by (Holtz-Eakin et al. 1988) is to use GMM in estimation which substitutes missing observations with zero but cannot solve the reduced observation problems associated with unbalanced panels. For all five models we used the first to fourth lag as instruments to improve the efficiency of our model.

We next performed our panel VAR estimates using the robust standard errors. The results are summarized in Tables 7 and 8. All Granger causality Wald test for Models 1, 3, 4 and 5 were consistently positive and significant at either the p<0.01 or p<0.05 level, indicating positive and significant bidirectional effects of the hourly counts of falsehood and correction tweets on each other in the next hour. In Model 2, the effect of falsehoods on corrections was significant at p<0.05, but the effect of corrections on falsehoods was not significant. The eigenvalues all lied inside the unit circle, indicating that the models were stable. The impulse response function (IRF) plots showing the 95% confidence intervals of the coefficient estimates based on 200

Monte Carlo simulations using Gaussian approximation also supported the results and were consistent across all models.

Table 6. Model lag selection results.

Criteria	Lag	Model 1 (Harvey and Florence)	Model 2 (Florence Only)	Model 3 (Harvey Only)	Model 4 (Environmental Tweets Only)	Model 5 (Social Tweets Only)
MBIC	1	400.718	174.768	130.164	6.203	-19.581
	2	105.657	35.410	10.040	-30.155	-37.232
	3	17.815	-2.071	-5.293	-18.135	-17.260
MAIC	1	481.066	249.856	198.078	60.237	26.044
	2	159.222	85.469	55.317	5.867	-6.815
	3	44.598	22.959	17.346	-0.124	-2.052
MQIC	1	453.159	223.190	173.217	39.303	7.847
	2	140.617	67.691	38.742	-8.088	-18.947
	3	35.296	14.070	9.058	-7.102	-8.118

We next performed our VAR estimates using the robust standard errors. The results are summarized in Tables 7 and 8. All Granger causality Wild test results were positive and consistent at p<0.001 in the same direction, indicating positive and significant bidirectional effects of the hourly counts of falsehood and correction tweets on each other in the next hour. The eigenvalues all lied inside the unit circle, indicating that the models were stable. The impulse response function (IRF) plots showing the 95% confidence intervals of the coefficient estimates based on 200 Monte Carlo simulations using Gaussian approximation also supported the results and were consistent across all panels.

Table 7. Bidirectional effects of hourly counts of falsehood and correction tweets.

Model	Lag	Falsehoods on falsehoods	Falsehoods on Corrections	Corrections on Falsehoods	Corrections on Corrections
	1	0.504*** (0.021)	0.027** (0.015)	0.042*** (0.016)	0.538*** (0.027)

Model 1 (Harvey and Florence, N_1=5,997)	2	0.243*** (0.022)	0.032** (0.017)	0.001 (0.014)	0.218*** (0.024)
	3	0.169*** (0.019)	-0.018 (0.015)	0.008 (0.016)	0.153*** (0.021)
Model 2 (Florence Only, N_2=3,870)	1	0.491*** (0.028)	0.042** (.016)	0.020 (0.023)	0.534*** 0.034
	2	0.211*** (0.028)	0.007 (0.018)	0.010 (0.020)	0.232*** 0.030
	3	0.169*** (0.024)	-0.020 (0.017)	0.003 (0.022)	0.131*** (0.026)
Model 3 (Harvey Only, N_3=2,127)	1	0.494*** (0.036)	0.004 (0.032)	0.070*** (0.021)	0.527*** (0.039)
	2	0.288*** (0.035)	0.078** (0.034)	-0.021 (0.021)	0.205*** (0.039)
	3	0.155*** (0.031)	-0.032 (0. 028)	0.015 (0.020)	0.169*** (0.033)
Model 4 (Environmental Tweets Only, N_4=671)	1	0.305*** (0.040)	0.025 (0.040)	-0.055* (0.028)	0.265*** (0.044)
	2	0.624*** (0.040)	0.037 (0.041)	0.140*** (0.030)	0.630*** (0.047)
Model 5 (Social Tweets Only, N_5=331)	1	0.378*** (0.053)	-0.074 (0.107)	0.002 (0.018)	-0.102 *** (0.146)
	2	0.573*** (0.057)	0.318** (0.126)	0.041** (0.016)	0.287** (0.112)

Notes: * p<0.1; ** p<0.05; ***p<0.01. Standard deviations in parentheses.

Table 8. Granger causality Wald test results and eigenvalues for stability.

Model	DF	χ^2 (Falsehoods on Corrections)	χ2 (Corrections on Falsehoods)	Eigenvalue (Falsehoods on Corrections)	Eigenvalue (Corrections on Falsehoods)
1	3	15.287***	16.889***	0.973 (Real) 0 (Imaginary)	0.917 (Real) 0 (Imaginary)
2	3	8.520**	1.943	0.947(Real) 0 (Imaginary)	0.908(Real) 0 (Imaginary)
3	3	11.089**	20.323***	0.985 (Real) 0 (Imaginary)	0.914(Real) 0 (Imaginary)
4	2	5.789***	24.339***	0.991 (Real) 0 (Imaginary)	0.875 (Real) 0 (Imaginary)
5	2	6.578**	10.020***	0.974 (Real) 0 (Imaginary)	0.397 (Real) 0 (Imaginary)

Notes: * p<0.1; ** p<0.05; ***p<0.01.

We also examined the forecast error variance decomposition (FEVD) of the effects modeled in the panel VAR (Table 9). It gives the unexpected variation in each variable that is produced by the shocks from other variables. The component measures the fraction in a variable explained by variations in the other variable. It also indicates the relative impact that one has on the other.

Table 9. Falsehoods-Corrections Forecast Error Variance Decomposition

RV (False)	**Impulse Variables**									
	Model 1		**Model 2**		**Model 3**		**Model 4**		**Model 5**	
Hour	False	Corr	False	Corr	False	Corr	False	Corr	False	Corr
0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	1	0	1	0
2	.999	.001	.999	.001	1	0	.999	.001	.991	.009
3	.998	.003	.997	.003	997	.003	.998	.003	.991	.009
24	.957	.043	.978	.023	.953	.047	.916	.084	.989	.011
∞	.916	.084	.970	.030	.895	.105	.842	.158	.989	.011
RV (Corr)	**Impulse Variables**									
Hour	False	Corr	False	Corr	False	Corr	False	Corr	False	Corr
0	0	0	0	0	0	0	0	0	0	0
1	.003	.997	0	1	.012	.988	.026	.974	.003	.997
2	.007	.993	.001	.999	.024	.976	.065	.935	.019	.991
3	.009	.991	.003	.998	.028	.972	.079	.921	.028	.972
24	.113	.887	.031	.970	.234	.766	.415	.585	.135	.865
∞	.203	.798	.041	.959	.451	.550	.623	.377	.180	.820

Notes: Falsehoods = False, Corrections = Corr, Response Variable =RV.

We report the results for each model in detail next.

Results for Model 1: Hurricane Harvey and Hurricane Florence

In Model 1, we examine the relationships between falsehoods and correction messages during Hurricanes Harvey and Florence. The panel consisted of falsehoods and correction messages on 21 topics. We included a year dummy to control for the differences due to year or hurricane and 23 hourly dummies to control the impact of the time of day on the hourly counts. The significant causality test results show that both falsehoods and correction messages Granger caused each other.

We plot the orthogonal IRF plots in Figure 4 to analyze how the hourly count of falsehoods or correction messages responded to a one standard deviation shock of the other, with all other effects held constant (Abrigo and Love 2015; Love and Zicchino 2006). The IRF Plot 4B shows that a one standard deviation shock or exogenous increase in the hourly count of correction tweets caused the hourly count of falsehoods to increase for over 20 hours. We also observe a sharp undulation within the first three hours. That is, there was a sharp increase (.04) in the first 2 hours and subsequently a decline (0.02) in increase in falsehoods in the third hour. A smoother but smaller increase continued after the third hour till the 20th hour, after which it gradually went back to zero. This means that, contrary to what FEMA and fact checking organizations intended, increasing the awareness that a news story was false increased the spread of the falsehood initially. It also shows that at some point, after the second hour, there was a decline in the increase in falsehoods for one hour. Similar effects existed for falsehoods on corrections, where a one standard deviation increase in the hourly count of falsehoods caused more corrections for the next 20 hours (Plot 4C). This interaction also saw a very small decline in hour two and then continued its positive trajectory for 18 hours. It should be noted that the response of falsehoods to corrections was constrained to zero in our initial period because of the ordering of our endogenous variables. In addition, the IRF plots show that falsehoods and corrections both caused significant increases of their own counts for about 30 hours (Plots 4A and 4D). These plots also show a much sharper wave during the first two hours before the effects gradually went back to zero. That is, in general, more falsehoods led to more falsehoods and more corrections led to more corrections, but the effects were more pronounced within the first two to three hours.

We examined the FEVD (Table 9) for 24 hours to account for potential changes throughout the day. For Model 1, corrections accounted for up to 4.3% of the forecast error in predicting falsehoods, while falsehoods accounted for up to 11% of the forecast error of corrections in the first 24 hours. Both steadily increased over time and were greater than zero based on simulated results. The effects of falsehoods on corrections were stronger than that of corrections on falsehoods. Though the effect were relatively small, they were significant and indicate important impacts of one on the other (Luo and Zhang 2013).

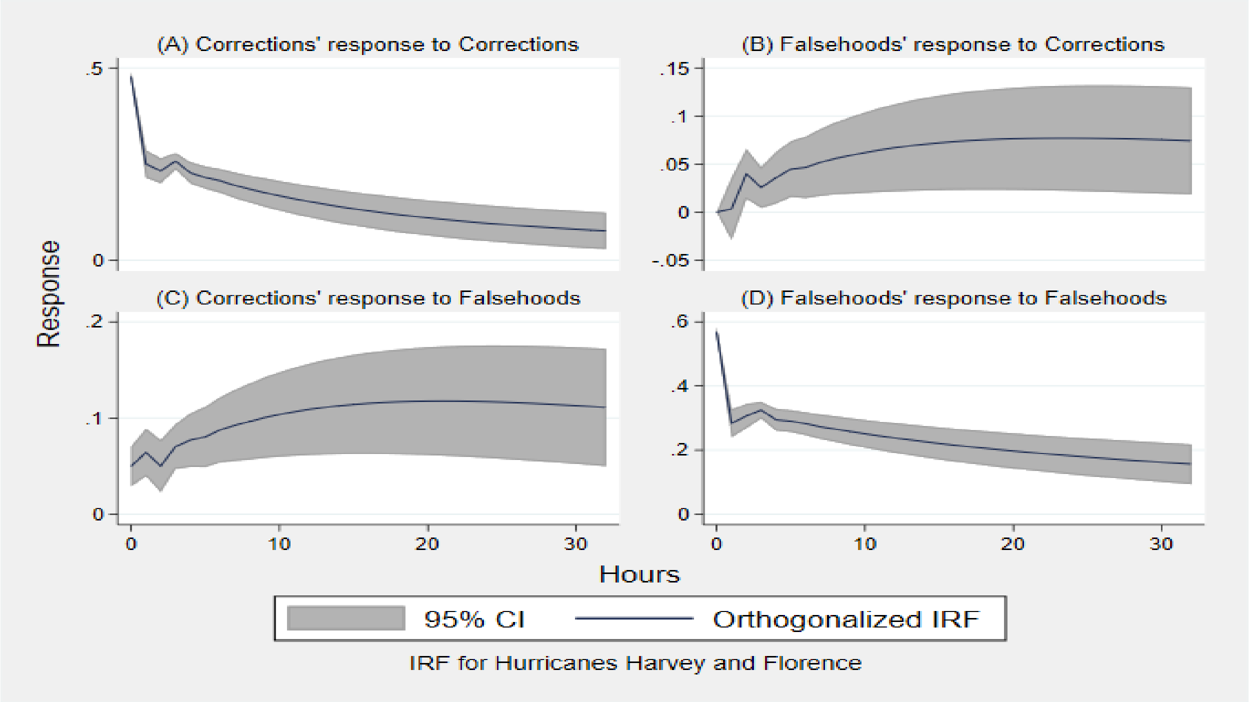


Figure 5. IRF plots of falsehoods' response to corrections and corrections' response to falsehoods for Hurricanes Harvey and Florence.

Results for Model 2: Hurricane Harvey Panel Analysis

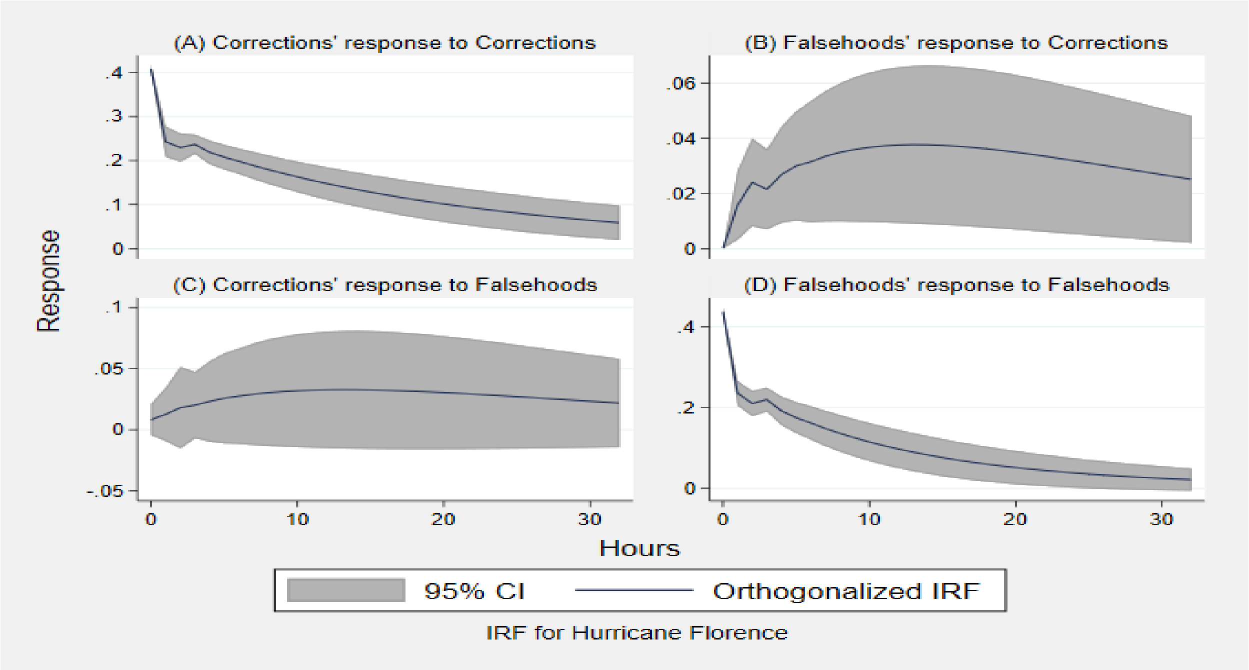


Figure 6. IRF plots of falsehoods' response to corrections and corrections' response to falsehoods for Hurricane Harvey.

The results of Model 2 for Hurricane Florence (Figure 5) are generally similar to those of Model 1. There was a positive impact of the hourly counts of falsehoods on corrections. However, the impact of the hourly count of corrections on falsehoods was nonsignificant because the confidence interval included the zero line (Plot C). This matched the Granger causality test result which showed that falsehoods did not cause a response in corrections. The FEVD results show that falsehoods accounted for up to 3.1% of the forecast error of corrections after the first 24 hours, which is greater than zero based on simulated results.

Results for Model 3: Hurricane Florence Panel Analysis

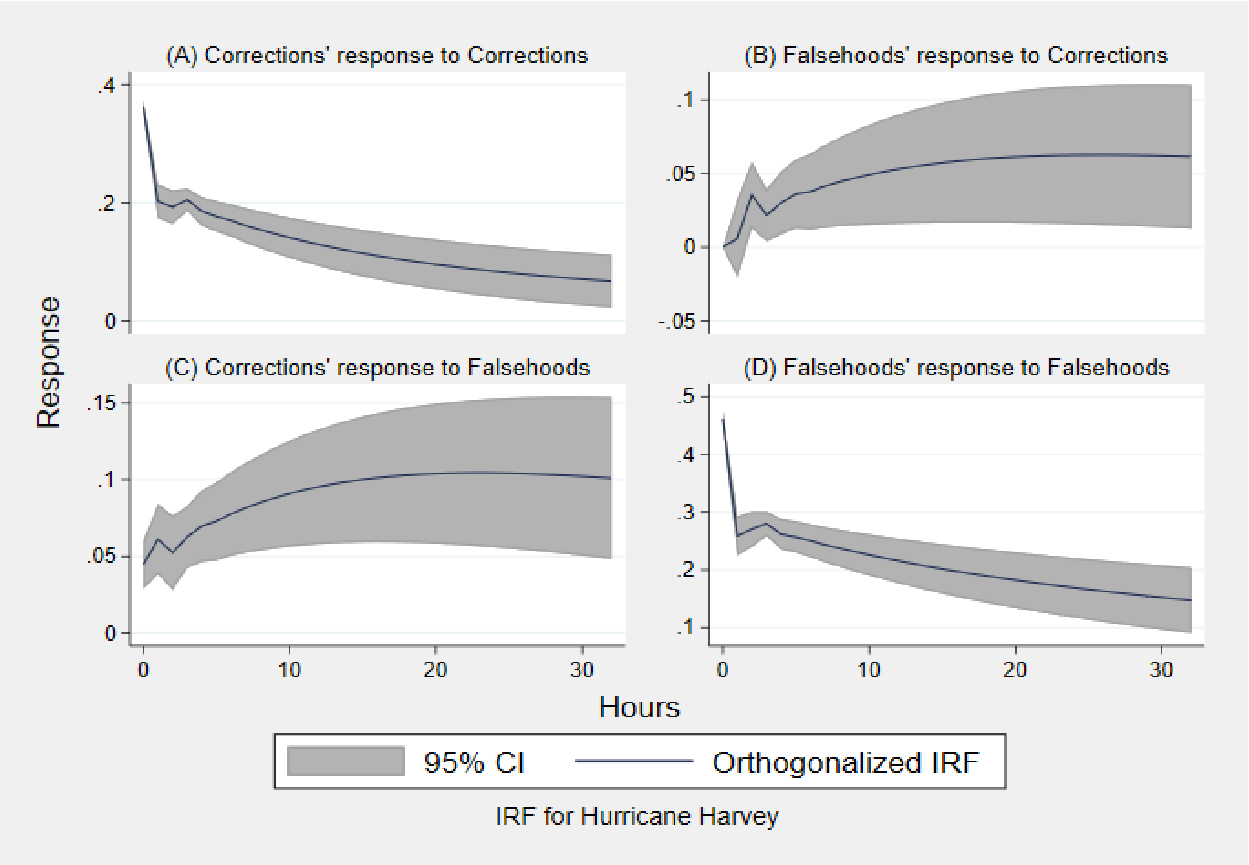


Figure 7. IRF plots of falsehoods' response to corrections and corrections' response to falsehoods for Hurricane Florence.

We obtained similar results in Model 3 for Hurricane Harvey (Figure 6) as those in Models 1 and 2. It shows positive relationships between falsehoods on corrections. That is, more subsequent falsehood or correction tweets as a result of an increase in the hourly count of the other. The initial effect of corrections on falsehoods was not significant at the zero-hour mark but became significant after the one-hour mark. The FEVD results show that corrections accounted for up to 4.7% of the

forecast error of falsehoods, while falsehoods accounted for up to 23.4% of the forecast error of corrections at the end of the 24-hour period. They were both greater than zero based on simulated results. The FEVD was stronger for falsehoods than for corrections.

Results for Model 4: Environmental Tweets Panel Analysis

In In addition to separate analyses by hurricane, we also investigated the effects of both dynamic loops for different types of message such as environmental news and social news during the two hurricanes.

The results of Model 4 (Figure 7) are consistent with previous results in Models 1, 2 and 3. There were positive relationships between the counts of falsehoods and corrections. That is, more falsehoods (or corrections) resulted in an increase in correction (or falsehood) tweets among environmental news in subsequent hours. Plot 7B also shows that the effect of corrections on falsehoods was not significant until after the 1-hour mark. The FEVD results show that corrections accounted for up to 8.4% of the forecast error of falsehoods, whereas falsehoods accounted for up to 41.5% of the forecast error of corrections in Model 4. Both were significantly greater than zero based on simulated results. The FEVD was stronger for falsehoods than for corrections.

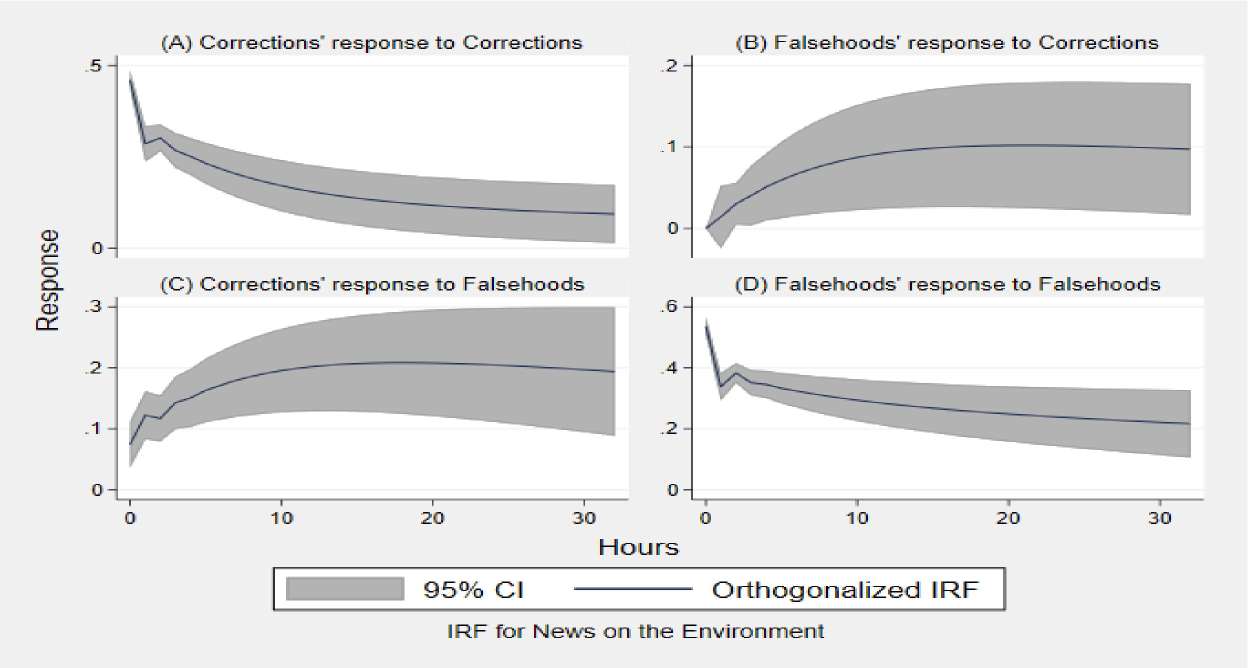


Figure 8. IRF plots for environmental false and correction tweets during both hurricanes.

The results from Model 5 (Figure 8) show a brief significant impact of corrections on falsehoods in the social news category around the 2-hour mark and after that the impact became nonsignificant because the confidence interval included the zero line (Plot 8B). In addition, more falsehoods and correction tweets led to more of their own messages in subsequent hours but was not significant for corrections at the zero hour. The effect of falsehood tweets on corrections in the social news category was not significant until after the 2-hour mark (Plot 8C). The FEVD results show that in the first 24 hours corrections accounted for up to 1.1% of the forecast error of falsehoods, and falsehoods accounted for up to 13.5% of the forecast error of corrections in Model 5 for social news. Both were significantly greater than zero based on simulated results. The FEVD was much stronger for falsehoods than for corrections for social news.

Results for Model 5. Social Tweets Panel Analysis

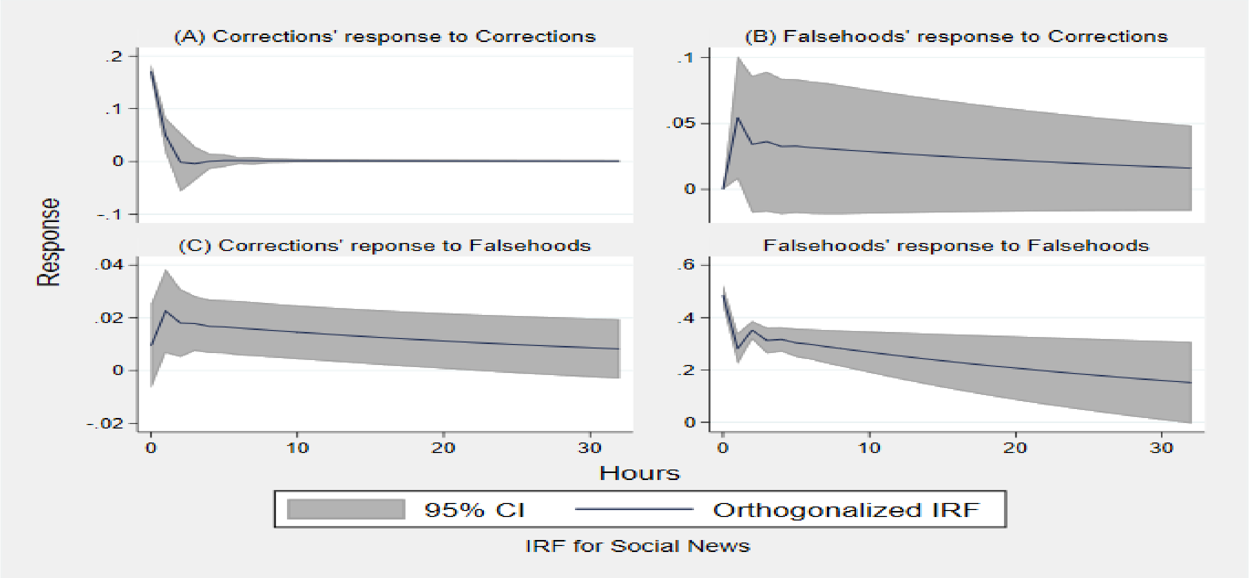


Figure 9. IRF plots for social false and correction tweets during both hurricanes.

The results from Model 5 (Figure 9) are consistent with those from Models 1 through 4 and show a positive relationship between falsehoods and corrections in the social news category. In addition, more falsehoods and correction tweets led to more of their own messages in subsequent hours. However, at the zero hour mark for example, the effect of correction tweets on falsehoods in the social news category was not significant until after the 2 hour mark (Plot C). The FEVD results show that in the first 24 hours corrections accounted for up to 1.2% of the forecast error of falsehoods, and falsehoods accounted for up to 1.9% of the forecast error of corrections in Model 5 for social news. Both were significantly greater than zero based on simulated results. The FEVD was stronger for falsehoods than for corrections for social news.

Robustness Checks

To show that effects of correction are robust to responses in Falsehoods, we calculated bihourly retweets counts using the total duration of both falsehoods and correction, from our main model (Model1). The results were consistent with previous results in which the IRF plots show that correction causes falsehoods to increase and that falsehoods also causes correction to increase. Also, the optimal lags of both our main model (Lag 3) and for both models using the 2 hours (Lag 2) show that the significance of the coefficients were consistent for our bihourly model.

Table 10. The bihourly effects of falsehood and correction tweets on each other

2 hours	N	Falsehoods on falsehoods	Falsehoods on Corrections	Corrections on Falsehoods	Corrections on Corrections
L1	3102	0.579*** (0.037)	0.072*** (0.023)	0.059*** (0.019)	0.614*** (0.032)
L2		0.278*** (0.035)	-0.034* (0.021)	-0.001 (0.016)	0.260*** (0.029)

Notes: * p<0.1; ** p<0.05; ***p<0.01.

Table 11. Falsehoods-Corrections FEVD for bihourly data

RV (False)	RV (Corr)
2hourly	

Hours	False	Corr	False	Corr
0	0	0	0	0
1	1	0	.007	.993
2	.997	.003	.015	.986
3	.996	.004	.020	.980
24	.969	.031	.122	.878
∞	.965	.036	.135	.865

Discussion

The current research examines the dynamic effects between the diffusion of falsehoods and correction messages on social media during shock events. Drawing on the competitive exclusion principle, we model falsehoods and correction messages as two specifies on Twitter and develop a theoretical model that illustrates different scenarios under which only one type of messages survive, both types of messages survive, or both die of. We further provide empirical validation of our theoretical modeling results using tweets data collected from Hurricanes Harvey and Florence. Our study has the following theoretical contribution and practical implications.

Theoretical Contributions

First, we contribute to the literature on information diffusion, rumors, misinformation, correction and competitive exclusion and bridge the gaps in literature by investigating the bidirectional relationships between both falsehoods and correction messages and how each affects the other in Twittersphere. Previous studies have examined the effects of either falsehoods or correction messages on human behavior (Ecker et al. 2010, 2011; Jolley and Douglas 2014) or the lasting effects of misinformation on the behavior of humans after exposure to falsehoods (Huang 2017). Some studies have also proposed methods to effectively debunk misinformation and present correction messages (Kim and Dennis 2017; Lewandowsky et al. 2012; Ozturk et al. 2015). However, these studies only focus on how one affects the other but not how both affect each other simultaneously. Our study is the first to examine both simultaneously and how falsehoods and correction messages affect each other's diffusion on social media. As a result, our research provides a more holistic picture of the different scenarios under which each type of messages survives or die off, how falsehoods and correction messages co-diffuse on social media, and how one affects the other's tweet count.

Second, we introduce the competitive exclusion principle to IS research and use to it develop our theoretical model of the diffusion of falsehoods and correction messages on social media. Deviating from previous studies that rely on epidemiology, we build our foundation on the principle that, when two species vying for the same limited resources encounter each other, there are three possible outcomes. First, the more dominant species eliminates the weaker one. In our model, the ideal scenario occurs when we see falsehoods die off because their replication or growth rate is considerably lower than their mortality rate and correction messages prevail with a higher replication or growth rate than mortality rate. This occurs when Twitter users share the correction messages more than falsehoods when they find these correction tweets to be more novel, news worthy or awe-inspiring. In contrast, when users find falsehoods to be more appealing, news worthy or believable, they would share them more than correction tweets, leading to the scenario when falsehoods prevail and correction messages die off. Second, the stronger species survives in the original ecosystem, and the weaker one engages in an evolutionary or behavioral shift and survives in a different niche so the two species are no longer in competition with each other. In the case of falsehoods and correction messages on Twitter, one may mutate or morph into a different message and then target a different audience. Prior research has shown that the weaker species may have a greater chance of survival if the overlapping niche is very small (Jaeger 1974). For example, falsehoods may mutate and become quite harmless, receive less tweets and shares and its target audience also changes. This is supported by empirical evidence where misinformation can mutate through textual changes and when tension from those messages dissolves, it resurfaces by repackaging itself as a different news to attract a different audience (Shin et al. 2018). Third, falsehoods and correction messages may both die off. On Twitter, this occurs when the users do not find the news or messages appealing enough to replicate or share. A typical example is our Topic 15 (Table 1) which had only 3 shares for falsehoods and 25 for corrections and lasted just 22 hours before both finally died off. A possible explanation is that the falsehoods were not important enough to the users and do not possibly contain anything intriguing to the recipients. Studies have shown that falsehoods usually contain less information but aim for novelty and some awe-inspiring effect on users (Osatuyi and Hughes 2018). Previous studies further collaborated this theory by showing that novel and awe-inspiring news go viral more than news that were not (Berger and Milkman 2012; Itti and Baldi 2009).

Third, we empirically validate our theoretical modeling results using falsehood and correction tweets data collected during Hurricanes Harvey and Florence. Our results are consistent across both hurricanes and different types of tweets. We find that in general, correction messages from FEMA and other fact checking organizations are not only ineffective but also help in spreading falsehoods on social media. We also find that falsehoods and correction messages will decline and eventually die out even without correction message albeit slowly. The results may be indicative of the flaws that can be attributed to FEMA and other fact checking organizations' responses to falsehoods on social media. Our result further contradicts those from studies on the effectiveness on using correction messages. For example, prior research (Huang 2017; Ozturk et al. 2015; Walter and Murphy 2018) suggests that exposing users to correction information that refutes falsehood reduces its spread. In contrast, our results indicate that the correction messages may be creating awareness and stimulating discussions and spread of falsehoods. Based on the competitive exclusion principle, this could be due to the higher replication rate of falsehoods as compared to correction messages. We propose increasing the replication rates of correction messages so as to overwhelm falsehoods. We also find that correction messages increase considerably when there is an introduction of false news. This can be attributed to the vigilance of correction agencies such as FEMA that introduce correction messages within the first few hours of experiencing falsehoods on Twittersphere. More importantly, false news increases instead of decreases when there is an introduction of correction messages. This is counterproductive to FEMA’s and other fact checking websites’ ultimate goal of reducing falsehood. Our results show that just leaving falsehoods to run its course is more effective than introducing correction messages. This phenomenon can be attributed to falsehoods losing their novelty over time if left alone. Prior studies have shown that the spread of falsehoods on social media is affected by the novelty of falsehoods such that falsehoods are more novel than real news and disseminate faster (King and Wang 2019). Moreover, studies on message framing showed that humans behaved consistently irrationally relying on several mental shortcuts to speed up our reasoning, which can make us remarkably sensitive to how things are framed (Tversky and Kahneman 1986). Recent studies have shown that we may be reinforcing beliefs when we attempt to warn of inherent misinformation such as during political elections without framing the correction accurately (Lakoff et al. 2004). This notion can also be attributed to the presentation and format of the news item as espoused by a recent study (Kim and Dennis 2017) that revealed certain changes in the way information is presented influences how users perceive and behave on the information.

Finally, our major theoretical and empirical results combined show how the diffusion of falsehood can eventually be reduced and removed by making the replication chance of correction greater than half of the product of replication chance of falsehoods using competitive exclusion principle from Ecology. Even though the ideal scenarios from our theoretical model suggests that one species will survive while the other one dies off, our empirical results suggest falsehoods and correction messages feed off each other on Twitter and each lead to more of the other in the next hour. Hence, the current state of correction messages has not reached the equilibrium state of reducing and eventually eliminating falsehoods. Given the ineffectiveness of the correction messages in reducing falsehoods, our theoretical modeling results show the conditions under which correction messages can effectively eliminate falsehoods. That is, when the growth or replication rate of correction messages are greater than its mortality rate, correction messages will expel falsehoods and prevail on Twitter.

Our robustness check shows that our models and analysis were robust to bihourly retweets on our major variables.

Practical Implications

This study has the following practical implications for government agencies and social media platforms. First, the results from our study can inform government agencies such as FEMA and policy makers on the ineffectiveness of current rebuttals and correction messages on social media, help them understand the relationships between falsehoods and correction messages and the impact of framing correction messages. This can go a long way in helping government agencies design more effective correction messages in the fight against misinformation. The findings from this study can assist these agencies and administrators in the design of effective emergency warning systems that can safeguard lives during emergency and crises situations such as earthquakes and hurricanes. Studies have shown that during emergency situations people are susceptible and fall prey to falsehoods (Silverman and Singer-Vine 2016). There are a few actionable recommendations we propose. For government agencies to effectively succeed in their combat against falsehoods on social media, they must first increase the replication rates (sharing) of correction messages. This may be achieved by restructuring and strategizing on their social media presence. Studies have shown that the majority of US citizens use social media for news (Reuters 2017). Government agencies should consider increasing their social media presence, for example, in every state. In addition, they can consider having their social media presence interconnected with other federal and state agencies across the country. These government agencies (federal and state) may then mirror rebuttals and corrections where appropriate. If such agencies work in unison, the replication rates for correction messages will dramatically increase. Moreover, government agencies may need to create incentives such as tax benefits for private entities and independent fact checking organizations to incentivize fact checking. Government agencies may also form partnerships with fact checking organizations and exchange information with each other when important falsehoods emerge. This collaboration ensures that shared knowledge can flow more freely, and the response time to debunking falsehoods is improved. It also provides a single voice in the fight against falsehoods. Finally, the agencies may need to consider developing more efficient correction messages that do not "amplify" the replication of falsehoods. This can be done by not repeating falsehoods in correction messages and creating interesting and innovative correction messages.

Second, the results from this study can help social media platforms in understanding the diffusion of falsehoods and correction messages and in combating the spread of the former. In order to reduce the replication rates of falsehoods, social media platforms can use more efficient automated detection technology to flag suspicious messages based on a combination of user, textual content and network features. These methods as a first line of defense may drastically reduce the total number of falsehoods being spread. In addition, empirical evidence suggests that bots are also responsible for sharing falsehoods (Vosoughi et al. 2018). Hence, it is important for social media platforms to effectively detect, thwart and remove automated malicious entities (King and Sun 2019). The adoption of such technologies will considerably reduce the replication of falsehoods by automated entities and ultimately reduce the spread and replication of falsehoods.

Conclusion and Limitations

We examine the bidirectional relationship between falsehoods and correction messages on social media. Our theoretical model and empirical validation show the scenarios under which each type of messages survive or die off and their impacts on each other. Our study has several limitations. First, we had to aggregate our tweet data at the hourly level to obtain the counts. Aggregation of data may cause loss of information at the granular level. Second, we only analyzed tweets during two crisis events from two periods. Future studies can investigate other news stories from other social media platforms to cross-validate our results. Third, we only included verifiable false news and correction messages. Future studies may examine other types of news such as rumors. Fourth, even though we showed the bidirectional relationships between falsehoods and correction messages, we were unable to show why that happened. Future studies may employ surveys or experiments to examine why correction messages are ineffective and to find out the optimal timing of correcting falsehoods.

CHAPTER IV

EFFECTS OF SENTIMENT ON THE MORPHING BEHAVIOR OF FALSEHOODS AND CORRECTION MESSAGES

Introduction

Deception in social media have received much attention from both academia and industry since it notably rose to global attention in 2016 during the Brexit votes and the US presidential election, where deception in the form of "fake news" was engineered as a deliberate campaign to wage war and influence user perception (Barthel et al. 2016). As a staging ground for modern movements, social networks have become the primary source of news (Reuters 2017), and it is no surprise as false news has deceived many people and created divisions in society (Silverman and Singer-Vine 2016). Moreover, falsehoods have potentially serious implications for public health and safety especially during shock crises situations such natural disasters as hurricanes and manmade events as terrorist activities. Traditionally, studies have concluded that falsehoods are propagated prevalently during what is termed as the "3Cs" which are during conflict events, crisis scenarios and catastrophes (Koenig 1985). An explanation for this is the level of anxiety or negative emotions that may be accompanied by some of these events. For example, during Hurricane Harvey in 2017, immigrants were adamant to evacuate disaster areas for fear of being rounded up and deported (Mendoza 2017). This belief stemmed from erroneous news stories that were spread on social media about the US Immigration and Customs agencies actively enforcing immigration laws. Many more of such incidents occurred during Hurricane Harvey where people were afraid of evacuating due to an inherent fear of contracting Tetanus, spurred on by news stories of the

disease. Finally, some false stories are unique in the manner at which they come back multiple times but with changes to their original textual content. An example is the incident of sharks on the freeway during Hurricane Harvey in Texas in 2017, which showed up again in Florida during Hurricane Irma in 2017 and later on during Hurricane Harvey in 2018 albeit with differences in content characteristics (McDonald 2018). The granularity of the changes in textual contents during the diffusion process of this evolving story makes it difficult for organizations such as the United States Federal Emergency Management Agency (FEMA) in their quest to protect lives and property to track and ultimately poses significant public health challenges and risks. Furthermore, though several studies have tried to investigate the concept of misinformation (Agrawal et al. 2013; Jindal et al. 2010; Oh et al. 2010), with several studies suggesting ways of addressing this phenomena (Friggeri et al. 2014; Torres et al. 2018, 2018), these studies have focused on content verification through reactive measures such as presentation and source credibility and their ratings (Kim and Dennis 2017; Oh et al. 2010) and through proactive methods such as detection (King and Sun 2018; Pérez-Rosas et al. 2017) and correction messages (Huang 2017; Ozturk et al. 2015).

In analyzing the spread or composition of the messages, studies have shown that falsehoods may be empowered by repetitions as they are shared on social media (DiFonzo and Bordia 2007). For example, one of the major characteristics of its virality is its ability to be recycled and mutate over time (Boyd et al. 2010). This is generally achieved by adding, deleting and shortening messages, motivated by the taking of ownership and the customization of the original message to suit one’s goals, which may be in part influenced by Twitter’s word limitations (Gil 2019; Kwak et al. 2010). Despite the growing interest on falsehoods on social media, few researches have examined how messages mutate or research fills this gap in the literature that has investigated falsehoods, real news and correction news as a static communication process and examines how the textual contents of these messages change over time on Twitter.

Because emotive and affective components are important factors in the virality of messages on social media(Osatuyi and Hughes 2018), we examine how affective components of a message may be a major factor in the morphing behavior of not just falsehoods but of correction messages. Using cosine similarity to measure the morphing of the messages, our empirical analysis show that emotive components affect the morphing of both falsehoods and correction messages and positive emotions are more influential as compared to negative and neutral emotions. We also show that emotionally charged messages in general morph more than neutral ones irrespective of the period. This finding helps us to understand one major factor that is responsible for the potential aggressive mutation of correction actions on falsehoods. Emotions spur the aggressive mutation of both correction and positive news as a response to countering falsehoods and negatively valence news.

Background Literature

Our study draws on two streams of research – sentiment and information morphing – and explores each of these as they apply to our study.

Sentiment

With the rise of Twitter as one of the most archetypal social media platforms for user-generated content, researchers in information systems and beyond have since relied on Twitter sentiments for inferring user behavior (Liang et al. 2016). These studies have ranged from the use of microblogs on unidirectional platforms such as Twitter which leads to asymmetrical connections (Stieglitz and Dang-Xuan 2013) to bidirectional platforms such as Facebook (Stieglitz and Dang-Xuan 2012). These studies have revealed the importance of sentiments, an intentional or unintentional affective or emotional state affecting a user's judgment of a particular topic which can be inferred from textual contents (Bollen et al. 2011), in understanding user behavior and reactive tendencies to information sharing (Bene 2017). However, results on the impacts of sentiments on user behavior on social media have been contradictory. For example, though a recent study showed that emotionally charged political messages are tweeted more (Stieglitz and Dang-Xuan 2013), other studies have alluded to the efficacy of mostly negative valence over positive ones in influencing virality, especially when it comes to news (Hansen et al. 2011). Some of the reasons alluded to this is the moderating effects of novelty or the newness of the news stories (Itti and Baldi 2009; Vosoughi et al. 2018). We summarize in Table 12 the relevant literature on sentiments.

Table 12. Literature on Sentiment

Topic	**Theory**	**Data & Methods**	**Results**
Sentiment analysis tools (Abbasi et al. 2014)	Literature on several sentiment analysis tools	Meta-Analysis of several tools for sentiment analysis	Structural issues such as sarcasm and jokes accounted for the largest percentage of highly erroneous tweets.
Consumer decisions (Di Muro and Murray 2012)	Literature on mood regulation	Two experimental studies using ANOVA	Users choose items that are congruent with their current mood.
Employee blogging and sentiments	Literature on blogs	OLS from three archival sources, fortune 500 IT firms, blog search engine	Negative posts act as catalyst and can increase the

(Aggarwal et al. 2012)		Technorati.com and daily XML feeds from Bloglines.	readership of employee blogs.
Sentiment and profitability (Constantinos Antoniou et al. 2013)	Literature on sentiment and stock price	Regression on common stocks from NYSE and AMEX and sentiment states	News that contradicts investors' sentiment causes cognitive dissonance, slowing the diffusion of such news.
Sentiment (Stieglitz and Dang-Xuan 2013)	Literature on Twitter and its use	Sentiment analysis on tweets and information sharing	Emotionally charged political tweets were shared quicker and more often than neutral ones.
Politics (Bene 2017)	Literature on communication	Negative binomial regression on Facebook posts of politicians in Hungary	Users are highly reactive to negative emotion-filled posts, influencing virality.
Emotions and online contents (Berger and Milkman 2013)	Literature on the virality of news articles	Sentiment analysis and probability	Emotional effect affects virality above and beyond other factors.
Social transmission (Jonah Berger 2011)	Short literature on transmission of information	Two experiments on emotion and sharing using ANOVA	Arousal-inducing contents are shared more than those others.
Emotions (Harris and Paradice 2007)	Literature on affective information and emotion	Laboratory experiments on 225 students using ANOVA	The higher the number of emotional cues from a sender, the higher the degree of emotions from responders.
Emotion (Kissler et al. 2007)	Literature on studies relating to event-related potential and emotional responses	Experiment on students from German university using an EEG and data analysis using ANOVA	Emotional words were associated with enhanced brain responses. Emotionally charged words were also better remembered than neutral words.

Emotion (Stieglitz and Dang-Xuan 2012)	Literature on Facebook and sentiment diffusion	Regression and sentiment analysis using LIWC on Facebook political public pages.	Positive and negative emotions in a post have a positive relationship with corresponding comments.

Information Morphing

Information processing in humans begins in the brain. Studies show that not only does the brain recollect events slightly different from what was actually perceived but that the brain and memory can be altered through suggestions and nudging (Braun-LaTour et al. 2006; Loftus and Pickerell 1995). This phenomena is often termed as recall accuracy (Marks 1973; Toglia 1999) and can lead to biases (Raphael 1987). However, information morphing based on social media interactions are often intentional and not directly related to our recall. Rather, they are due to a user’s need to either create an awe-inspiring tweet or even more plausible to “personalize” the tweet (Boyd et al. 2010). This feat can usually be achieved by adding, subtracting and substituting characters in the text. Those characters may not usually change in meaning but may convey the original tweeter’s beliefs and a state of mind.

We define *information morphing* on social media as the constant change in textual constants from its original message over time. Despite an abundance of research on information and rumor diffusion, the focus of the extant literature has been on the diffusion rate and its contributors with none of the studies focusing on their evolution during the diffusion process (Cheng et al. 2013; Jindal et al. 2010). There have been very few studies, if any, that have attempted to understand how news evolve over time. For example, Friggeri et. al (Friggeri et al. 2014) analyzed rumor evolution on Facebook in the form of memes. They found that rumors do not particular die out but persist in low frequencies and come back after a while. According to Kim et al. (Kim and

Dennis 2017), changes in the news presentation formats have significant impacts on how the recipients perceive and interpret the news. Furthermore, using political tweets, a recent study analyzed the average change in corpus of false and real tweets when they resurfaced and found that on average falsehoods change at an average of 0.5 when they are reintroduced, while real news was not investigated because they were not observed to return (Shin et al. 2018). The researchers suggested the need for future studies on the length of text and sentiments as morphing occurs. Also scholars argue that false news gains its strength through repetitions and its repeatability (DiFonzo and Bordia 2007).

A prominent study using experiments showed that messages similar to information get distorted over time as they flow through a channel (Treadway and McCloskey 1987). A classic case of information distortion was when psychologists in an experimental setting showed subjects a picture of several people on a subway train, which included a white man holding an open razor and having an altercation with a black man. When prompted to elaborate on what they witnessed, more than fifty percent of the subjects stated they saw the black man instead brandishing the razor at the white man (Woocher 1977).

As a microblogging site, Twitter depends on a directed friendship or followership even though reciprocity is not required (Marwick and boyd 2011). Retweeting, which is basically reposting an original post, can introduce the content to a new audience and such retweeted messages can usually be modified so that they lose any reference to the original or even posted to a different social network (Boyd et al. 2010). This propels tweets to go even further without the knowledge of the original tweeter as they reach a wider audience (Marwick and boyd 2011). This implies that the morphing of Twitter messages can take about any form, as Twitter contains emojis and allows modification to whatever extent that suits the re-tweeter’s ideology.

Individuals who do not have the same ideologies as the original tweet thus have an opportunity to retweet a response to the contrary, which may also propel the original tweet but using different contents and context, some of which are usually in the form of corrections and rebuttals. As news travels from one person or place to another, messages are either accepted or rejected based on an individual's cognitive homogeneity (Torres et al. 2018). Re-tweeters may thus opt to include texts, emojis and images to appeal to affective states of other readers. In the current study, we use cosine similarity to measure how similar a tweet is to the original tweet.

Research shows that rebuttals and corrections at times can be very effective in addressing misinformation on social media by reducing the credibility of the refuted content (Huang 2017). Study further shows that message or rumor-correcting tweets are more propagated or spread more than the rumors themselves (Chua et al. 2017). This is very important as it shows the power of rebuttals, coupled with the fact that such rebuttals that are retweeted can be altered and modified. This study thus allows us to have a better understanding of the mechanics on how false news morphs over time and the role rebuttals play in the evolution framework. Our study is quite different from previously mentioned studies and places relatively less emphasis on the generality of the spread of the underlying phenomenon. In addition, these previous studies tend to treat the mutability of misinformation as a corpus instead of in granularity. On the other hand, our study takes an alternative perspective, which views misinformation as verifiable false news that are mutable and robust as they diffuse. We explore this idea using a fixed effects model with multiple time series levels while controlling for the word counts' increase/decrease and variability.

While prior studies were relegated to analyzing specific and limited number of political rumors, we bridge the gap in the literature by employing a large panel dataset comprised of 14 different topics collected within a shock event in 2018. Furthermore, unlike previous studies that focus only on rumors, we compare and attempt to understand the differences between falsehoods, real news and rebuttals.

Finally, unlike previous studies, we attempt to understand not just false news but the differences between the mutability of false and correction news using their entire life cycles. We also cross-checked our results using a brief exploratory analysis.

Research Questions

Our study differs significantly from previous research (Friggeri et al. 2014; Kim and Dennis 2017; Shin et al. 2018) in several ways. First, our study places emphasis on the granularity of the evolution of tweets in every text and not as a corpus. Secondly, we place emphasis on the daily, hourly evolution of the messages and the type of messages and control for the word count. Our study aims at investigating the diffusion of verified news items as an evolving phenomenon focusing on the changes in pattern as they each diffuse on social media. To achieve this goal, we employ a unique panel analysis on 14 verified falsehoods and corrections topics that circulated on Twitter during Hurricane Harvey in 2017. We also seek to understand the role sentiment plays in the nomology of things. We attempt to answer the following research questions:

1. How do false and correction messages morph on social media?
2. What are the effects of sentiments on the morphing of false news and correction messages?

Research Hypotheses

For quite some time researchers have argued that rumors and falsehoods are infamously effective in causing disruptions due to their ability to cause reactions from their highly emotional contents (Bene 2017; Berger and Milkman 2013). In sharing news, these reactions may be manifested in several ways, including the modification of the said news item in synched to the user's current affective state. A recent study shows that due to character limitations from Twitter, users are known to perform several of the following modifications: first they do so by shortening tweets through deleting, preserving and adapting tweets for their own purposes and the use of authorship and attribution (Boyd et al. 2010). The use of these methods leads to changes in the original content but not necessarily the context irrespective of the magnitude of change. A recent study revealed the virality of positive news stories (Berger and Milkman 2012). The authors posit that even when controlling for novelty or usefulness of news items, positive news is usually shared more than neutral ones. Furthermore, the researchers argued on the causal impacts of emotions as the driving force behind this behavior. We argue that due to the virality and the emotional connotation, positive news will therefore morph more than other news types. We therefore hypothesize:

H1: *An original tweet's sentiment is positively associated with its morphing.*

Studies have shown that emotionally charged messages influence reactivity in receivers as compared to neutral ones (Stieglitz and Dang-Xuan 2013). This may be because they influence the affective components in the brain and induce reactions without the user extending their cognitive process. Studies have since tried to show that those affective components trigger a peripheral thought process (Angst and Agarwal 2009; King and Sun 2018; Osatuyi and Hughes 2018) but not their cognitive process, and this may lead to irrational negative behaviors (King and Sun 2018). A study using electroencephalogram (Kissler et al. 2007) showed that emotional words influence high amount of brain responses as compared to neutral ones. In general, we argue that emotive tweets will cause users to react and change the contents of tweets before sharing in order to synchronize and personalize their own feelings as compared to neutral tweets. Hence, we have:

H2: *An original tweet's sentiment squared is positively associated with its morphing.*

Bad news, emotions or events have long held sway over those that were inherently good, as a general principle across a broad range of psychological phenomena (Baumeister et al. 2001). Fake news and real news can be categorized as good and bad, and in this instance fake news stories have been shown to be more viral and influential in sharing behavior as compared to other types of news (Vosoughi et al. 2018). The novelty of such fake news stories abound and entice users on social networks to take ownership of them in order to increase their social media standing (Itti and Baldi 2009). A study has shown that when a user takes possession of such a tweet they are more likely to engage in authorship attrition and/or the preservation and adaptation of the original message (Boyd et al. 2010). This adaptation is what leads users to shortening or deleting part of the tweets and adapting them for their own purpose and writing style. When this happens the similarities between the original and the retweet will change. The same has not been said for real news stories as they are generally not considered as novel as fake news (Friggeri et al. 2014; Vosoughi et al. 2018) and hence may not contain enough novelty to warrant such zealous modifications. Nor are they known to cause such reactivity. However, correction news are very different both in tone and intensity from real news as they rebut falsehoods and usually do so in the strongest possible terms. The strength of correction messages may lie in their strongly worded context and how they differ from the falsehoods. When arguing against a topic, one is usually expected to imply the topic in question and modify the argument against. While real news does not contain novel information, correction messages may contain more novel information as to efficiently rebut the argument in question. This means that correction messages may stimulate more interest and be more modified more than false new. As a result, we hypothesize:

H3*: An original tweet's veracity is positively associated with its morphing.*

Through analyzing news articles in the New York Times, a recent study revealed that positive affections highly influenced virality (Berger and Milkman 2010). This can be because people and their decision making are geared towards maintaining a sense of positivity as they go about their everyday tasks (Di Muro and Murray 2012). As a result, individuals are more likely to maintain and even increase the status quo when modifying a text. Such modifications could include improving on a positive tweet to include jokes and emoticons that may increase morphing, while at the same time increasing the positivity of the previous tweet's sentiment. We argue that, with each tweet, users over time will try to upend the positivity of the previous tweet and thus as time goes by the morphing may increase over time. On the other hand, the modification of tweets with negative sentiments may not be sustainable as time goes by due to the loss of newness or surprise value. We therefore hypothesize:

H4: *The positive association between sentiment and morphing gets stronger over time.*

Considering emotionally charged tweets are expected to influence virality more than neutral tweets (Stieglitz and Dang-Xuan 2013), we expect that group emotional contagion may in fact assist in the transfer of moods and emotions (Barsade 2002). This means that if there are no emotions or the emotional valence of the tweet is neutral, it may not receive much attention and as such may not be retweeted more. Just as the original message may convey such emotions, positive and negative emotions will be transferred to the recipient and their modifications would then be a direct reflection of their emotional state. The user's modification of the tweet whether positive or negative can then be easily seen from the modification of the text. And as time goes by and more users receive the tweet, the emotions are transferred to and from and expressed by the modification of the textual contents. Thus over time, the emotional aspects will cause several modifications as an expression of transferred emotions over time. Moreover, as time goes by and the novelty in a tweet decreases, a neutral tweet will quickly lose traction and be modified less. In contrast, a tweet with a positive or negative emotion will be able to better withstand the test of time due to the emotion contained in the message and continue to morph as times goes by. We therefore hypothesize:

H5: *The positive association between sentiment squared and morphing gets stronger over time.*

As correction news morphs faster than fake news due to the desire to confront the “fakeness” of a news source, we argue that it is more likely to also morph more as time goes on. For example, since studies have shown that fake news in general may diffuse faster and morph at a higher rate in a short amount of time than other news (Vosoughi et al. 2018). Correction news may make up for this deficiency by being more emotional and aggressive in their response to falsehoods. This reaction will give way to a more aggressive morphing behavior as time goes by. Also, considering that falsehoods must first be introduced in the nomology for correction messages to even exist, we argue that the mechanisms underlying correction messages may be playing “catchup” and as such need to increase their morphing behavior over time. We therefore foresee that over time due to the aggressive stances employed in rebutting falsehoods, correction messages may increase morphing behavior at both the short term and the long run more than falsehoods. We argue that as times goes by morphing may increase more for correction news than false news. We therefore hypothesize:

H6: The positive association between veracity and morphing gets stronger over time.

Sample and Methodology

Sample

In this study we investigate the morphing of tweets by first identifying and collecting all tweets for each day from Hurricane Harvey's formation on August 17 through its aftermath on September 27, 2017 using Twitter's filter/streaming query command. We only retained verifiable false and correction tweets based on FEMA's rumor control page and three fact checking websites including Factcheck.org, Snopes.com, Truth or fiction. We obtained 28 original tweets consisting of 14 fake tweets and 14 correction tweets as a result. Next, we collected all their retweets for a 5-week period based on their original topics and tweets. We obtained a total of 150,907 tweets and retweets for our first-step exploratory analysis on the morphing hazard rates of falsehoods and correction messages.

Next, we leveraged SpaCY and the natural language processing libraries in Python to calculate the sentiments of the tweets as SpaCy provides a fast and accurate syntactic analysis following an approach by (King 2020). We marked up words in our corpus as corresponding to a part of speech using its meaning and its association with related words in the sentence. In text analysis, though the wrapper method is more commonly used, the filter method was the more appropriate due to its efficacy with large datasets as recommended by (Huei Chou et al. 2010).The polarity and subjective scores for each sentence were then saved and used for our analysis. The polarity score is the raw sentiment orientation of the textual content, which ranges from 1 to 99.99 for positive sentiment, 0 for neutral, and -1 to -99.99 for negative sentiment. Since our independent variable is the change in characters of a tweet, we controlled for word count and used time in hour as an exogenous variable in order to reduce endogeneity. We obtained a total of 133,319 verified tweets and retweets for our second-step empirical analyses on the factors affecting the morphing of falsehoods and correction messages.

Cosine Similarity

In this study, we analyze original tweets and their corresponding retweets over a period to understand how they morph during diffusion. An efficient way of measuring the similarities or differences in data and documents with textual contents such as tweets is the use of clustering techniques. Clustering in itself is simply dividing data into various groups based on object similarity (Berkhin 2006). Agglomerative clustering is a type of hierarchical clustering method used in data mining that begins at some point and repeatedly combines two or more suitable clusters (Berkhin 2006). Cosine similarity is an agglomerative clustering technique that has been used intensively in face detection (Nguyen and Bai 2010) and web clustering (Strehl et al. 2000) and has been proven to be very effective in business use as a means for cataloging and documenting large corpuses of documents (Cutting et al. 1992; Steinbach et al. 2000). The cosine similarity between vectors X and Y is denoted as $Cosine(X; Y) = X*Y/X|Y$.

For this study, we define morphing as the change of characters in a tweet that does not change the original meaning of a tweet. This requires the use of text distance measures. Some of the more popular text distance measures include using the Hamming distance, the Jaccard distance, the Levenshtein distance and the cosine distance. The Hamming distance compares every letter of two strings with respect to their position (Jayram et al. 2008; Norouzi et al. 2012). This means that the first letter of the first word will be compared to the first letter in the string of the second word. One major advantage of this method is that it is relatively fast and perform. However, it is unable to accurately compute two strings with an uneven number of letters. For a media outlet such as Twitter that relies on slangs, short words and different jargons meaning the same things, relying on this method will not produce an optimal or accurate output.

Another measure of similarity is the Jaccard distance, which measures the dissimilarity between sample sets (H.Gomaa and A. Fahmy 2013; Kim et al. 2020; Niwattanakul et al. 2013). The Jaccard distance is calculated by finding the Jaccard index and subtracting it from 1, or alternatively dividing the differences by the intersection of the two sets. It is calculated by finding the number in both sets, divided by the number in either set, multiplied by 100. This will produce a percentage measurement of the similarity between the two sample sets. The Jaccard distance is then calculated by simply subtracting the percentage value from 1 or as the inverse of its coefficient (Niwattanakul et al. 2013). This measure is mainly used in convolutional neural networks in image identifications and the coactualization of object detection.

Another distance measure is the Levenshtein distance and it is the number of operations needed to convert one string to another (H.Gomaa and A. Fahmy 2013). This distance measure penalizes for every edit and as such every edit needed will add 1 to the Levenshtein distance when inserting, adding, deleting or substituting characters (Gooskens and Heeringa n.d.; Niewiarowski 2019). This distance measure is intuitive but is computationally intensive, and its algorithm is difficult to implement accurately.

The cosine distance is a term-based similarity measure that considers the distance between two documents and is commonly used in natural language processing. It applies to the vector representation of documents, and the cosine similarity vectorizes the text by converting them into numerical data (H.Gomaa and A. Fahmy 2013). It is calculated by computing the dot product of two vectors divided by the norm of a times the norm of “b”. To calculate the morphing of false news, we calculated the cosine similarity of the word vectors. This is a common agglomerative clustering technique used for analyzing the similarities or differences between textual contents and has been used extensively in face detection (Nguyen and Bai 2010), web processing (Strehl et al. 2000), and cataloging large corpus of data (Cutting et al. 1992; Steinbach et al. 2000) . This method is also used to divide the data into various groups based on object similarity (Berkhin 2006). It is able to show the distances between corpuses of tweets that are in a multidimensional term vector space which is defined by the cosine of the angles (Shin et al. 2018). The cosine similarity for the tweets begins when the initial tweet is assigned a numerical value of 0 and then its similarity is compared with subsequent tweets and assigned values based on similarities. The initial value assigned is a comparative between the initial tweet on itself and should show no differences and is assigned a value of zero. The closer the cosine similarity value is to zero, the more similar the tweets are. The larger the cosine similarity, the more different the newer tweet is from the original tweet. Hence, the cosine similarity between two vectors X and Y is given by Cosine (X; Y) = X*Y/X|Y.

Because a low cosine similarity means low mutation or morphing from the original tweet and vice versa, we expect the cosine similarity of positive and correction tweets to be higher than false news over time.

Exploratory Survival Analysis

Survival analysis analyzes the occurrence of an event as a failure process starting from a certain point in time and the factors associated with the occurrence of the event (David A. Freedman 2008; Michael G. Akritas 2004; Oakes 2000). It relies on the expected duration of time until one or more events occur. Survival analysis has been applied in IS to study behavioral patterns such as the diffusion of technologies (B. Baesens et al. 2005; GARETH O. ROBERTS and LAURA M. SANGALLI 2010; Massimo G. Colombo and Rocco Mosconi n.d.). Treating the mutation or morphing of a tweet on the same topic with the same veracity as an event, we analyzed the hazard functions for the morphing of both falsehoods and correction messages. Because mutation can occur multiple times for the same original tweet, we treated the mutation of falsehoods and correction messages as independent recurring event where the characters change in an original (first) tweet over time (Cox 1972).

The Andersen-Gill (AG) model, an extension of the Cox proportional hazards model, is the most frequently used model to examine the occurrence of recurrent events (Andersen and Gill 1982). It relates the intensity function of event recurrences to the covariates multiplicatively and treats each subject as a multi event with independent increments which has a common baseline hazard function for all recurring events. The (AG) is appropriate for our analysis because it assumes that each tweet ad retweet is independent and does not rely explicitly on previous events before they occur.

The hazard function $\lambda_{ik(t)}$ for the k^{th} event of the i^{th} subject is denoted as:

$$\lambda_{ik}(t) = \lambda_0(t)e^{Xik\beta}. \qquad (1)$$

We assume that the similarity spreads or changes as a result of the message contacts between users of the network per topic. In our analysis, during the diffusion process, a 1 means that there was a change in the original tweet or mutation, while a 0 means the observation was censored and was not observed to morph during the period of the analysis.

We analyzed our tweet data of 150,907 observations and present the Nelson cumulative hazard functions for falsehoods and correction messages using the AG model Figure 1. The Nelson cumulative hazard function for recurring events represents the expected number of events for a unit that has been observed for the given amount of time. The results indicate that although falsehoods had a slightly higher initial morphing rate, correction messages morphed faster than falsehoods after the first 60 hours. This might be as a result of competition between both falsehoods and correction. Lastly, falsehoods also morphed 20 hours longer than correction messages as no events were observed after about 680 hours for correction messages.

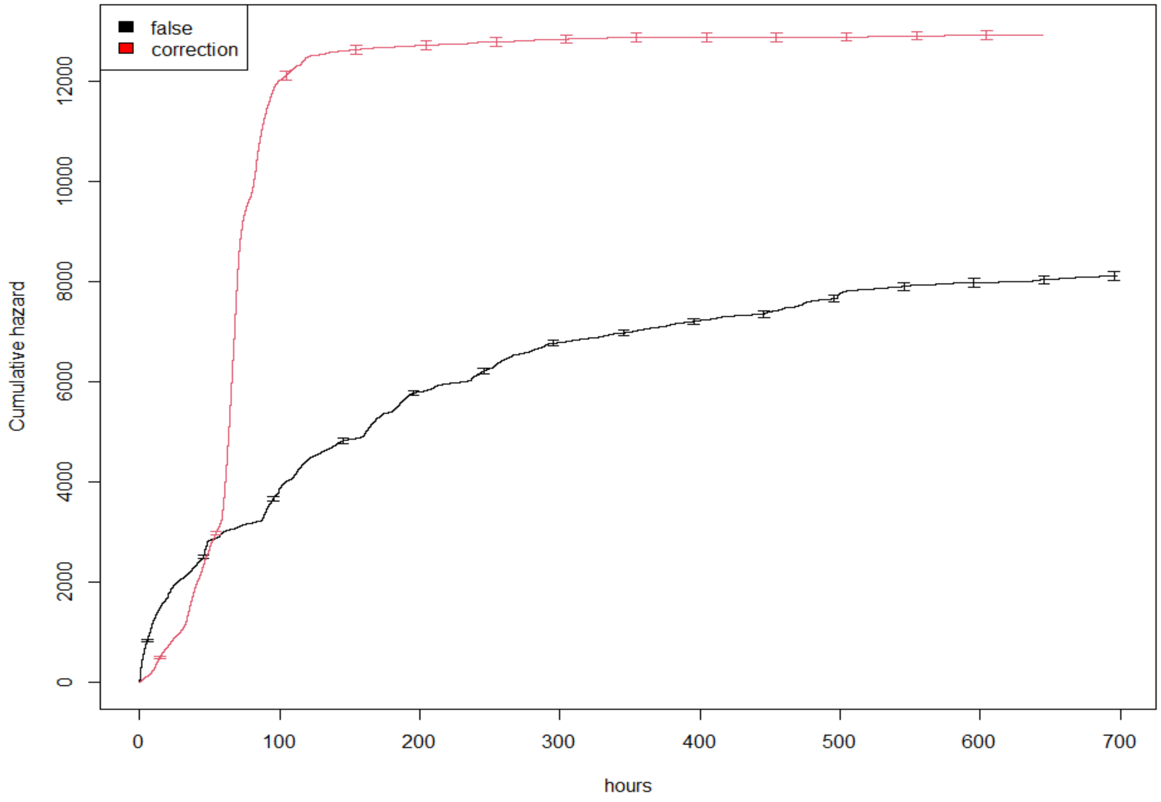


Figure 10. Nelson cumulative hazard functions for false and correction tweets.

Empirical Analyses

Variable Definition

We summarize our variable definitions in Table 13. In addition to our dependent and independent variables, we also included three control variables including the word count and variation to control for the length of the tweet and the morphing history on the morphing of a tweet at time t. We performed both the Breusch Pagan and the white's test for heteroskedasticity and obtained the white heteroscedastic-consistent robust estimates which corrects for and is efficient for large samples for our analysis. Table 14 summarizes the sample descriptive statistics.

Table 13. Variables and definitions

Variable	**Definition**
Dependent Variable	
Morphing	Cosine similarity score between 0 and 99.999 for each tweet or retweet.
Independent Variables	
Veracity	1 if the original tweet is a verified true or correction tweet, and 0 if verified false.
Sentiment	The raw score of the sentiment of the tweet from -20 to 20.
Time	The number of hours that had elapsed since the original tweet on the same topic.
Control Variables	
Word Count	The number of words in a tweet.
Variation	The average cosine similarity from the second tweet to the last tweet on the same topic with the same veracity.

Table 14. Sample descriptive statistics (N=133,319)

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Cosine similarity	4.873	1.446	0	29.292
Sentiment	-0.468	3.585	-20	18.75
Veracity	0.505	0.500	0	1
Time	72.960	59.895	0	637
Word count	18.767	4.873	1	111
Variation	4.755	0.476	1.079	5.931

Model Specification

Equation 1 specifies our empirical model to examine the morphing an original tweet $X_{i,0}$ to $X_{i,t}$ at time t.

Cosine (X_0; X_t) = β_0+β_1Sentiment$_i$ + β_2Sentiment$_i^2$ + β_3Veracity$_i$+ β_4 *t + β_5Sentiment$_i$*t + β_6Sentiment$_i^2$*t + β_7Veracity$_i$*t +β_8 WordCount$_i$+ β_9Variation$_{i,(2,t-1)}$ + $\varepsilon_{i,t}$.   (1)

Table 15 summarizes the results of our empirical analyses . All our independent variables had variance inflation factors less than 4 with a mean value of 2.17.

Table 15. Results of robust model during Hurricane Harvey (N=133,319)

	Model 1	**Model 2**	**Model 3**
Intercept	1.506*** (0.078)	-0.643*** (0.094)	-0.913*** (0.120)
Sentiment		0.059*** (0.002)	0.035*** (0.012)
Sentiment^2		0.006*** (0.0002)	0.002*** (0.001)
Veracity		0.797*** (0.008)	0.867*** (0.037)
Time	0.003*** (0.00007)	0.004*** (0.00007)	0.003*** (0.0001)
Word Count	0.021*** (0.001)	0.040*** (0.001)	0.041*** (0.001)
Variation	0.576*** (0.015)	0.855*** (0.019)	0.911*** (0.025)
Sentiment*time			0.0004* (0.0002)
Sentiment^2*time			0.00006**

			(0.00002)
Veracity*time			-0.001 (0.001)
RMSE	1.4093	1.337	1.3355
R-Squared	0.0508	0.1456	0.1476

Note: RMSE: Root mean square error. *p<0.10; ** p<0.05; ***p<0.01.

Our first model depicts the cosine similarity (morphing) of a tweet as a function of our control variables. This is our baseline model which has time and two control variables: the word count and variation. We find that the intercept and all variables were significant at the 0.01 level.

In Model 2, we added our independent variables including sentiment, sentiment^2, and veracity. The coefficients for veracity, sentiment and sentiment^2 were all positive and significant at the 0.01 level. This is consistent with both H1, H2 and H3 but inconsistent with the previous literature where positive sentiments and real news are less likely to evoke emotions and hence may not encourage much reactivity to engage in modification of text (Berger and Milkman 2010). Result on sentiment^2 was however consistent with previous studies, which had shown that emotionally charged messages were most likely to evoke reaction and cause virality and, in our case, leading to users changing the textual contents of the tweets.

In Model 3, we added the interaction terms between time and the independent variables. The coefficient estimates and their significance levels of the independent and control variables were consistent with those from Model 2. Among the newly added interaction terms, the coefficient estimate for the interaction term between sentiment and time was positive and significant at the 0.10 level, supporting H4. The coefficient estimates of 0.0004 means that for each additional hour the effect of sentiment on the cosine similarity increases by 0.0004.

The coefficient estimates for the interaction term between sentiment^2 and time was positive and significant at the 0.05 level, supporting H5. The coefficient estimates of 0.00006 means that for each additional hour the effect of sentiment^2 increased by 0.00006. This result suggests that the relationship between emotionally charged tweets and cosine similarity increased over time.

The coefficient estimate for the interaction term between veracity and time was not significant. Hence, H6 was not supported.

Discussion

Using data collected during Hurricane Harvey, this research examines the interactions between sentiments and the morphing behavior of fake and correction messages on Twitter. Our research has the following theoretical contributions and practical implications.

Theoretical Contribution

This study makes several contributions to the literature on fake news, sentiments, information diffusion and information morphing on social media. First, we provided a visualization of the hazard rates of morphing of both fake news and their correction messages on Twitter using survival analysis. Our results show that correction messages morph more aggressively than falsehoods. However, over time there is not much difference in their morphing hazard rates after 650 hours when correction messages end.

Second, we developed an empirical model for predicting the morphing of messages on Twitter. Despite increasing interest in academia on the diffusion of fake news, to the best of our knowledge, this is the first time a research has been conducted with a high level of granularity on information morphing on social media.

Third, we identified factors such as sentiment, the square of sentiment and veracity that may influence the morphing of both false and correction messages. Prior research has already showed us that the virality of tweet messages are usually accompanied by the mutation of the news item over time (Boyd et al. 2010). Our results show that sentiments influence morphing behavior. Though most studies claim negative valence in news may influence sharing behavior and cause reactivity (Stieglitz and Dang-Xuan 2012), a study showed that positive news may influence sharing and virality (Berger and Milkman 2010). This is consistent with our results. A possible explanation could be that during extreme events positive news may retain some novelty and thus may cause individuals to not only share but change the textual contents before sharing. Also previous literature shows that positive messages sometimes are more readily shared and cause more reactions than negative or neutral ones. Our findings showed that certain contents that end up evoking a lower form of arousal like sadness ended up being less viral. A recent study lends credence to our findings and showed that positive emotions affects profitability and influences momentum in the financial arena (Constantinos Antoniou et al. 2013). The researchers showed that momentum was positively affected by optimism rather than other emotions. This level of optimism leads to a built in momentum which rallies and encourages positive actions. In the case of Twitter it is the sharing and mutation of content. We also find that in general, tweets that are emotionally charged (positive and negative) have a positive effect on morphing and are more likely to cause reactivity and content changes. This is consistent with the previous literature that showed that emotionally charged messages were more likely to be shared than neutral messages (Stieglitz and Dang-Xuan 2013). A possible explanation is that those contents may be able to induce cognitive and arousal-related effects which might compel reactivity. It is this reactivity that influences users to want to make a tweet more personal, thereby modifying the tweet to synchronize with their current affect state. Also a recent study showed that emotions in general elicit the social sharing of emotions (Rimé 2009). According to that study, the type of emotions in a message is usually transferred to the receiver, which means messages with neutral emotions would garner lukewarm attitude and elicit very little if any reactivity, unlike emotionally charged messages. Our results also showed that correction news morphed more than false news. This result is inconsistent with the previous literature on virality, which showed that false news may diffuse a lot more than news that is inherently not false (Vosoughi et al. 2018). The authors were unable to attribute its virality to user characteristics nor network properties but provided an alternative explanation to which novelty may be what propelled falsehoods. That means that irrespective of the news item, a major hallmark for virality may be linked to novelty or an element of surprise and depending on the right conditions, either type of news contents may end up being more viral. Following this reasoning, during extreme events where there may be an abundance of falsehoods, users may attempt to correct such news stories with fervor such that they may keep modifying the news stories more than the competing falsehoods. Another possibility could be that positive news or correction news may attempt to exaggerate positivity of an event already posited as bad by false news contents to sway users and lift their spirit high. This gives a sense of hope during crisis situations. For example, a recent study (Shin et al. 2018) showed that rumor resurgence often accompanied changes in textual contents and were mostly in the direction of exaggeration. Finally, users who share positive news during extreme events may want to personalize the message so that it is seen by the receivers as originating from them. That way they would be perceived as novel disseminators of the news, and it might improve their standing in the network.

Fourth, we also compare the impacts of the above mentioned factors on morphing behavior over time. Our results show that not only do positive sentiments morph faster, the morphing also accelerates over time. It shows that positively charged news are more likely to garner more changes in content than both neutral and negative news as time goes by. This can be due to the novelty of the news item in the earlier stages of crisis situations. However at the later stages, even though the novelty may wear off (Itti and Baldi 2009), it does not take away from the positivity that encourages people and may give them hope. Hence, the morphing continues. The morphing may also continue as the more the messages change it may seek to exaggerate the positive nature to make up for the loss in novelty by including emojis to influence affective emotions like happiness, anticipation, joy and trust unlike negative news which may exhibit some novelty in the initial stages which is usually accompanied by awe inspiring and fear induced contents (Vosoughi et al. 2018). However, after these types of emotion wears off, there is no longer a need to reshare with that much fervor and as such it may evolve slowly. A recent study showed some negative and false rumors in the form of memes have the uncanny characteristic to persist in low frequencies, sometimes becoming dormant months un-end before flaring up again (Friggeri et al. 2014). Furthermore, we show that emotionally charged tweets morphs faster than neutral tweets as time goes. This is consistent with the previous literature that showed that emotionally charged tweets are more likely to go viral than neutral tweets (Stieglitz and Dang-Xuan 2013). A possible explanation for this is that emotionally charged (positive or negative) tweets affect the emotions of the receiver. As such they may influence users' desire to make changes to the contents of the original tweet. Also, the strength of the influence of this emotion may tend to be the same and exogenous irrespective of the time, which may be unlike that of neutral tweets. Besides, neutral tweets do not usually contain any awe-inspiring contents compared with emotionally charged messages. In addition, studies have shown that emotional messages were also spontaneously better remembered than neutral words (Kissler et al. 2007). This means that such emotions would linger on and over time increase as time goes by compared to neutral ones.

Fifth, we show the importance of including control variables such as word count and variation. The word count captures the number of words in the tweet. Twitter utilizes a method where several tweets can be included into a single tweet called new lines. Hence, our word count is not the count of words in the sentences but the count of words in the tweets. A recent study showed that one of the ways in which users personalize tweets is to delete and add news words to the original tweet (Boyd et al. 2010). Due to the nature of our data we included the variation in the cosine similarity to capture the dynamics of the changes in textual contents to capture the impact of past morphing behavior of tweets on the same topic on the current tweet.

Practical Implications

This study has several practical implications. First, our research provides not only a better understanding of the morphing behavior of false news and correction messages but also provides insights on how sentiments affect morphing on social media. Understanding how both false news and correction messages evolve over time and how emotions affect such mutations may help us further our understanding on how fake news influences user behavior over time. Furthermore, this study can be applied to any communication process and helps us further understand the role veracity plays in the transfer of emotions on social media. This study also shows that it may be possible to measure the transfer of emotions by using information morphing as a proxy. Our results show that users change the textual contents of messages aggressively when the contents are emotionally charged. To minimize the aggressive nature of tweets, social media administrators need to tamper the original tweets with more neutral tweets to reduce aggression but not to reduce or dilute the true meanings of the original posts. This may be beneficial in reducing the falsification and misrepresentation of messages as they flow through a network. Furthermore, understanding how the length of tweets or textual contents impact the morphing of messages may help social media administrators better design their platforms. For example, limiting the number of words and/or characters allowed in a review platform to a bare minimum may help users express their dis/pleasure in a more a constructive manner without enabling toxicity. A typical example is Twitter switching from 140 to 280 characters in 2017 and later finding out users did not really need the entire 280 characters to express their emotions (Moon 2018). Rather, the amount of emotional salutations increased but did not change the meaning or context. Our results showed that morphing increases with an increase in textual contents, therefore limiting the number of characters in those platforms can help create a safer environment devoid of toxicity.

Second, social media administrators leveraging this framework can monitor and control the overflow of negative emotions that has the potential of becoming toxic over time. They can do this by limiting the duration of negative interactions on their platforms such as muting forums after an intense period of engagement. Our study shows that as time goes by, both positive and negative sentiments cause an increase in morphing. This may be used as a proxy for measuring and setting thresholds on the appropriate levels of toxicity that is allowed, and beyond this has the potential of causing disruptions in an otherwise conducive and productive environment if such negative engagements are allowed to persist.

Third, social media administrators and government agencies can combat the spread of falsehoods and rumors alike by designing and deploying more positively charged correction messages to counter their spread. Considering our results showed that positively charged tweets morph more than falsehoods and correction tweets influence morphing more, a possible solution to combating the spread of falsehoods on social media is that government agencies and social platform administrators design and deploy effective positively charged correction messages that morphs more to counter and possibly dampen the spread and morphing of falsehoods on social media. This would ultimately increase the virality of positive news and have a ripple effect in encouraging positive emotions on the social media platforms.

Fourth, the adaptation of this framework by content creators, advertisers and marketing executives as a marketing strategy may help strengthen their overall approach to the marketing mix. This approach (positivity) when used in advertising may influence more sharing behavior and potentially impact profitability. It would also help users develop a more lasting positive view of the organization as studies have showed that people inherently like to be associated by positivity in their everyday lives (Di Muro and Murray 2012). Social media users may adopt this strategy and promote more real and positive news to help serve as a catalyst in spreading and transferring positive energy using their online social media presence. This may overall reduce the public’s reliance on falsehoods which are considered inherently negative and lacing in substance.

Finally, the steady morphing of information over time may indicate normality but a rapid increase in morphing behavior may indicate other more serious underlying negative issues such as dissatisfaction towards a policy, service or product or even intentional manipulation by competitors. This could therefore be used as an early warning signal in the arrest of potentially negative security behaviors at the organizational level. Furthermore, as the morphing continues so does the repetition and spread. Social media administrators and government agencies may be able to use this to monitor and address issues that has the potential to promote dissent at the state, public or organizational level

Conclusion

This study investigates the differences in morphing behavior for both correction messages and falsehoods on Twitter. Using a combination of textual analysis and econometric techniques, we show the morphing hazard rates of both falsehoods and correction messages on Twitter. Our empirical results reveal that correction messages, positive tweets and emotionally charged tweets morph faster. Furthermore we show that tweets with positive sentiment or are emotionally charged morph faster over time. Word count and past morphing history also positively affect morphing behavior

CHAPTER V

CONCLUSION

The proliferation of false news contents has created headaches for industry and academic alike. Combating this phenomenon has proven to be quite a challenging task. Even though several studies have attempted to investigate this canker, results have mostly been inconclusive, inconsistent or nonexistent. We attempt to bridge the gap in research by understanding the various factors that may influence the diffusion of information on social media. Furthermore, we investigate the bidirectional effects of falsehoods and correction messages on social media. Lastly, we investigate the differences in morphing behavior for both correction messages and falsehoods on Twitter. Our results show that fake news, novel news, negative news and tweets with lower lexical density propagated more on social media. We also show the that the impacts of sentiment were different for fake news than real news and that environmental tweets were shared less than the baseline tweets. Furthermore, our studies show the counter intuitive nature of current correction endeavors by FEMA and other fact checking organizations in combating falsehoods. Specifically, we show that even though fake news causes an increase in correction messages, they influenced the propagation of falsehoods. Using a combination of textual analysis and econometric techniques, we show the morphing hazard rates of both falsehoods and correction messages on Twitter. Our empirical results reveal that correction messages, positive tweets and emotionally charged tweets morph faster. Furthermore we show that tweets with positive sentiment or are emotionally charged morph faster over time. Word count and past morphing history also positively affect morphing behavior.

.

REFERENCES

Abbasi, A., Zhou, Y., Deng, S., and Zheng, P. 2018. "Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective," MIS Quarterly (42:2), pp. 427–464. (https://doi.org/10.25300/MISQ/2018/13239).

Abramowitz, M., and Stegun, I. . A. 1964. Handbook of Mathematical Functions, (10th ed., Vol. 10), 55, Washington D.C., District of Columbia.

Abrigo, M. R. M., and Love, I. 2015. "Estimation of Panel Vector Autoregression in Stata: A Package of Programs," in 2015 International Panel Data Conference, Central European University, Budapest, Hungary, p. 29.

Abrigo, M. R. M., and Love, I. 2016. "Estimation of Panel Vector Autoregression in Stata," The Stata Journal: Promoting Communications on Statistics and Stata (16:3), pp. 778–804. (https://doi.org/10.1177/1536867X1601600314).

Adomavicius, G., Bockstedt, J., and Gupta, A. 2012. "Modeling Supply-Side Dynamics of IT Components, Products, and Infrastructure: An Empirical Analysis Using Vector Autoregression," Information Systems Research (23:2), pp. 397–417. (https://doi.org/10.1287/isre.1120.0418).

Agrawal, M., Rao, H. R., and Oh, O. 2013. "Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises," MIS Quarterly (37:2), pp. 407–426. (https://doi.org/10.25300/MISQ/2013/37.2.05).

Albuquerque, P. H. M., do Valle, D. R., and Li, D. 2019. "Bayesian LDA for Mixed-Membership Clustering Analysis: The Rlda Package," Knowledge-Based Systems (163), pp. 988–995. (https://doi.org/10.1016/j.knosys.2018.10.024).

Aldenderfer, M. S., and Blashfield, R. K. 1984. "Cluster Analysis," Sage, Thousand Oaks.

Allcott, H., and Gentzkow, M. 2017. "Social Media and Fake News in the 2016 Election," The Journal of Economic Perspectives (31:2), pp. 211–235.

Allcott, H., Gentzkow, M., and Yu, C. 2018. "Trends in the Diffusion of Misinformation on Social Media," ArXiv:1809.05901 [Cs, Econ, q-Fin]. (http://arxiv.org/abs/1809.05901).

Allport, G. W., and Postman, L. 1946. "An Analysis of Rumor," Public Opinion Quarterly (10:4), p. 18. (https://doi.org/10.1093/poq/10.4.501).

Andrews, D. W. K., and Lu, B. 2001. "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," Journal of Econometrics (101:1), pp. 123–164. (https://doi.org/10.1016/S0304-4076(00)00077-4).

Aral, S., and Dhillon, P. 2016. "Unpacking Novelty: The Anatomy of Vision Advantages," MIT. (http://dx.doi.org/10.2139/ssrn.2388254).

Astapova, A. 2017. "In Search for Truth: Surveillance Rumors and Vernacular Panopticon in Belarus," The Journal of American Folklore (130:517), p. 276. (https://doi.org/10.5406/jamerfolk.130.517.0276).

Balijepally, V., Iyengar, K., and Mangalaraj, G. 2011. "Are We Wielding This Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research," Journal of the Association for Information Systems (12:5), pp. 375–413. (https://doi.org/10.17705/1jais.00266).

Barthel, M., Mitchell, A., and Holcomb, J. 2016. "Many Americans Believe Fake News Is Sowing Confusion," Pew Research Center (15). (https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/).

Bauer, R. A., and Gleicher, D. B. 1953. "Word-of-Mouth Communication in the Soviet Union," PUBLIC OPINION QUARTERLY, p. 15.

Bello-Orgaz, G., Jung, J. J., and Camacho, D. 2016. "Social Big Data: Recent Achievements and New Challenges," Information Fusion (28), pp. 45–59. (https://doi.org/10.1016/j.inffus.2015.08.005).

Bene, M. 2017. "Go Viral on the Facebook! Interactions between Candidates and Followers on Facebook during the Hungarian General Election Campaign of 2014," Information, Communication & Society (20:4), pp. 513–529. (https://doi.org/10.1080/1369118X.2016.1198411).

Berger, J., and Milkman, K. 2010. Social Transmission, Emotion, and the Virality of Online Content.

Berger, J., and Milkman, K. L. 2012. "What Makes Online Content Viral?," Journal of Marketing Research (49:2), pp. 192–205. (https://doi.org/10.1509/jmr.10.0353).

Blei, D. M. 2003. Latent Dirichlet Allocation, p. 30.

Botha, E. 2014. "A Means to an End: Using Political Satire to Go Viral," Public Relations Review (40:2), pp. 363–374. (https://doi.org/10.1016/j.pubrev.2013.11.023).

Bowler, M., Halbesleben, J., Stodnick, M., Seevers, M. T., and Little, L. M. 2009. "The Moderating Effect of Communication Network Centrality on Motive to Perform Interpersonal Citizenship Author(s):," Journal of Managerial Issues (21:1), pp. 80–96.

Boyd, D., Golder, S., and Lotan, G. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in 2010 43rd Hawaii International Conference on System Sciences, Honolulu, HI: IEEE, January, pp. 1–10. (https://doi.org/10.1109/HICSS.2010.412).

Brose, U. 2008. "Complex Food Webs Prevent Competitive Exclusion among Producer Species," Proceedings of the Royal Society B: Biological Sciences (275:1650), pp. 2507–2514. (https://doi.org/10.1098/rspb.2008.0718).

Brose, U. 2008. "Complex Food Webs Prevent Competitive Exclusion among Producer Species," Proceedings of the Royal Society B: Biological Sciences (275:1650), pp. 2507–2514. (https://doi.org/10.1098/rspb.2008.0718).

Brown, J. H. 1971. "Mechanisms of Competitive Exclusion Between Two Species of Chipmunks," Ecology (52:2), pp. 305–311. (https://doi.org/10.2307/1934589).

Cameron, K., Rooff, M., Friesema, E., Brown, T., Jonanovic, B., Hauber, S., and Baker, D. 2013. "Patient Knowledge and Recall of Health Information Following Exposure to 'Facts and Myths' Message Format Variations," Patient Education and Counseling (92), pp. 381–387. (https://doi.org/10.1016/j.pec.2013.06.017).

Chen, X. 2016. The Influences of Personality and Motivation on the Sharing of Misinformation on Social Media, iSchools, March 10. (https://doi.org/10.9776/16145).

Cheng, J.-J., Liu, Y., Shen, B., and Yuan, W.-G. 2013. “An Epidemic Model of Rumor Diffusion in Online Social Networks,” EDP Sciences, Societ`a Italiana Di Fisica, Springer-Verlag 2013. (https://doi.org/10.1140/epjb/e2012-30483-5 An epidemic).

Chua, A. Y. K., Tee, C.-Y., Pang, A., and Lim, E.-P. 2017. “The Retransmission of Rumor and Rumor Correction Messages on Twitter,” American Behavioral Scientist (61:7), pp. 707–723. (https://doi.org/10.1177/0002764217717561).

Deng, S., Huang, Zhijan. J., Zhao, H., and Sinha, A. 2018. “The Interaction Between Microblog Sentiment and Stock Returns: An Empirical Examination,” MIS Quarterly (42:3), pp. 895–918. (https://doi.org/10.25300/MISQ/2018/14268).

Di Muro, F., and Murray, K. B. 2012. “An Arousal Regulation Explanation of Mood Effects on Consumer Choice,” Journal of Consumer Research (39:3), pp. 574–584. (https://doi.org/10.1086/664040).

DiFonzo, N., and Bordia, P. 2007. “Rumor, Gossip and Urban Legends,” Diogenes (54:1), pp. 19–35. (https://doi.org/10.1177/0392192107073433).

Dunn, H. B., and Allen, Charlotte. A. 2005. “RUMORS, URBAN LEGENDS AND INTERNET HOAXES,” Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators, p. 7.

Ecker, U. K. H., Lewandowsky, S., and Apai, J. 2011. “Terrorists Brought down the Plane!—No, Actually It Was a Technical Fault: Processing Corrections of Emotive Information,” Experimental Psychology (64:2). (https://doi.org/10.1080/17470218.2010.497927).

Ecker, U. K. H., Lewandowsky, S., and Tang, D. T. W. 2010. “Explicit Warnings Reduce but Do Not Eliminate the Continued Influence of Misinformation,” Memory & Cognition (38:8), pp. 1087–1100. (https://doi.org/10.3758/MC.38.8.1087).

Ehrenberg, R. 2012. “Social Media Sway: Worries over Political Misinformation on Twitter Attract Scientists’ Attention,” Science News (182:8), pp. 22–25. (https://doi.org/10.1002/scin.5591820826).

Faraj, S., Kudaravalli, S., HEC Paris, Wasko, M., and University of Alabama at Birmingham. 2015. “Leading Collaboration in Online Communities,” MIS Quarterly (39:2), pp. 393–412. (https://doi.org/10.25300/MISQ/2015/39.2.06).

Friggeri, A., Adamic, L., Eckles, D., and Cheng, J. 2014. “Rumor Cascades,” in Eighth International AAAI Conference on Weblogs and Social Media, AAAI Publications, pp. 101–110. (https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8122/8110).

Gause, G. F. 1932. “Experimental Studies on the Struggle for Existence,” Journal of Experimental Biology (9:1), pp. 389–402.

Hair, J. B., Babin, B. J., Anderson, R. E., Black, W. C., and Tatham, R. L. 2016. Multivariate Data Analysis, (Sixth.), Upper Saddles River, NJ: Pearson Prentice-Hall.

Halliday, M. A. K. 1989. Spoken and Written Language, (2nd ed.), Oxford University Press.

Harris, R. B., and Paradice, D. 2007. “An Investigation of the Computer-Mediated Communication of Emotions,” INSInet Publication (3:12), p. 11.

Hoang, T. B. N., and Mothe, J. 2018. “Predicting Information Diffusion on Twitter – Analysis of Predictive Features,” Journal of Computational Science (28), pp. 257–264. (https://doi.org/10.1016/j.jocs.2017.10.010).

Holtz-Eakin, D., Newey, W., and Rosen, H. S. 1988. "Estimating Vector Autoregressions with Panel Data," Econometrica (56:6), p. 1371. (https://doi.org/10.2307/1913103).

Honeycutt, C., and Herring, S. C. 2009. "Beyond Microblogging: Conversation and Collaboration via Twitter," in 2009 42nd Hawaii International Conference on System Sciences, Waikoloa, Hawaii, USA: IEEE, pp. 1–10. (https://doi.org/10.1109/HICSS.2009.89).

Hovland, C. 1959. "Reconciling Conflicting Results Derived from Experimental and Survey Studies of Attitude Change.," American Psychological Association (14:1), pp. 8–17. (http://dx.doi.org/10.1037/h0042210).

Huang, H. 2017. "A War of (Mis)Information: The Political Effects of Rumors and Rumor Rebuttals in an Authoritarian Country," British Journal of Political Science (47:02), pp. 283–311. (https://doi.org/10.1017/S0007123415000253).

Huei Chou, C., Sinha, A., and Zhao, H. 2010. "A Hybrid Attribute Selection Approach for Text Classification," Journal of the Association for Information Systems (11:9), pp. 491–518. (https://doi.org/10.17705/1jais.00236).

Itti, L., and Baldi, P. 2009. "Bayesian Surprise Attracts Human Attention," Vision Research (49:10), pp. 1295–1306. (https://doi.org/10.1016/j.visres.2008.09.007).

Jaeger, R. G. 1974. "Competitive Exclusion: Comments on Survival and Extinction of Species," BioScience (24:1), p. 33. (https://doi.org/10.2307/1296657).

Jin, F., Dougherty, E., Saraf, P., Cao, Y., and Ramakrishnan, N. 2013. "Epidemiological Modeling of News and Rumors on Twitter," The 7th SNA-KDD Workshop.

Jindal, N., Liu, B., and Lim, E.-P. 2010. "Finding Unusual Review Patterns Using Unexpected Rules," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10, Toronto, ON, Canada: ACM Press, p. 1549. (https://doi.org/10.1145/1871437.1871669).

Jolley, D., and Douglas, K. M. 2014. "The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions," PLoS ONE (9:2), (R. Tripp, ed.), p. e89177. (https://doi.org/10.1371/journal.pone.0089177).

Jonah Berger. 2011. "Arousal Increases Social Transmission of Information," Psychological Science (22:7), pp. 891–893. (https://doi.org/DOI: 10.1177/0956797611413294).

Karimi, H., Roy, P., Saba-Sadiya, S., and Tang, J. 2018. "Multi-Source Multi-Class Fake News Detection," in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico: International Conference on Computational Linguistics, p. 12.

Kazienko, P., and Chawla, N. 2015. Applications of Social Media and Social Network Analysis, 2190-5436, Switzerland: Springer, Cham. (https://doi.org/10.1007/978-3-319-19003-7).

Ketchen Jr., D. J., and Shook, C. L. 1996. "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique.," Strategic Management Journal (17:6), pp. 441–458. (https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G).

Killins, R. N., Egly, P. V., and Escobari, D. 2017. "The Impact of Oil Shocks on the Housing Market: Evidence from Canada and U.S," Journal of Economics and Business (93), pp. 15–28. (https://doi.org/10.1016/j.jeconbus.2017.07.002).

Kim, A., and Dennis, A. R. 2017. “Says Who?: How News Presentation Format Influences Perceived Believability and the Engagement Level of Social Media Users,” SSRN Electronic Journal. (https://doi.org/10.2139/ssrn.2987866).

Kim, A., Moravec, P. L., and Dennis, A. R. 2019. “Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings,” Journal of Management Information Systems (36:3), pp. 931–968. (https://doi.org/10.1080/07421222.2019.1628921).

Kim, J., Bae, J., and Hastak, M. 2018. “Emergency Information Diffusion on Online Social Media during Storm Cindy in U.S.,” International Journal of Information Management (40), pp. 153–165. (https://doi.org/10.1016/j.ijinfomgt.2018.02.003).

King, K. K., and Sun, J. 2018. “Investigating User Disclosure of Sensitive Information: An ELM Theory,” in 24th Americas Conference on Information Systems (Vol. 1), New Orleans: AIS Electronic Library, p. 12.

King, K. K., and Sun, J. 2019. “Catch Bots with a Bot: An Automated Approach to Misinformation Detection,” in 4th International Conference on Design Science Research in Information Systems and Technology, Worcester, MA: Springer, May, p. 6.

King, K. K., and Wang, B. 2019. Diffusion of False versus Real News on Social Media: An Analysis of Twitter.

Koenig, F. 1985. Rumor in the Marketplace: The Social Psychology of Commercial Hearsay, (1st ed.), Dover, MA: Auburn House: Praeger.

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. “Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning,” Journal of Management Information Systems (35:1), pp. 350–380. (https://doi.org/10.1080/07421222.2018.1440758).

Lakoff, G., Dean, H., and Hazen, D. 2004. George Lakoff -The Essential Guide for Progressives, (1st ed.), White River Junction, Vermont: Chelsea Green Publishing.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. 2018. “The Science of Fake News,” Science (359:6380), pp. 1094–1096. (https://doi.org/10.1126/science.aao2998).

Lee, K., Mahmud, J., Chen, J., Zhou, M., and Nichols, J. 2014. Who Will Retweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information, p. 10.

Lee, T.-H., and Crompton, J. 1992. “Measuring Novelty Seeking in Tourism,” Annals of Tourism Research (19:4), pp. 732–751. (https://doi.org/10.1016/0160-7383(92)90064-V).

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. 2012. “Misinformation and Its Correction: Continued Influence and Successful Debiasing,” Psychological Science in the Public Interest (13:3), pp. 106–131. (https://doi.org/10.1177/1529100612451018).

Liang, N. (Peter), Biros, D. P., and Luse, A. 2016. “An Empirical Validation of Malicious Insider Characteristics,” Journal of Management Information Systems (33:2), pp. 361–392. (https://doi.org/10.1080/07421222.2016.1205925).

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. 2016. Fake News Detection Through Multi-Perspective Speaker Profiles, p. 5.

Love, I., and Zicchino, L. 2006. “Financial Development and Dynamic Investment Behavior: Evidence from Panel VAR,” The Quarterly Review of Economics and Finance (46:2), pp. 190–210. (https://doi.org/10.1016/j.qref.2005.11.007).

Luo, X., and Zhang, J. 2013. “How Do Consumer Biuzz and Traffic in Social Media Marketing Predict the Value of the Firm.,” Journal of Management Information Systems (30:2), pp. 213–238.

Luo, X., Zhang, J., and Duan, W. 2013. “Social Media and Firm Equity Value,” Information Systems Research (24:1), pp. 146–163. (https://doi.org/10.1287/isre.1120.0462).

Lütkepohl, H. 2005. New Introduction to Multiple Time Series Analysis, Berlin: New York : Springer.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M. 2016. “Detecting Rumors from Microblogs with Recurrent Neural Networks,” in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), New York, USA, pp. 3818–3824.

Malhotra, A. 2018. “Introduction to Libraries of NLP in Python — NLTK vs. SpaCy,” Medium, , June. (https://medium.com/@akankshamalhotra24/introduction-to-libraries-of-nlp-in-python-nltk-vs-spacy-42d7b2f128f2, accessed November 24, 2018).

McGuire, W. . J. 1964. “Inducing Resistance to Persuasion. Some Contemporary Approaches,” Advances in Experimental Social Psychology (1), pp. 191–229. (https://doi.org/10.1016/S0065-2601(08)60052-0).

McPeek, M. A. 2014. “Limiting Factors, Competitive Exclusion, and a More Expansive View of Species Coexistence,” The American Naturalist (183:3), iii–iv. (https://doi.org/10.1086/675305).

Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. 2017. “A Survey on Information Diffusion in Online Social Networks: Models and Methods,” Information (8:4), p. 118. (https://doi.org/10.3390/info8040118).

Nesi, H. 2001. “A Corpus-Based Analysis of Academic Lectures across Disciplines,” Continuum Press, Language across Boundaries, p. 25.

Nesia, B. H., and Ginting, S. A. 2014. “Lexical Density of Englis Reading Texts For Senior High School,” Langauge Teaching and Learning FBS, p. 14.

Oh, O., Kwon, K. H., and Rao, H. R. 2010. “An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010.,” in ICIS 2010 Proceedings 231, Saint Louis, Missouri, USA,: AIS Electronic Library, pp. 1–13.

Oliver, J. E., and Wood, T. J. 2014. “Conspiracy Theories and the Paranoid Style(s) of Mass Opinion: CONSPIRACY THEORIES AND MASS OPINION,” American Journal of Political Science (58:4), pp. 952–966. (https://doi.org/10.1111/ajps.12084).

Osatuyi, B., and Hughes, J. 2018. “A Tale of Two Internet News Platforms-Real vs. Fake: An Elaboration Likelihood Model Perspective,” in Proceedings of the 51st Hawaii International Conference on System Sciences, Hawaii, pp. 3986–3994. (URI: http://hdl.handle.net/10125/50388).

Ozturk, P., Li, H., and Sakamoto, Y. 2015. “Combating Rumor Spread on Social Media: The Effectiveness of Refutation and Warning,” in 2015 48th Hawaii International Conference on System Sciences, HI, USA: IEEE, January, pp. 2406–2414. (https://doi.org/10.1109/HICSS.2015.288).

Pastor-Satorras, R., and Vespignani, A. 2001. “Epidemic Spreading in Scale-Free Networks,” Physical Review Letters (86:14), pp. 3200–3203. (https://doi.org/10.1103/PhysRevLett.86.3200).

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. 2017. “Automatic Detection of Fake News,” ArXiv:1708.07104 [Cs]. (http://arxiv.org/abs/1708.07104).

Punj, G., and Steward, D. W. 1983. Cluster Analysis in Marketing Research: Review and Suggestions for Application., p. 16.

Rácz, É. V. P., and Karsai, J. 2006. “The Effect of Initial Pattern on Competitive Exclusion,” Community Ecology (7:1), pp. 23–33. (https://doi.org/10.1556/ComEc.7.2006.1.3).

Reuters. 2017. “Most American Adults Get News From Social Media,” Tech 40 under 40, , September 8. (fortune.com/2017/09/08/facebook-twitter-snap-news/, accessed May 7, 2018).

Rodrıguez, G. 2013. Models for Count Data With Overdispersion, p. 7.

Rosenstiel, T., Sonderman, J., Locker, K., Ivancin, M., and Kjarval, N. 2015. Twitter and the News: How People Use the Social Network to Learn about the World, p. 44.

Schwarz, N., Sanna, L. J., Skurnik, I., and Yoon, C. 2007. “Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns,” in Advances in Experimental Social Psychology (Vol. 39), Elsevier, pp. 127–161. (https://doi.org/10.1016/S0065-2601(06)39003-X).

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., and Menczer, F. 2018. “The Spread of Low-Credibility Content by Social Bots,” Nature Communications (9:1), p. 4787. (https://doi.org/10.1038/s41467-018-06930-7).

Shin, J., Jian, L., Driscoll, K., and Bar, F. 2018. “The Diffusion of Misinformation on Social Media: Temporal Pattern, Message, and Source,” Computers in Human Behavior (83), pp. 278–287. (https://doi.org/10.1016/j.chb.2018.02.008).

Shu, L., Long, B., and Meng, W. 2009. “A Latent Topic Model for Complete Entity Resolution,” in 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China: IEEE, March, pp. 880–891. (https://doi.org/10.1109/ICDE.2009.29).

Silverman, C., and Singer-Vine, J. 2016. “Most Americans Who See Fake News Believe It, New Survey Says,” BuzzFeed News, , December 6. (https://www.buzzfeed.com/craigsilverman/fake-news-survey?utm_term=.bhnzwog00G#.hjpN2eMqqg, accessed May 7, 2018).

SMWG_Countering-False-Info-Social-Media-Disasters-Emergencies. 2018. DHS.

Spenkuch, Jj. L., and Toniatti, D. 2015. “Political Advertising and Election Outcomes,” SSRN Electronic Journal. (https://doi.org/10.2139/ssrn.2613987).

Stieglitz, S., and Dang-Xuan, L. 2013. “Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior,” Journal of Management Information Systems (29:4), pp. 217–248. (https://doi.org/10.2753/MIS0742-1222290408).

Sun, Q., Li, R., Luo, D., and Wu, X. 2008. Text Segmentation with LDA-Based Fisher Kernel, p. 4.

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., and Takayasu, H. 2015. “Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study,” PLOS ONE (10:4), (R. M. H. Merks, ed.), p. e0121443. (https://doi.org/10.1371/journal.pone.0121443).

Tear, A., and Southall, H. 2019. “Social Media Data.Pdf,” in Data in Society: Challenging Statistics in an Age of Globalisation (1st ed.), Bristol University Press, pp. 47–60. (https://www.jstor.org/stable/j.ctvmd84wn.12).

To, V., Fan, S., and Thomas, D. 2013. "Lexical Density and Readability: A Case Study of English Textbooks," Internet Journal of Language, Culture and Society (37), pp. 61–71. (https://doi.org/327-774X).

Torres, R. R., Gerhart, N., and Negahban, A. 2018. Combating Fake News: An Investigation of Information Verification Behaviors on Social Networking Sites, p. 10.

Tsou, M.-H. 2015. "Research Challenges and Opportunities in Mapping Social Media and Big Data," Cartography and Geographic Information Science (42:sup1), pp. 70–74. (https://doi.org/10.1080/15230406.2015.1059251).

Tversky, A., and Kahneman, D. 1986. "Rational Choice and the Framing of Decisions," The University of Chicago Press (59:No. 4 part 2 The Behavioral Foundations of Economic Theory), The Behavioral Foundations of Economic Theory, p. 29.

Twitter. 2019. Twitter Objects, Twitter. (https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html).

Ure, J. 1971. "Lexical Density and Register Differentiation," in Applications of Linguistics, pp. 443–452.

Vosoughi, S., Roy, D., and Aral, S. 2018. "The Spread of True and False News Online," Science (359:6380), pp. 1146–1151. (https://doi.org/10.1126/science.aap9559).

Walter, N., and Murphy, S. T. 2018. "How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation," Communication Monographs (85:3), pp. 423–441. (https://doi.org/10.1080/03637751.2018.1467564).

Wang, W. Y. 2017. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," ArXiv:1705.00648 [Cs]. (http://arxiv.org/abs/1705.00648).

Wei, C.-P., Hu, P. J.-H., and Lee, Y.-H. 2009. "Preserving User Preferences in Automated Document-Category Management: An Evolution-Based Approach," Journal of Management Information Systems (25:4), pp. 109–144. (https://doi.org/10.2753/MIS0742-1222250404)

APPENDIX A

APPENDIX A

DERIVATION OF VAR

Using μ_i Using μ_i as the rate of decay (mortality) of message (species) i ($i = 1,2$), K as the number of users, $N_i(t)$ as the number of tweets and retweets at time t with fraction $n_i(t) = \frac{N_i(t)}{K}$, r_i as the per-capita growth rate of message (species) i ($i = 1,2$), and $\beta_{i,j}$ as the competition rate of specious j to i ($i\ , j = 1,2;\ i\ \neq j$), we have:

1)

$$\frac{d\,N_1}{dt} = r_1 N_1 \left[(1 - \frac{N_1}{K}) - \beta_{12}\frac{N_2}{K}\ \ \right] - \mu_1 N_1$$

$$\frac{d\,N_2}{dt} = r_2 N_2 \left[(1 - \frac{N_2}{K}) - \beta_{21}\frac{N_1}{K}\ \ \right] - \mu_2 N_2$$

Thus,

$\frac{d\,n_1(t)}{dt} = r_1 n_1(t)\left[(1 - \frac{\mu_1}{r_1}) - n_1(t) - \beta_{12}\, n_2(t)\right]$	(1)
$\frac{d\,n_2(t)}{dt} = r_2 n_2(t)\left[(1 - \frac{\mu_2}{r_2}) - n_2(t) - \beta_{21}\, n_1(t)\right]$	(2)

Next, we will perform a stability analysis of the system of equations (1-2).

Stability Analysis

At stability of the fraction of the two species $\frac{d\,n_i(t)}{dt}$ =0. Thus, that ensuing system of algebraic equations has the following four possible solutions (that are called equilibria):

1) n_1 = 0, and n_2 = 0;

2) n_1 = 0, and n_2 = $(1-\frac{\mu_2}{r_2}) = R_2$, which exists only if $\mu_2 < r_2$;

3) n_1 = $(1-\frac{\mu_1}{r_1}) = R_1$, and n_2 = 0, which exists only if μ_1 < r_1; or

4) $n_1 = \frac{\left(1-\frac{\mu_1}{r_1}\right)-\beta_{12}\left(1-\frac{\mu_2}{r_2}\right)}{1-\beta_{12}\,\beta_{21}} = \frac{R_1-\beta_{12}R_2}{1-\beta_{12}\,\beta_{21}}$, and $n_2 = \frac{\left(1-\frac{\mu_2}{r_2}\right)-\beta_{21}\left(1-\frac{\mu_1}{r_1}\right)}{1-\beta_{21}\,\beta_{12}} = \frac{R_2-\beta_{21}R_1}{1-\beta_{21}\,\beta_{12}}$, which exists only when $0 \leq n_1, n_2 \leq 1$.

To study the local stability of the equilibria of the system, let

$$f_1(r_1, R_1, \beta_{12}) = r_1 n_1 [R_1 - n_1 - \beta_{12}\, n_2] \tag{3}$$

$$f_2(r_2, R_2, \beta_{21}) = r_2 n_2 [R_2 - n_2 - \beta_{21}\, n_1] \tag{4}$$

Then the Jacobian matrix of the system is given by:

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial n_1} & \frac{\partial f_1}{\partial n_2} \\ \frac{\partial f_2}{\partial n_1} & \frac{\partial f_2}{\partial n_2} \end{bmatrix} = \begin{bmatrix} r_1[R_1 - 2n_1 - \beta_{12}\, n_2] & -r_1\beta_{12}\, n_1 \\ -r_2\beta_{21}\, n_2 & r_2[R_2 - 2n_2 - \beta_{21}\, n_1] \end{bmatrix}.$$

The first equilibrium n_1 = 0, and n_2 = 0 is always not stable since the Jacobian

$$J_1 = \begin{bmatrix} r_1 R_1 & 0 \\ 0 & r_2 R_2 \end{bmatrix},$$

which is only stable if the eigenvalues of J_1 are negative.

That is, when $\mu_1 > r_1$ and $\mu_2 > r_2$ or when the two messages die faster than they grow.

For the second equilibrium, n_1 = 0, and $n_2 = R_2 > 0$

$$J_2 = \begin{bmatrix} r_1[R_1 - \beta_{12}\, R_2] & 0 \\ -r_2\beta_{21}\, R_2 & -r_2 R_2 \end{bmatrix},$$

which is stable only if $\frac{R_1}{R_2} < \beta_{12}$ and is when the rebuttal message is aggressive enough to put an end to the falsehood. The third equilibrium, $n_1 = R_1$, and $n_2 = 0$ with a Jacobian J_3 is similarly stable when $\frac{R_2}{R_1} < \beta_{21}$ as the falsehood is stronger than that threshold of $\frac{R_2}{R_1}$. These two equilibria are the competitive exclusion cases in which one of the two message puts an end to the other. The last equilibrium is the co-existence of the two messages when $n_1 = \frac{R_1 - \beta_{12} R_2}{1 - \beta_{12}\,\beta_{21}}$, and $n_2 = \frac{R_2 - \beta_{21} R_1}{1 - \beta_{21}\,\beta_{12}}$. It is stable when

$$J_4 = \begin{bmatrix} r_1\left[R_1 - 2\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}} - \beta_{12}\,\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}}\right] & -r_1\beta_{12}\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}} \\ -r_2\beta_{21}\,\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}} & r_2\left[R_2 - 2\frac{R_2 - \beta_{21}R_1}{1 - \beta_{21}\,\beta_{12}} - \beta_{21}\,\frac{R_1 - \beta_{12}R_2}{1 - \beta_{12}\,\beta_{21}}\right] \end{bmatrix}$$

In this model, we find a scenario where both falsehoods and correction messages either eventually die off or survive after competition.

VAR

Following (Abrigo and Love 2016), we modeled the diffusion of our two time series of falsehoods and correction messages $\mathbf{Y}_{i,t}$ as a (2 x 1) vector of dependent variables including the number of falsehoods diffused and the number of correction messages on topic i; $i = 1,2, \dots ,21$, during hour $t \in \{1, \dots , T_i\}$. The (2 x 1) vector $\boldsymbol{\gamma}_{i,t}$ is made up from dummy variables representing the year or hurricane, and $\boldsymbol{h}_{i,t}$ is a (2 x 23) matrix of dummy variates representing the hour of the day. Finally, $\mathbf{u}_i$ and $\boldsymbol{\epsilon}_{i,t}$ are (2 x 1) vectors of dependent variable-specific panel fixed effect and idiosyncratic errors, respectively. The (2x2) matrices $\mathbf{A}_1$, $\mathbf{A}_2$, . . . , $\mathbf{A}_{P-1}$, $\mathbf{A}_P$ are the parameters to be estimated as recommended by (Abrigo and Love 2015), where P is the lag order (e.g. 1, 2, …) to be estimated empirically. The assumptions are that the innovations may be denoted by: $\mathbf{E}(\boldsymbol{\epsilon}_{i,t})\mathbf{=0}$, $E(\boldsymbol{\epsilon}_{i,t}\;\boldsymbol{\epsilon}_{i,t}{}')$

= $\mathbf{\Sigma}$, and E($\boldsymbol{\epsilon}_{i,t}$ $\boldsymbol{\epsilon}_{i,s}{}'$)= $\mathbf{0}$ for all t > s, which is a white noise multivariate process of our two variables. Following studies by (Holtz-Eakin et al. 1988), we assume that the cross-sectional units share the same underlying data generating process, with the reduced-form parameters $\mathbf{A}_1$, $\mathbf{A}_2$, . . . , $\mathbf{A}_{p-1}$, $\mathbf{A}_p$ common among them. The structural representation can be given as:

$\mathbf{Y}_{i,t} = \mathbf{A}_1\mathbf{Y}_{i,t-1} + \mathbf{A}_2\mathbf{Y}_{i,t-2} + \ldots + \mathbf{A}_{P-1}\mathbf{Y}_{i,t-P+1} + \mathbf{A}_P\mathbf{Y}_{i,t-P} + \boldsymbol{\beta}_1\boldsymbol{\gamma}_{\boldsymbol{i,t}} + \boldsymbol{\beta}_2\boldsymbol{h}_{\boldsymbol{i,t}} + \mathbf{u}_i + \boldsymbol{\epsilon}_{i,t}$ for $t \geq p$.

RELATIONSHIP BETWEEN EMPIRICS AND THEORETICAL EVIDENCE

In this section we derive an approximation of the VAR model

$\mathbf{Y}_{it} = \mathbf{A}_1\mathbf{Y}_{it-1} + \mathbf{A}_2\mathbf{Y}_{it-2} + \ldots + \mathbf{A}_{P-1}\mathbf{Y}_{it-P+1} + \mathbf{A}_P\mathbf{Y}_{it-P} + \boldsymbol{\beta}_1\boldsymbol{\gamma}_{\boldsymbol{it}} + \boldsymbol{\beta}_2\boldsymbol{h}_{\boldsymbol{it}} + \mathbf{u}_i + \boldsymbol{\epsilon}_{i,t}$

from the continuous time model in the system of equations (1-2).

Let $\boldsymbol{X}(t) = \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix}$ be the solution of the system of equations (1-2).

which could be written as

$\frac{d}{dt}\, \boldsymbol{X}(t) = F(\boldsymbol{X}(t))$

where $F(\boldsymbol{X}(t)) = \begin{bmatrix} f_1(\boldsymbol{X}(t)) \\ f_2(\boldsymbol{X}(t)) \end{bmatrix}$. Using forward Euler formula (FE), we can write

$$\boldsymbol{X}(t_{n+1}) = \boldsymbol{X}(t_n) + \Delta t\, F(\boldsymbol{X}(t_n))$$

for $t_1 < t_2 < \cdots < t_K$ that make a step size of $\Delta t = t_i - t_{i-1}$ for all i. FE can be applied iteratively for P times to get

$$\boldsymbol{X}(t_{n+1}) = \boldsymbol{X}\big(t_{n-p}\big) + \Delta t \sum_{i=0}^{P} F(\boldsymbol{X}(t_{n-i}))$$

for $n \geq p$. By linearizing $F(\boldsymbol{X}(t_{n-i}))$ about some point $\boldsymbol{X}_i^*$ close to $\boldsymbol{X}(t_{n-i})$,

$$F(\boldsymbol{X}(t_{n-i})) \approx F(\boldsymbol{X}_i^*\,) + J(\boldsymbol{X}_i^*\,)\,(\boldsymbol{X}(t_{n-i}) - \boldsymbol{X}_i^*)$$

where J is the Jacobian matrix. Thus,

$$\boldsymbol{X}(t_{n+1}) \approx \boldsymbol{X}\big(t_{n-p}\big) + \Delta t \sum_{i=0}^{P}\Big(F(\boldsymbol{X}_i^*) + J(\boldsymbol{X}_i^*)(\boldsymbol{X}(t_{n-i}) - \boldsymbol{X}_i^*)\Big)$$

That will lead to

$$\boldsymbol{X}(t_{n+1}) \approx \Delta t \sum_{i=0}^{P}(F(\boldsymbol{X}_i^*) - J(\boldsymbol{X}_i^*)\boldsymbol{X}_i^*) + \boldsymbol{X}\big(t_{n-p}\big) + \Delta t \sum_{i=0}^{P} J(\boldsymbol{X}_i^*)\boldsymbol{X}(t_{n-i})$$

Calling $\boldsymbol{X}(t_n)$ by $\boldsymbol{Y}_n$, it can be finally written as

$$\boldsymbol{Y}_{n+1} \approx C_p + \sum_{i=0}^{P} A_i \boldsymbol{Y}_{n-i}$$

where $C_p = \Delta t \sum_{i=0}^{P}(F(\boldsymbol{X}_i^*) - J(\boldsymbol{X}_i^*)\boldsymbol{X}_i^*)$ and $A_i = \Delta t\, J(\boldsymbol{X}_i^*)$ for $i = 0, \dots, P-1$ and $A_p = \boldsymbol{I} + \Delta t\, J\big(\boldsymbol{X}_p^*\big)$ where **I** is the 2x2 identity matrix. A forward Euler-Maruyama can then be used to extend that linearization as a way to linearize a stochastic differential equation version of the deterministic system that leads to the noise term in the PVAR. So, in relation to the model in equations (1-2) the matrices are given by

$$A_i = \Delta t \begin{bmatrix} r_1\big[R_1 - 2\, x_{1,i}^* - \beta_{12}\, x_{2,i}^*\big] & -r_1\beta_{12}\, x_{1,i}^* \\ -r_2\beta_{21}\, x_{2,i}^* & r_2\big[R_2 - 2\, x_{2,i}^* - \beta_{21}\, x_{1,i}^*\big] \end{bmatrix}$$

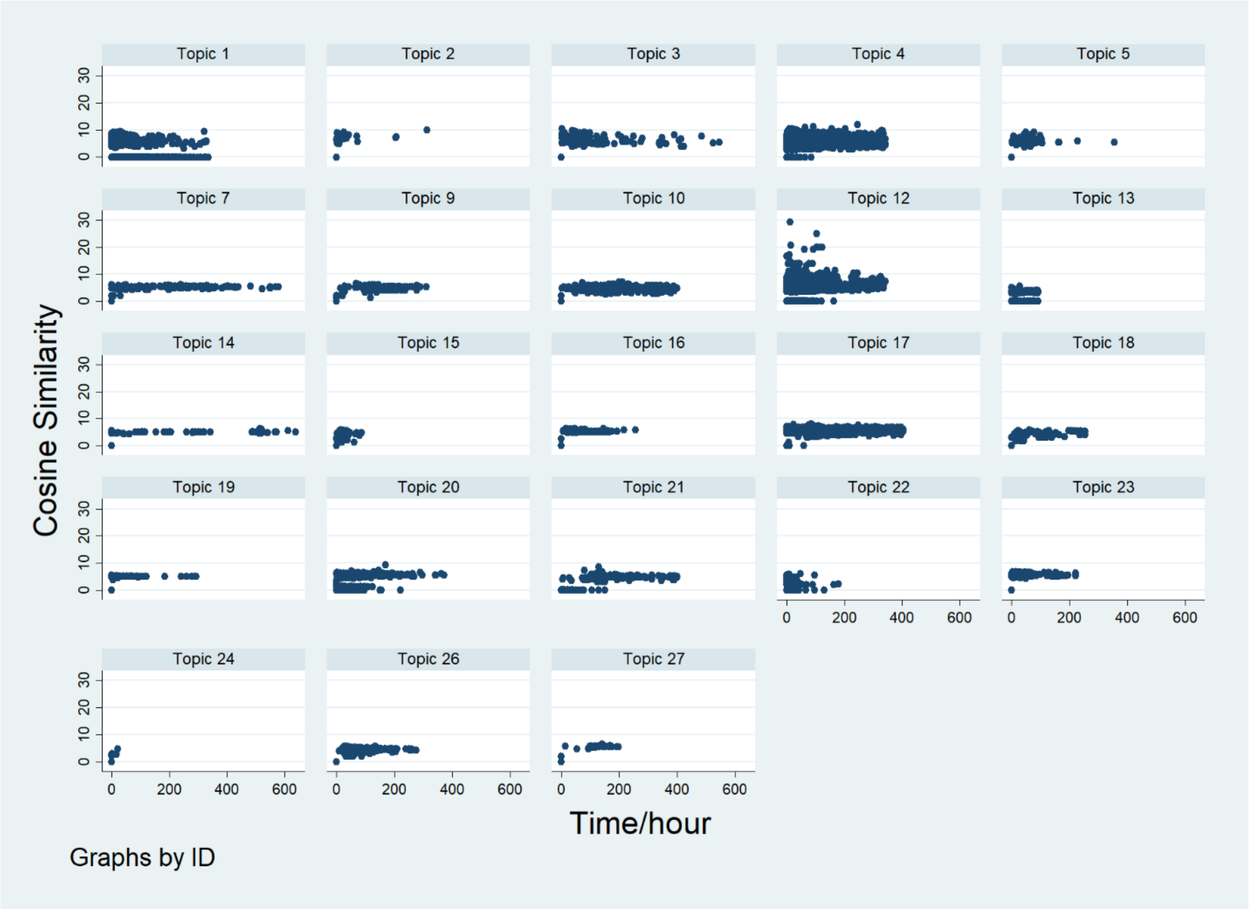
for $i = 0, \dots, P-1$, and

$$A_p = \begin{bmatrix} 1 + \Delta t\, r_1\big[R_1 - 2\, x_{1,P}^* - \beta_{12}\, x_{2,P}^*\big] & -r_1\beta_{12}\, x_{1,P}^* \\ -r_2\beta_{21}\, x_{2,P}^* & 1 + \Delta t\, r_2\big[R_2 - 2\, x_{2,P}^* - \beta_{21}\, x_{1,P}^*\big] \end{bmatrix}.$$

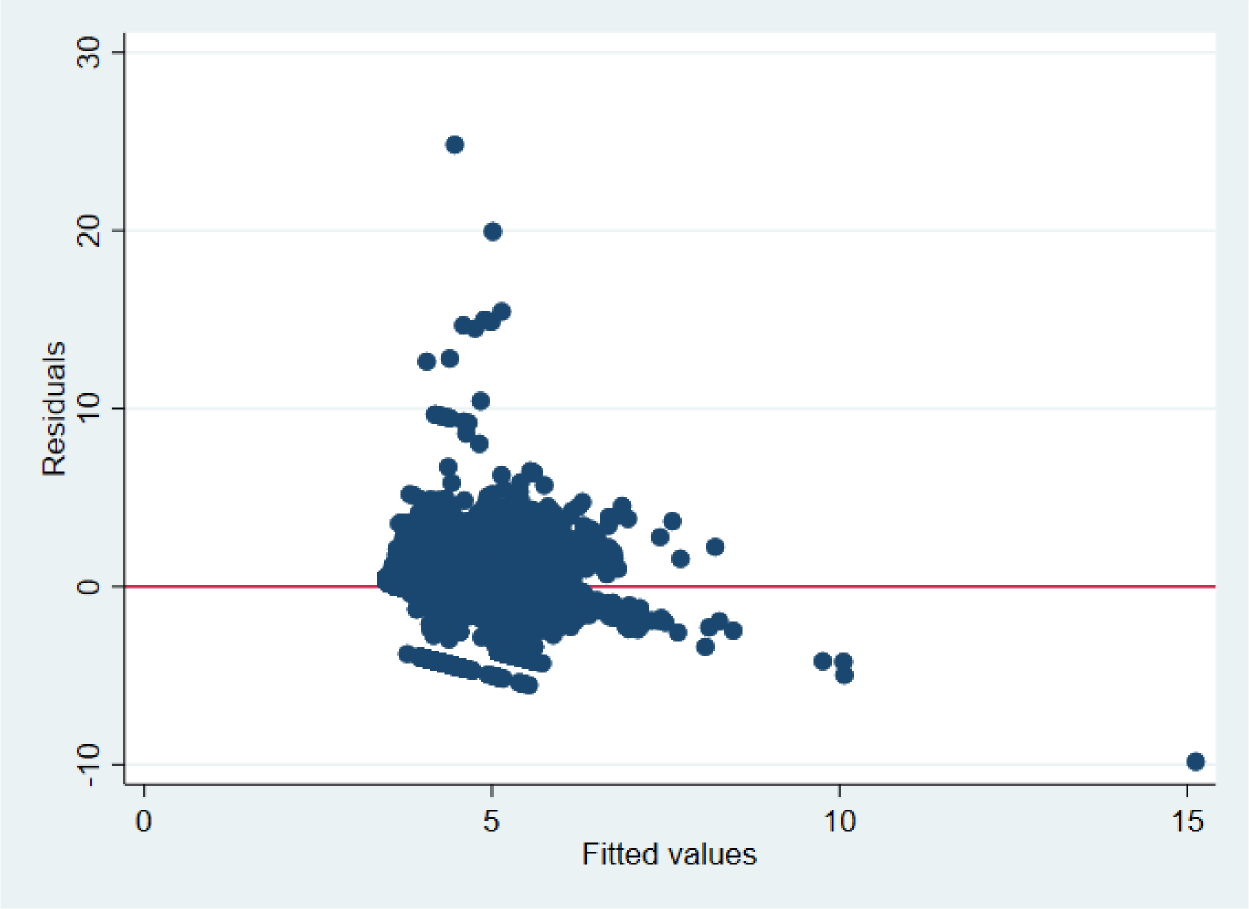
APPENDIX B

APPENDIX B

SCATER PLOTS FOR SIMILARITY



WHITES TEST FOR HETEROSKCEDASTICITY



BIOGRAPHICAL SKETCH

Kelvin Kizito King

University of Texas, Rio Grande Valley

1201 W University Dr., Edinburg, TX 78539

kkk279@nyu.edu

Kelvin Kizito King received his Ph.D. in Information Systems in August 2020 at the University of Texas, Rio Grande Valley. He received his master's degree from New York University in 2015. Kelvin has produced multiple conference research publications and has multiple journal submissions on the diffusion of fake news on social media and mechanisms to detect and counter falsehoods. His research has received many awards including the first place award in the 2018 UTRGV Graduate Research Symposium, the UTRGV Ph.D. Student Excellence Award in Information Systems 2018, a National Science Foundation travel grant in 2019, and the Best Poster Award at the 14th International Conference on Design Science Research in Information Systems and Technology in 2019. At the time pf graduating he was designated a Ph.D. Project – Baruch Fellow.

Kelvin has also acted as a reviewer for several conferences and journals, including AMCIS, HICCS and JMIS. His primary research interests include, information diffusion on social media, health informatics, investigating the dark web and detecting fake and manipulated reviews. His primary tools are machine learning algorithms, econometrics and mathematics. He has industry experience as a data conversion specialist, a data scientist and as a consultant.