Dynamics of Information Distribution on Social Media Platforms during Disasters

by

Eunae Yoo

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2018 by the
Graduate Supervisory Committee:

Elliot Rabinovich, Co-Chair
Bin Gu, Co-Chair
William Rand
John Fowler

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

When preparing for and responding to disasters, humanitarian organizations must run effective and efficient supply chains to deliver the resources needed by the affected population. The management of humanitarian supply chains include coordinating the flows of goods, finances, and information. This dissertation examines how humanitarian organizations can improve the distribution of information, which is critical for the planning and coordination of the other two flows. Specifically, I study the diffusion of information on social media platforms since such platforms have emerged as useful communication tools for humanitarian organizations during times of crisis.

In the first chapter, I identify several factors that affect how quickly information spreads on social media platforms. I utilized Twitter data from Hurricane Sandy, and the results indicate that the timing of information release and the influence of the content's author determine information diffusion speed. The second chapter of this dissertation builds directly on the first study by also evaluating the rate at which social media content diffuses. A piece of content does not diffuse in isolation but, rather, coexists with other content on the same social media platform. After analyzing Twitter data from four distinct crises, the results indicate that other content's diffusion often dampens a specific post's diffusion speed. This is important for humanitarian organizations to recognize and carries implications for how they can coordinate with other organizations to avoid inhibiting the propagation of each other's social media content. Finally, a user's followers on social media platforms represent the user's direct audience. The larger the user's follower base, the more easily the same user can extensively broadcast information. Therefore, I study what drives the growth of humanitarian organizations' follower bases during times of normalcy and emergency using Twitter data from one week before and one week after the 2016 Ecuador earthquake.

For my family and in loving memory of Dr. Janet Bell

ACKNOWLEDGMENTS

First and foremost, I would like to thank my main advisor and co-chair of my dissertation committee, Dr. Elliot Rabinovich. I am so grateful for all of the hours that he has dedicated to me over the past nine years to meet me, counsel me, and mold me into the researcher I am today. He taught me to always challenge myself and to relentlessly pursue what I am passionate about. These lessons I will carry with me for the rest of my life, not only in academia but also in my personal life.

I would also like to acknowledge and express my gratitude towards the rest of my committee members. Although Dr. Bin Gu is from another department, he embraced me and agreed to co-chair my dissertation. I am very thankful to him for making himself available for me to provide his perspectives on the dissertation and to teach me new methods. In addition, thank you to Dr. William Rand for being a part of my dissertation committee. Dr. Rand was instrumental in helping me publish the first part of my dissertation and has always been willing to discuss my research with me and give advice. Last but not least, I would like to thank Dr. John Fowler, who has been a part of my educational journey since serving on my undergraduate honors thesis committee. I am indebted to him also for paving the way for me to join the doctoral program.

I must also acknowledge the Department of Supply Chain Management for accepting and supporting me. I am sincerely grateful to Dr. Michele Pfund for helping me make my dream of becoming a doctoral student a reality. Thank you to Dr. Choi, Dr. Kull, Dr. Webster, and Dr. Yin for teaching the doctoral seminars and providing me with foundational knowledge of the field. I would also like to thank Eddie Davila for showing me how to be a teacher and for always having my back. Thank you to Dr. Mahyar Eftekhar for always being generous with his time and especially for his contributions to the first part of the dissertation. I also appreciate Dr. G, the chair of our department, for greeting

me with a smile and for his care. To my cohort – Sining Song and Sina Golara – thank you for going through these past five years with me and allowing me to lean on you for support and friendship. I would also like to thank my other cohort – Yousef Abdulsalam, Sangho Chae, Zhongzhi Liu, and Zac Rogers, for adopting me and letting me tag along with them. I am so blessed to have Rachel Balven in my life as my sister from another department and fellow nerd. Thank you for always being available for me for tea time and lunch dates.

From the bottom of my heart, I thank my family for being with me through all the ups and downs during the time I was a doctoral student. I am so grateful for my husband for vicariously living through the doctoral program with me and for teaching me to program. I would also like to thank my mom and dad for immigrating to America to expand my opportunities as well as for all the food, ice cream, laundry, and hugs they have given me. To my sister, I am thankful for distracting me and making me laugh. I would like to also give thanks to my TFC family for praying for me and for being my extended family. Thank you to Hari Mohanraj for the countless chats about food, random subreddits, and data science. Jen Chen, thanks for all of the work + affogato sessions and for graduating with me again. To Debbie Jang, thank you for checking on me and giving me things to think about and learn from outside of my research.

Thank you to the Departments of Supply Chain Management and Information Systems at the W. P. Carey School of Business for providing me with research funding. I would also like to thank the Center for Services Leadership for awarding me a dissertation grant and Arizona State University's Office of Knowledge Enterprise Development, the GPSA, and Graduate Education for awarding me the Graduate Research and Support Program award. Lastly, I would like to acknowledge the AWS Cloud Credits for Research program and Lorena Costanzo for supporting my work.

Above all, I thank God for providing for me and for giving me strength.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

PREFACE xii

CHAPTER

1 EVALUATING INFORMATION SPEED AND ITS DETERMINANTS IN SOCIAL MEDIA NETWORKS DURING HUMANITARIAN CRISES 1

Abstract 1

1. Introduction 2

2. Information Diffusion on Social Media Networks: Background, Theory, and Propositions 4

2.1. The Effect of Influential Originators on the Diffusion of Cascades 7

2.2. The Effect of Content Promoting Situational Awareness on the Diffusion of Cascades 8

2.3. The Effect of Timing in the Launch of Cascades on the Diffusion of Cascades 9

3. Research Methodology 10

3.1. Context: Twitter and Hurricane Sandy 10

3.2. Data Collection 11

3.3. Operational Measures 15

3.3.1. Dependent Variable 15

CHAPTER Page

3.3.2. Determinants 20

3.3.3. Control Variables 21

4. Empirical Analysis 21

4.1. Statistical Modeling 23

4.2. Results 24

5. Discussion of Results and Conclusions 26

Acknowledgements 31

2 THE INTERACTION OF SIMILAR CONTENT ON SOCIAL MEDIA PLATFORMS DURING DISASTERS 32

Abstract 32

1. Introduction 33

2. Background 38

3. Point Process Model for the Diffusion of Cascades 43

4. Data 48

4.1. Sample 48

4.2. Parallel Cascades 51

5. Model Estimation 55

6. Results of Model Estimation 58

7. Analysis of Competitive vs. Cooperative Effects by Parallel Cascades 61

8. Conclusion 64

CHAPTER Page

3 EXPANDING THE REACH OF HUMANITARIAN ORGANIZATIONS ON SOCIAL MEDIA PLATFORMS 68

Abstract 68

1. Introduction 69

2. Literature Review 73

2.1. Information Management in Humanitarian Operations 73

2.2. Social Media Platforms and Operations Management 75

3. Mechanisms for the Formation of Follower Links 77

4. Structural Model 80

4.1. Stage 1: Consumption 80

4.2. Stage 2: Follow Decision 82

5. Data 85

6. Internal and External Link Analysis 89

7. Structural Model Analysis 93

7.1. Model Estimation 97

7.2. Results 100

8. Robustness Checks 104

9. Conclusion 106

Acknowledgements 109

REFERENCES 110

CHAPTER Page

APPENDIX

A DESCRIPTION OF THE AGENT-BASED MODEL....122

B INTEGRATION OF THE CONDITIONAL INTENSITY FUNCTION 134

C DESCRIPTIVE STATISTICS OF PARAMETER ESTIMATES BY DISASTER....137

D DETERMINING NEW FOLLOWERS AS INTERNAL OR EXTERNAL LINKS...139

E TEXT CLASSIFICATION....143

F RESULTS FROM ROBUSTNESS CHECKS....147

LIST OF TABLES

Table		Page
1.	Irrelevant Tweets	12
2.	Breakdown of Cascade Categories	14
3.	Variable Operationalization	15
4.	Descriptive Statistics for MAPE Values	19
5.	Correlations and Descriptive Statistics	23
6.	GLM Results	25
7.	Information on Sampled Disasters	49
8.	Keywords and Phrases in Queries	50
9.	Examples of Cascades and Their Near-Duplicates	54
10.	Breakdown of Sample Size and Retweets by Disaster	54
11.	Descriptive Statistics for Parameter Estimates	58
12.	Descriptive Statistics for Determinants of α_{21}^{i}	62
13.	OLS Regression Results	63
14.	Categorization of Suppliers Listed by Twitter Handles	86
15.	External and Internal Links	93
16.	Summary of Notation and Variable Operationalization	94
17.	Classification of Tweets and their Content	95
18.	Descriptive Statistics for Key Variables	100
19.	Results of the Weighted Maximum Likelihood Estimation	101

LIST OF FIGURES

Figure		Page
1.	Cumulative Distribution of Cascades over Time	23
2.	Self-Exciting Point Process for a Sample Cascade	46
3.	Count of Cascades Initiated over Time	51
4.	Kernel Density Plot of Cascades' Retweet Counts	55
5.	Kernel Density Plot of Logged Follower Counts for Retweeters	59
6.	Sample Cascade's Arrivals and Intensity Based on Estimated Parameters	61
7.	The Internal Mechanism	79
8.	Count of (a) Tweets and (b) Retweets	87
9.	Magnitude of Audiences for Suppliers' Content	88
10.	Cumulative Distribution Functions of Logged Follower Counts for Suppliers and Retweeters	88
11.	Locating New Followers in Scraped Follower Lists	91
12.	Cumulative Count of New Followers	92

PREFACE

Humanitarian operations management is concerned with the coordination and delivery of resources that can alleviate the suffering of those affected by a disaster. Like the commercial sector, humanitarian operations must run efficient supply chains to be successful (Van Wassenhove 2006), but key differences distinguish humanitarian from commercial operations. First, the mission of humanitarian operations is not necessarily to minimize operational costs but, rather, to minimize human suffering (Holguín-Veras et al. 2013). Humanitarian organizations (HOs) face extreme levels of variability from the demand side since disasters cannot always be predicted as well as from the supply side since HOs are dependent on the availability of a highly uncertain resource amounts under varying lead times. Moreover, the operating environment is turbulent due to destabilized infrastructure and the convergence of many stakeholders (e.g., local government, military, and other HOs) with goals that may not be aligned toward a common objective (Kovács and Spens 2007, Van Wassenhove and Pedraza Martinez 2012).

Despite these challenges, humanitarian operations must fulfill their objective of distributing all required resources and services to beneficiaries. Examples of commonly demanded resources and services include food, water, and medical services. Another vital resource is information, especially since information facilitates the sourcing and delivery of other resources and services to beneficiaries and other stakeholders. In fact, the effective management of information is one of the most critical factors in determining the success of humanitarian operations (Long and Wood 1995). With accurate information about beneficiaries' needs, for instance, HOs can allocate resources such that the right products can reach the right population at the right time. HOs also issue donor appeals and exchange information with collaborating HOs to enhance coordination and avoid redundant efforts.

However, the management of information has been reported as a major struggle for humanitarian operations. As noted previously, the operational environment during a disaster is volatile due to factors like a damaged physical landscape, population migration, and disrupted economic and political states (Holguín-Veras et al. 2012). This means that decision parameters related to the operational environment are changing constantly, and what may have been relevant or accurate information yesterday is no longer so today. For example, the number of beneficiaries that a HO expects to serve may change suddenly due to notices of mandatory evacuations. Because information is highly perishable in the humanitarian context (Meier 2015), HOs require a robust information network that can quickly diffuse information among the appropriate stakeholders.

Social media platforms have emerged as a useful tool to address this need, and many HOs maintain an active presence on these platforms. HOs have found social media platforms to be valuable because information is shared in real-time and propagates rapidly through platforms' sharing functions. Using these platforms, HOs broadcast information about their available services and share updates about their projects. Furthermore, HOs employ social media platforms to collect information from beneficiaries that post first-hand knowledge of conditions at disaster sites (Gao et al. 2011). The purpose of this dissertation is to develop insights into how social media platforms can disseminate information during times of crisis by answering the following three research questions:

1. What user-related and content-related factors affect the diffusion speed of information on social media platforms in a disaster?
2. How is the dissemination rate of social media content affected by the concurrent diffusion of other content?
3. What mechanisms drive the growth of HOs' social media networks in periods with and without a disaster?

CHAPTER 1

Evaluating Information Diffusion Speed and its Determinants in Social Media networks during Humanitarian Crises[1]

Abstract

The rapid diffusion of information is critical to combat the extreme levels of uncertainty and complexity that surround disaster relief operations. As a means of gathering and sharing information, humanitarian organizations are becoming increasingly reliant on social media platforms based on the Internet. In this paper, we present a field study that examines how effectively information diffuses through social media networks embedded in these platforms. Using a large dataset from Twitter during Hurricane Sandy, we first applied Information Diffusion Theory to characterize diffusion rates. Then, we empirically examined the impact of key elements on information propagation rates on social media. Our results revealed that internal diffusion through social media networks advances at a significantly higher speed than information in these networks coming from external sources. This finding is important because it suggests that social media networks are effective at passing information along during humanitarian crises that require urgent information diffusion. Our results also indicate that dissemination rates depend on the influence of those who originate the information. Moreover, they suggest that information posted earlier during a disaster exhibits a significantly higher speed of diffusion than information that is introduced later during more eventful stages in the disaster. This is because, over time, participation in the diffusion of information declines as more and more communications compete for attention among users.

---

[1] This paper was previously published. The citation is as follows: Yoo, E., Rand, W., Eftekhar, M. and Rabinovich, E., 2016. Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *Journal of Operations Management*, *45*, pp.123-133.

1. Introduction

The management of humanitarian operations during disasters is often highly complex due to the extreme uncertainty and diversity of stakeholders involved in these crises (Van Wassenhove 2006). In such instances, gathering and sharing timely information regarding infrastructure, supply of resources, and needs is critical to develop an understanding of existing conditions and coordinate an effective response (Pettit and Beresford 2009). To that end, researchers have stressed the importance of rapid information diffusion for humanitarian organizations (HOs) to gather intelligence about conditions in affected communities (e.g., Oloruntoba and Gray 2006) and for HOs to distribute information among stakeholders in order to foster collaboration (Altay and Pal 2014).

Internet-based social media hosted on platforms like Twitter or Facebook may help facilitate information diffusion because they provide the means through which stakeholders can upload and share information with others in real-time and at virtually no cost. Many HOs have recognized the value of social media platforms and have started using them to access and share information from various sources. This includes data from informants with first-hand knowledge of what is occurring in affected areas (Gao et al. 2011), and recently, HOs have aggregated these data to create crisis maps showing landmarks like damaged infrastructure and shelters (Meier 2012). HOs have also used social media to share capacity levels and resource availability to enhance coordination among stakeholders (Sarcevic et al. 2012).

Despite these experiences, and calls by experts for additional research on the use of social media for humanitarian operations (e.g., Holguín-Veras et al. 2012, Kumar and Havey 2013), the literature on this subject is still at an embryonic stage. Most of this work has focused on descriptions and characterizations of social media responses to

humanitarian crises (e.g., Kaigo 2012, Kogan et al. 2015) and has yet to rigorously consider the dynamics of information dissemination during these events and their influence on humanitarian operations.

Our paper addresses this deficiency by analyzing diffusion dynamics of information in social media from a disaster case. To that end, we follow Ellison et al. (2007) and focus on a network representation of social media platforms on the Internet in which users can forge connections and share information directly with each other, as well as indirectly through other users. These connections will form social media networks in which information produced by a user (i.e., an originator) will create cascades when those connected directly to her receive it and, in turn, share it with those with whom they are connected. These information cascades will continue to spread as long as more users join these cascades by sharing the information they receive with those connected to them.

To address this objective, we develop and test a set of theoretical propositions regarding the role played by three key determinants of information diffusion dynamics in social media networks. Although past work has discussed the importance of these determinants in the crisis informatics literature (e.g., Ringel Morris et al. 2012, Starbird and Palen 2010, Vieweg et al. 2010), their impact on information diffusion across social networks remains undetermined. The first determinant focuses on the *influence* that information cascade originators have in these networks as a function of their social connections. The second one focuses on the *type of content* being shared in these networks and whether it contributes to improving situational awareness during a crisis. The third determinant corresponds to the *timing* in the introduction of information in these networks with respect to the progression of disaster events. Since the propositions focus on characteristics of cascades, the unit of analysis in our study is a cascade.

Our results show that information can spread faster when it originates from users that are influential in these networks. They also indicate that the timing when information is initially posted by an originator relative to a disaster's development of events will impact the information's rate of diffusion across social media networks. Information that is originally posted later, as a disaster intensifies, will spread at a lower rate than information that is posted at earlier stages of the disaster because, over time, participation in the diffusion of information cascades declines as more cascades compete for attention among users. This phenomenon underscores a paradox in which as a disaster's effects build up, there will be more cascades contributed by originators, but the information in those cascades will spread more slowly.

In the next section, we expand on theoretical explanations for the diffusion of information on social media networks and develop the propositions that guided our study. In Section 3, we detail how we collected the data and operationalized the variables to test the propositions. We then present the empirical model and the results pertaining to the evaluation of the propositions in Section 4, followed by a discussion of the results, implications, and conclusions in Section 5.

2. Information Diffusion on Social Media Networks: Background, Theory, and Propositions

Research based on Information Diffusion Theory has relied on different types of models of adoption to explain the dynamics of information cascades' diffusion in network settings. Two of the seminal models are the *Independent Cascade* (IC) model developed by Goldenberg et al. (2001) and Kempe et al. (2003) and the *Linear Threshold* (LT) model developed by Granovetter (1978). These models assume each member contributes monotonically to the diffusion of information (i.e., there is no dis-adoption or forgetting of the information). In these models, information diffusion proceeds iteratively over time

starting from a set of members that contribute information to be subsequently distributed by other members across the network (Guille et al. 2013). IC and LT models also account for information diffusion due to a member receiving information from sources external to the network or internally from those informed participants that are adjacent to her in the network (Myers et al. 2012).

IC and LT models, however, differ from each other in several aspects. IC models assume that an informed member has one chance at a time of independently sharing information with one uninformed member adjacent to her in the network (Kempe et al. 2003). Thus, at any point in time, an uninformed member has a likelihood, q, of becoming aware of the information when at least one of her neighbors in the network has already become aware of the information. But, in many versions of the IC model (Goldenberg et al. 2001), there is also a probability, p, that the individual will become aware of this information from external sources. High values for q and p will denote a high information diffusion rate throughout the network due to the internal influence of network connections or influence of sources external to the network, respectively (Guille et al. 2013).

In LT models, it is assumed that a participant will share information with her uninformed neighbors in the network if, over time, the number of informed members adjacent to her in the network exceeds her own influence threshold (Granovetter 1978). The lower this threshold across the network, the faster the participant will share information with her uninformed neighbors and the faster information will diffuse internally throughout the network. In prior work, this threshold is denoted by ϕ (Watts and Dodds 2007). In our paper, we operationalize this threshold by setting $\phi = 1 - q$. This allows us to maintain a relationship consistency with the IC model where high values of q indicate faster diffusion, and low values of q indicate slower diffusion. In some prior work, the q parameter is fixed for all individuals, while in other contexts it is chosen from a

distribution for each individual (Watts 2002). Traditionally, the LT model has not incorporated a p parameter, instead relying on the initial seeds of the network to propagate the information (Kempe et al. 2003, Watts and Dodds 2007), but a p parameter playing the same role that it does in the IC model can be added to this model instead of an initial seed (Dodds and Watts 2005).

Though previous work has created a generalized model that incorporates both the IC and LT models (Dodds and Watts 2005), we developed a framework that allows for versions of both the IC and LT models to be described using the same two parameters of p and q. To that end, we modeled the user decision process in the following sequential steps:

(1) **Effect of p:** Independent of the adoption model (LT or IC), each agent who has not yet adopted the information adopts the information with probability p due to discovering the information from a source of information diffusion outside the network structure.

(2) **Effect of q:** Depending on the adoption model, users take different actions.

   a. ***q in the LT model:*** Each user who has not adopted observes the number of neighbors who have adopted divided by the total number of neighbors they have. If that ratio exceeds ϕ, the focal user adopts the information (Watts and Dodds 2007).

   b. ***q in the IC model:*** Each user who adopted information in the most recent previous time step has q probability of transmitting the information to any neighbor who has not adopted the information (Goldenberg et al. 2001).

Though each of these models has found success in analyzing diffusion processes (e.g., Goldenberg et al. 2001, Guille et al. 2013, Rand et al. 2015, Watts and Dodds 2007), it is not obvious whether both models can be used jointly in studying information diffusion

on social media networks in the same context. As part of our contribution to the literature, we will first examine how IC and LT models explain these cascades' diffusion dynamics within the same context. Then, we will use this analysis to focus our line of inquiry on the effects of the three diffusion determinants we introduced in Section 1. We will expand on these determinants' effects below.

2.1. The Effect of Influential Originators on the Diffusion of Cascades

The diffusion of an information cascade will depend on the level of *influence* that the cascade's originator carries in the social network. An originator's influence is particularly relevant to the context of cascades in social media networks during humanitarian crises since users previously reported having significant concerns about the credibility of disaster information they received through social media (Ringel Morris et al. 2012). While influence can be assessed in a number of different ways, prior results from information diffusion models concentrate on influence measured by a user's number of social connections and suggest that users with large network audiences are perceived to have superior credibility (Bhattacharya and Ram 2012). These perceptions will allay concerns about trustworthiness and induce individuals to conform to cascades launched by influential originators (Goldenberg et al. 2009). Based on this evidence, we expect that users will be inclined to join cascades originated by network members with extensive influence, and as a result, these cascades will exhibit greater rates of internal diffusion.

Moreover, research has relied on the principle that influential cascade originators usually have numerous social connections that will expose large audiences to their cascades soon after they are launched (Kempe et al. 2003). This implies that if a cascade's originator is well-connected, the cascade will diffuse rapidly because a wider audience will be exposed early on to the cascade. We anticipate that this principle will also apply in the context of information diffusion in social media networks during a disaster. Hence, we

conjecture that an information cascade's diffusion may experience a surge soon after a highly influential user exposes the cascade's information to her network links. This will contribute to the cascade's overall rate of diffusion throughout the social media network. Proposition 1 summarizes this argument for our setting.

Proposition 1: In the context of cascades carrying disaster-related information throughout social media networks, the influence of a cascade's originator contributes positively to the cascade's speed of diffusion.

2.2. The Effect of Content Promoting Situational Awareness on the Diffusion of Cascades

Research shows that diffusion rates will increase if network members perceive that cascades' contents are informational and that sharing these contents will be helpful to others (Rogers-Pettite and Herrmann 2015). Based on this evidence, we argue that, during humanitarian crises, network members are more inclined to participate in cascades carrying informational content that is seen as useful to disaster relief operations. For many of these members, the decision to join cascades conveying informational content related to disaster relief will follow altruistic and emotional motivations to help victims. In joining these cascades, these members anticipate no material gains. Instead, they look to obtain rewards resulting from their cooperation with other cascade participants and from offering support to others in need (Fowler and Christakis 2010).

In a humanitarian context, these information cascades will convey content that will heighten *situational awareness*. Situational awareness, in itself, is defined as a complete and coherent understanding of what is going on during emergencies, and it is gained from information that helps to assess the situation at hand (Sarter and Woods 1991, Vieweg et al. 2010). In humanitarian operations, information supporting situational awareness is vital because decision parameters are highly dynamic (Holguín-Veras et al. 2012). Hence,

situational awareness is required to make decisions that are well-informed and reflective of current events.

Given the value of situational awareness, we expect that network members will have a greater disposition to join cascades that carry information that could improve situational awareness. Our expectation follows evidence showing that cascades with information that improves situational awareness exhibit greater participation among social media users (Vieweg et al. 2010). Thus, messages meant to improve situational awareness during a crisis are likely to strengthen the diffusion of information cascades across social networks. Proposition 2 formalizes this argument.

Proposition 2: In the context of cascades carrying disaster-related information throughout social media networks, speed of diffusion will be higher for cascades carrying information that heightens situational awareness than for cascades carrying other types of information.

2.3. The Effect of Timing in the Launch of Cascades on the Diffusion of Cascades

Past work on information diffusion has underscored the role played by temporal patterns in the dissemination of information across networks. As part of this body of work, Boyd et al. (2010) identified a preference by participants in social media networks to share time sensitive information with others. This is particularly relevant in a humanitarian context, in which participants will be motivated to share urgent information that will help address directly their own needs and those of others in the network.

Leskovec et al. (2009) argued that the level of motivation among network participants to share time-sensitive information will contribute to the likelihood of certain topics gaining initial traction among network participants and eventually forming a cascade. These topics, for example, may comprise the development of urgent news events during a humanitarian crisis. At an early stage during a disaster, cascades addressing such

topics will spread quickly as more participants imitate one another in sharing information. But over time, the rate of participation in the diffusion of cascades will decline as newer topics compete with older ones for attention. As a result, the diffusion of new cascades is likely to become increasingly difficult, regardless of the urgency embedded in an information cascade. Cascades that are launched at later stages during the course of a crisis are therefore expected to diffuse at a lower rate than cascades launched at earlier stages. That is, the diffusion of information cascades on social media networks will decline as a disaster unfolds. Proposition 3 formalizes this argument.

Proposition 3: In the context of cascades carrying disaster-related information throughout social media networks, the speed of diffusion will be lower for cascades that are launched later than for cascades launched earlier during the progression of a disaster event.

3. Research Methodology

3.1. Context: Twitter and Hurricane Sandy

We focused on Twitter to test our propositions. Social networks on Twitter are based on directional links between users. On Twitter, a user can follow, or track, the messages (or "tweets") of another user or be followed by other users (called "followers"). Users can receive the tweets of those they follow and broadcast all of their own tweets to their followers. Twitter also gives a user the ability to "retweet" original tweets or other retweets posted by users that she follows in order to share these messages with her own followers. A user's retweets preserve the contents of the original message, and these retweets may be shared in turn by the user's own followers, who may or may not be a part of the network of the user who uploaded the original tweet.

Our study focused on Twitter data associated with Hurricane Sandy, a disaster for which Twitter usage has received some research attention (e.g., Rand et al. 2015).

Hurricane Sandy is considered to be the largest Atlantic hurricane on record in the United States (U.S.). It began as a tropical storm in the Caribbean in October of 2012, grew into a Category 3 hurricane at its peak, and impacted the Eastern U.S. We determined Hurricane Sandy to be an appropriate disaster case for our study for two reasons. First, the hurricane's major effects were felt in a densely populated, highly developed area. Because of the hurricane's magnitude and Twitter's popularity in this area, a large volume of tweets were posted in relation to this event, creating a rich dataset for empirical analyses. Second, as the main effects of Hurricane Sandy were felt in the U.S., tweets were mostly sent in English. This eliminated the need for translation to address our research objectives.

3.2. Data Collection

Our data contain original tweets and retweets posted from October 26 until October 30, 2012. These dates correspond to the periods before, during, and after Hurricane Sandy effects were experienced in the U.S and overlap with the stages when preparation and response activities to the hurricane occurred. Preparation and response stages are usually the most relevant for humanitarian operations in many disasters as high levels of uncertainty and volatility in conditions on the ground are pervasive at these times (Van Wassenhove 2006).

The collected data include the actual contents of the tweets and retweets, information about the users responsible for these posts, and the date and time, to the second, when each of the posts appeared on Twitter. The data were gathered in real-time using Twitter's Search API, an interface through which one can program queries to collect tweets and retweets posted within the past seven days. Twitter limits the amount of data that can be downloaded per IP address using the Search API. To overcome this limit, a script using the Search API was run constantly on ten different machines with a rule that would pull tweets and retweets containing the keywords "Sandy," "hurricane," "storm,"

and/or “superstorm”. Based on the volume of data downloaded, we were confident that the Search API extracted a high percentage of the tweets and retweets that contained our search keywords during our data collection period. Nevertheless, we decided to evaluate the completeness of the data gathered through the Search API by comparing it against a sample we acquired from Gnip, a Twitter subsidiary with access to the entire Twitter firehose (i.e., all activity ever posted on Twitter). To draw the Gnip sample, we used identical keywords and date ranges to those specified for the Search API sample. Our comparison demonstrated that the Search API only missed 7.81% of the messages in the Gnip dataset. This suggests that our sample contains a vast majority of the tweets and retweets posted during Hurricane Sandy and with the selected keywords.

Subsequently, we used a program to separate the original tweets from the retweets that the Search API extracted. We manually reviewed all of the original tweets and filtered out those that we deemed irrelevant along with their retweets. Although they contained the chosen keywords, irrelevant tweets included jokes, song lyrics, emotional responses, and discussions of topics unrelated to Hurricane Sandy. Please refer to Table 1 for more detail on irrelevant tweets. After removing the irrelevant messages, we were left with 18.27% of the original tweets in the sample along with their retweets[2]. In total, these tweets and retweets corresponded to 333,968 messages.

Table 1: Irrelevant Tweets

Irrelevant Category	**Example Tweet**
Emotional Response	“actually really scared of the hurricane coming :(“
Joke	“Hurricane Sandy sounds like a delicious mixed drink.”
Not Related to Sandy	“Yay!!(: hanging out with my bestfriend @strong_sandy”
Opinion	“I get the feeling this hurricane in gonna be just like irene and barley [sic] hit us..”
Song Lyric	“The voice that calmed the sea would call out through the rain and calm the storm in me... -Casting Crowns. I love this song! #whoami”
Vague Forecast	“Sandy is coming“

[2] The process of cleaning and categorizing the cascades took approximately 45 hours to complete.

Because our propositions dealt with information cascade effects, the unit of analysis for our study is the cascade. In view of this, we organized the tweets and retweets in the dataset into cascades. We followed the lead of authors who have previously conceptualized information cascades in Twitter as retweet chains (e.g., Galuba et al. 2010, Lerman and Ghosh 2010). Each original tweet represented the start of a cascade, and retweets by additional users signaled participation in a cascade. In Twitter, the text in all retweets is usually identical to the text in the original tweet that launched the cascade since Twitter makes retweets possible through the push of a single button. Retweets are also marked at the beginning by "RT@username," followed by the original tweet's text. The username following "RT@" identifies the user that posted the original tweet and launched the cascade.

Based on these attributes, we compiled cascades in our data by identifying and grouping retweets that shared the same text and embedded originator usernames. Then, to ensure that each group of matching retweets constituted an actual cascade and not background conversations among select users, we confirmed that each cascade consisted of at least ten retweets issued at varying intervals. This process generated 5,683 cascades. We chose a threshold of ten retweets because cascades on Twitter usually do not require many retweets to develop (Lerman and Ghosh 2010).

We then developed a program to examine in detail the original tweets that began each cascade. Through this program, we isolated the username embedded in the beginning of each retweet's "RT@username" and separated the original tweet's text that followed. Then, the program searched through the dataset and pulled each original tweet with the matching username and content. In this process, we found that 249 cascades (comprised of 19,558 retweets) could not be matched to their original tweet because they had missing

information about the originating users[3]. This prevented us from identifying the time when each of these cascades started, and therefore, we were unable to examine their diffusion. Although this left us with no option but to drop these cascades from our sample, the removal of these cascades had a negligible impact on our results since they constituted only 4.3% of our observations. After we filtered these cascades, we were left with a final sample of 311,429 retweets forming 5,434 cascades to evaluate our propositions. Table 2 shows the distribution of the cascades across six content categories.

Table 2: Breakdown of Cascade Categories*

Category	Count of Cascades	Description	Sample Retweet
Advisories	2,024	Transportation shutdowns, evacuation warnings, survival/safety tips, and updates on hurricane intensity/trajectory	RT @Timcast: Reports that all NYC bridges will be closing at 7pmEST via @NYScanner #Sandy #Frankenstorm
Business	445	Reports of business-related shutdowns and forecasts of economic impacts	RT @Reuters_Biz: Stock bond markets shut on Tuesday may reopen Wednesday http://t.co/JL6fEHea
Declarations	141	Declarations of emergencies by states	RT @USNationalGuard: So far governors in MD VA NY DC PA CT NC NJ DE MA and VT have declared states of emergency ahead of #Hurricane #Sandy.
Forecasts	640	Forecasts of weather and hurricane effects	RT @twc_hurricane: BREAKING: TWCs experts now expect localized wind gusts of 90+ mph near the coast of NJ NYC and Long Island later today. #Sandy
Humanitarian	246	Information related to shelters, relief efforts, and deployment of aid	RT @femaregion2: #Sandy Search for open shelters by texting: SHELTER + a zip code to 43362 (4FEMA). Ex: Shelter 01234 (std rates apply)
Reports	1,938	Status updates of weather, damage, outages, etc.	RT @News12LI: As of 10:32am LIPA is reporting 15695 outages across Long Island. #Sandy

*Adapted from Olteanu et al. (2014) and Vieweg et al. (2010)

[3] This information may be missing from the data because privacy settings chosen by the originators did not allow the Search API to access this information or because the original tweet was posted before the start of the data collection.

3.3. Operational Measures

In this section, we expand on the operationalization of the variables introduced in the propositions. Moreover, we introduce a set of control variables to be used as part of the empirical testing of these propositions. Table 3 lists the variables in the propositions and the control variables along with their operationalization.

Table 3: Variable Operationalization

Construct	Variable Label	Operationalization
Information Cascade's Diffusion Speed	*DIFFUSION*	Ratio of *q/p* values obtained from the *IC* model
Cascade Originator's Influence	*INFLUENCE*	Number of users following the cascade originator (at the time of cascade launch)
Cascade Content's Contribution to Situational Awareness	*AWARENESS*	Dummy variable coded 1 if the cascade content contributed to situational awareness; 0 otherwise
Lateness in the Launch of the Cascade during the Disaster	*LATENESS*	Lag in the launch of the cascade relative to the start of the data collection (measured in hours)
Incidence of Cascade Boosts by Originator	*BOOST*	Dummy variable coded 1 if originator boosted the cascade; 0 otherwise
Misleading Cascade	*FALSE*	Dummy variable coded 1 if the cascade content was misleading; 0 otherwise

3.3.1. Dependent Variable

To measure the cascades' diffusion speed on Twitter's network, we followed Rand et al. (2015)'s approach and ran an agent-based model (ABM) to evaluate how well the IC and LT models we introduced in Section 2 represented the cascade data. This generated an overall adoption rate of information at discrete time steps. ABM offers a robust understanding of information diffusion on social networks since it represents not only the properties of the individual agents but also their communication channels via local network connections. Rand and Rust (2011) identify up to six properties of a system that make it useful to model using ABM: (1) a medium number of agents, (2) local and potentially complex interactions among agents, (3) agents' heterogeneity, (4) rich environments, (5) temporal aspects, and (6) agents' adaptability. Information diffusion on social media features all six of these properties to an extent, making ABM a suitable

method for our study. Please refer to Part I of Appendix A that accompanies this paper for a more detailed discussion of the appropriateness of ABM. The ABM was constructed, verified, and validated following the guidelines of Rand and Rust (2011). Parts II through IV of the appendix contain supplemental information of model construction, verification, and validation beyond the details given below, and Part VI shows the natural language version of the code used to create the ABM.

There were two basic entities in the ABM: (1) a Twitter user interested in receiving and transmitting information and (2) the relationship between each pair of users in a cascade, i.e., a social tie or a link. Ties between users enabled the transmission of information across each cascade. In Twitter, two users are connected to each other if one of the users follows the other and/or vice versa. Thus, the agents in the ABM possessed a set of links that corresponded to the social links of each user to other users based on their "following" relationships. We patterned these relationships against the links observed across a sample of 4,076 participants in the longest cascade in our dataset. Using Twitter's RESTful API, we identified the users followed by each cascade participant at the time of Hurricane Sandy. This yielded a total of 1,322,814 links, of which 3,315 served to cascade the information by being direct connections between users who were part of the cascade. Because of the rate limits on Twitter's RESTful API, it would have taken a prohibitive amount of time to pull all of the networks for each cascade. Therefore, we used the network for the longest cascade as the pattern for all of the other cascades we examined. While this decision simplified the modeling process, it is not a major limitation since Twitter exhibits scale-free properties, meaning that subnetworks are similar to their corresponding larger networks (Kwak et al. 2010).

Since the main observation in the ABM was the overall adoption of information at each time step for each cascade, agents had a property that specified whether or not they

had adopted new information. By adoption of new information, we mean the joining of a cascade by retweeting. In addition, agents had both a coefficient of external influence (p) and a coefficient of internal influence (q) that controlled the rate of adoption of a new piece of information in each cascade following external or internal stimuli, respectively. At the beginning of the ABM, all agents started in an "un-adopted" state, and a directed social network linking the agents was formed based on the empirical Twitter networks described above. Then, at each time step, any agents that still had not adopted the information decided whether to adopt the information based on p, q, and the state of their neighbors in the network. Agents followed the unified model discussed in Section 2 to make these decisions. The agents first chose whether to adopt based on external influence. To do this, they drew a random number from the uniform distribution of [0,1). If that number was less than p, they then adopted that information. This decision rule for external influence was identical regardless of whether the LT or IC models were considered.

The role of internal influence of network links was subsequently considered. In the LT model, each agent counted the number of neighbors that had adopted the information and divided this sum by the total number of neighbors. The agent then compared this number to $\phi = 1 - q$, and if the ratio was higher than ϕ they proceeded to adopt the information. In the IC model, each agent who adopted the information in the most recent time step transmitted the information to all of its neighbors who had not adopted. These uninformed agents drew a random number from the uniform distribution of [0,1), and if the number was less than q, they adopted the information. After all of the non-adopting agents had considered whether or not to adopt according to the rules described above, statistics on the number of adoptions that occurred during that time step were calculated. The model then iterated again until every agent in the network had adopted the information. We calibrated our model so that a time step was roughly one minute. This

enabled a seamless comparison to the observed data, which was also set in a resolution of one-minute increments.

The ABM provided observations for each cascade on the adoption of information at each time step for the IC and the LT models. We then compared this information to the empirical data to determine for each adoption model and each cascade which values of p and q best matched the empirical data. To complete this task, we used a simulated annealing (SA) approach. This method works by generating iterative values of p and q and measuring the performance of the model between the time series of the model data and the observed data for each cascade until identifying the parameter values for p and q that optimize this performance. We chose to use SA since a full search of the parameter space was precluded by the computational cost, and SA provides a robust way to search the space quickly for a set of parameters that minimizes errors. For technical details on the number of runs and implementation of the SA algorithm, please refer to Part V of the electronic appendix.

To estimate the performance measure from each model run for each cascade network, we obtained values for *Y(t)*, the number of agents in the network who had adopted the information at each time step, t. Next, we compared *Y(t)* to the actual number of adopters per time step from our empirical data, *Empirical(t)*, using the Mean Absolute Percentage Error (MAPE). As Equation 1 shows, the MAPE is equal to the absolute difference between the empirical value of information adoption observed at time step, t, throughout the duration of the cascade and the ABM's value at that same time step, divided by the empirical value at time step t and averaged over all values (n).

$$MAPE=100\times\frac{1}{n}\sum_{t=0}^{n}\frac{|Empirical(t)-Y(t)|}{Empirical(t)} \quad (1)$$

We then averaged the MAPE across k runs. For a sample of the cascades, we observed that the average MAPE did not change markedly with more than ten runs for a

given parameter setting. Thus, we chose to use ten model runs to provide an adequate estimate of the underlying adoption patterns for a given cascade network and a given set of parameters. It was this average MAPE value over ten runs that was then used by the SA approach to optimize the parameter values.

Table 4 provides a distribution of the MAPE values across all the cascades for the IC and LT models. A comparison of the MAPE values for p and q across the cascades revealed that the MAPE values for p and q were consistently low across the IC and the LT models and similar to values identified for this metric in previous studies (Rand et al. 2015). Since both the IC and LT models performed well, we chose to focus on the IC model to operationalize information cascades' diffusion speed as our dependent variable (henceforth labeled as *DIFFUSION*). This is because the IC model allowed for a more direct measurement of *DIFFUSION* as the ratio of q/p values obtained from the model's output[4]. By operationalizing the dependent variable as q/p, we were able to account for diffusion forces due to sources internal and external to the networks underlying the cascades.

Table 4: Descriptive Statistics for *MAPE* Values

Model	Median	Mean	SD	Min	Max
IC	13.36	22.91	94.58	0.61	5,611.07
LT	12.97	21.89	86.26	1.08	4,821.76

[4] In the IC model, q represents a probability of internal influence, i.e., adoption due to internal influence is q multiplied by the fraction of neighbors who have adopted. Therefore, q/p describes the difference in spreads due to internal influence vs. external influence. However, in the LT model, q is a measure of how low the threshold to adoption is due to internal influence. This is different than a probability of adoption. Hence, q in the LT model is not directly comparable to p in the LT model since p is a direct measure of the probability of adoption due to external influence. This makes it difficult to make direct claims about the rate of internal vs. external adoption in the LT model based on these parameters. Nevertheless, since we developed the ABM under both IC and LT models, the ABM could serve to evaluate which rules cause the agents to adopt, and, from that, count up the number of agents that adopt due to internal influence and external influence in the LT model and compare those numbers to gauge diffusion speed indirectly.

3.3.2. Determinants

We are interested in investigating three determinants: (1) the cascade originator's influence, (2) the cascade content's contribution to improving situational awareness, and (3) the timing of the launching of the cascade. To measure a cascade originator's influence, we followed Cha et al. (2010), who explained that an agent is influential when it acts as an information channel to a large audience. This is consistent with opinion leadership models that support the notion that individuals are influential when they have a high number of connections with others (Bonacich 1972). Thus, we measured a cascade originator's influence as the number of the originator's followers on Twitter at the time the cascade was launched (*INFLUENCE*).

The second explanatory variable serves to identify those cascades that spread information related to situational awareness. To identify whether a cascade included this type of content, we created a dummy variable using the categorization scheme introduced in Section 3.2. This dummy (*AWARENESS*) equals 1 if the cascade belonged to advisories, humanitarian, or reports categories since, as detailed in Table 2, all dealt with information about safety, shelters, or the functional state of the affected areas. Otherwise, *AWARENESS* equals 0. We validated this operationalization by having four raters independently classify whether a randomly sampled set of 100 cascades pertained to situational awareness as defined in this study. We then checked the inter-rater agreement of our and the raters' classifications using Fleiss' kappa (Fleiss 1971). The kappa statistic was equal to 0.68, which indicates substantial agreement (Landis and Koch 1977).

Finally, the third explanatory variable captures the timing of each cascade's launch during the disaster. To that end, we measured the difference in hours between each cascade's launch and the time when we began our data collection. By calculating these intervals, we

captured how late a cascade was launched during the disaster. We labeled the variable for this measure as *LATENESS*.

3.3.3. Control Variables

As part of our empirical model, we accounted for instances in which cascade originators attempted to artificially increase the rate of diffusion of information in their cascades. Therefore, our first control variable accounts for instances when users boosted (or bumped up) those cascades that they themselves originated in order to increase the cascades' visibility on Twitter. A user may attempt to give a cascade that she originated a "boost" by reposting, at least once, the same tweet that initiated a cascade. However, in doing so, the originator may contribute to artificially distorting the cascade's growth pattern and its rate of diffusion. We controlled for this effect by using a binary indicator (*BOOST*) that specified which cascades in our sample were boosted by their originators or not. We set *BOOST* to 1 if a cascade was boosted by its originator or 0 otherwise.

Moreover, we controlled for whether the information conveyed in a cascade was misleading. Prior studies have documented the circulation of manufactured information in online social networks during disasters (e.g., Kaigo 2012). In our sample, some cascades contained information that purposefully exaggerated the size of the hurricane while others conveyed messages designed to convey outlandish claims about damages caused by the hurricane. Because such reports can generate a sense of panic among users (Gupta et al. 2013), they may artificially increase the rate of diffusion of information in these cascades. We controlled for this effect with a dummy variable (*FALSE*) that is set to 1 if the cascade's contents were false and 0 otherwise.

4. Empirical Analysis

We used regression analysis to test the propositions based on Equation 2. The use of regression analysis enabled us to specify the rate of diffusion for a cascade, *i*, as a

function of the explanatory and control variables discussed in Section 3.3 in addition to an error term, u_i.

$$DIFFUSION_i = \beta_0 + \beta_1 INFLUENCE_i + \beta_2 AWARENESS_i + \beta_3 LATENESS_i + \beta_4\, BOOST_i + \beta_5 FALSE_i + u_i \quad (2)$$

Figure 1 shows a cumulative distribution of the cascades' originations over time, and Table 5 lists the descriptive statistics for the variables in Equation 2. Since the mean for *DIFFUSION* (37.28) is statistically higher than 1 ($p<0.01$), our data suggest that internal information diffusion on social media networks advances at an average rate that significantly exceeds the average speed at which information originates from external sources. Please note that we limited the range of our parameters to historically observed values (Chandrasekaran and Tellis 2007). Thus, it might be argued that we did not explore a large enough range to observe model fits with very large *p* values. As a robustness check, we examined the number of cascades where the optimal *p* values were at the maximum range of exploration we allowed. Out of 5,434 cascades, only 648 of the IC model fits had *p* values at their maximum value, and of those 648, only 12 had the minimal *q* values. This means that for at least approximately 88% of our cascades, the best model fit was one where internal influence of network connections was much higher than external influence. In fact, removing the runs where *p* reached its maximum value changes the mean for *DIFFUSION* to 40.25, which illustrates how strong a role internal influence plays in the vast majority of these cases.

(Figure 1 and Table 5 on next page)

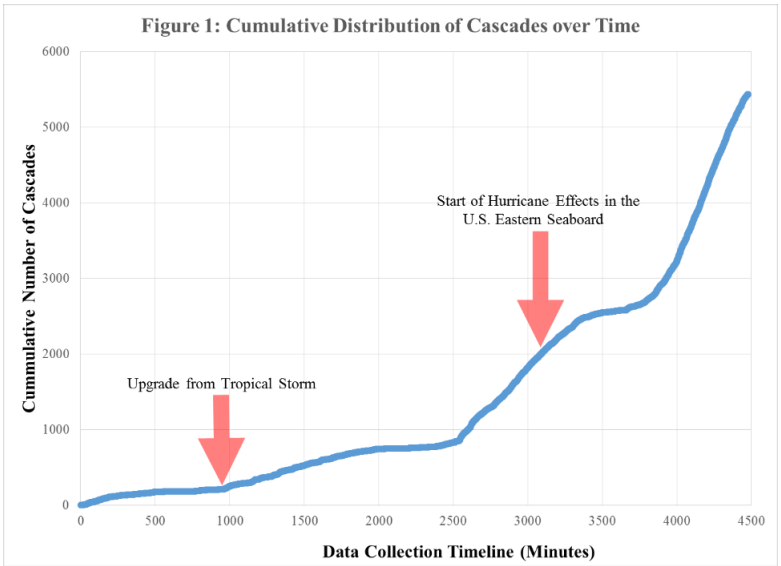


Table 5: Correlations and Descriptive Statistics

	1	*2*	*3*	*4*	*5*	*6*
1. DIFFUSION	1					
2. INFLUENCE	0.03*	1				
3. AWARENESS	-0.01	-0.04**	1			
4. LATENESS	-0.25**	0.01	0.16**	1		
5. BOOST	0.28**	0.01	0.01	-0.09**	1	
6. FALSE	-0.01	-0.04**	0.11**	0.16**	-0.03	1
Mean	37.28	234,447.66	0.78	55.34	0.02	0.04
Std. Deviation	55.63	848,551.55	0.42	18.04	0.13	0.19
Minimum	12.67	0	0	0.00	0	0
Maximum	737.80	9,133,950	1	74.65	1	1

** p<0.05, **p<0.01*

4.1. Statistical Modeling

We used a Generalized Linear Model (GLM) with a gamma distribution to model Equation 2. This approach was suitable for our model because *DIFFUSION* only took on positive values and displayed a right-skewed distribution of values. Also, after probing the relationship between *DIFFUSION* and *INFLUENCE* and *LATENESS*, we observed that the variance of *DIFFUSION* increased with the mean. This is consistent with the gamma distribution ($Var\,[Y_i] = \mu^2/\nu$). Separate plots of *DIFFUSION* versus *INFLUENCE* for each of the two categories in *AWARENESS* also revealed that there were some outlying

DIFFUSION values at extreme *INFLUENCE* values, which is another property consistent with the gamma distribution (Dobson and Barnett 2008).

To ensure an appropriate use of GLM, we also followed several additional steps. First, we used a Pearson Chi-Squared estimation method to estimate the GLM scale parameter (McCullagh and Nelder 1989). Second, we examined a log link function and an identity link function as possible alternatives to transform the dependent variable to estimate the GLM. Although the GLM results were fully consistent across both link functions, the identity link function provided significantly better Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) fit measures than the log link function. Thus, the results we report in this paper correspond to those obtained using the identity link function (Hardin and Hilbe 2007). The results obtained using the log link function are available upon request. Third, we used the Huber-White sandwich estimators to estimate standard errors that are robust to possible misspecification of the variance and link functions in the GLM. Finally, we checked for multicollinearity among the explanatory and control variables and found that almost all correlations among these variables were fairly small (Table 5).

4.2. Results

Table 6 presents the results from the GLM. To generate these results, we used a hierarchical approach. We first considered a restricted model in which we regressed the dependent variable only upon the control variables (GLM 1). Then, we regressed the dependent variable on the control variables as well as the explanatory variables in the propositions (i.e., unrestricted model or GLM 2). The results from likelihood ratio chi-squared test of GLM 2 indicate that the group of explanatory variables is statistically significant. Significant reductions of the AIC, BIC, and Deviance measures for GLM 2 also confirm that the addition of the predictors in GLM 2 makes a statistically significant

contribution in explaining our dependent variable's variance, above and beyond the contribution made by the control variables (Coxe et al. 2013, Hardin and Hilbe 2007).

Table 6: GLM Results

	GLM 1 Coeff. (Std. Errors)	**GLM 2 Coeff. (Std. Errors)**
INFLUENCE		1.05E-6 (4.92E-7)*
AWARENESS		-0.27 (1.25)
LATENESS		-0.75 (0.05)**
BOOST	124.31 (20.37)**	100.65 (18.12)**
FALSE	-1.76 (1.64)	9.13 (1.61)**
Intercept	35.36 (0.68)**	77.15 (3.36)**
Scale Factor	1.84	1.45
Likelihood Ratio Chi-Square	181.51**	612.32**
AIC	49864.94	49320.01
BIC	49884.74	49359.61
Deviance	3170.49	2619.56
Obs.	5434	5434

**p< 0.05, ** p< 0.01*

From Table 6, the coefficient for *INFLUENCE* was positive and statistically different from zero ($p<0.05$). Therefore, Proposition 1 is confirmed: as a cascade originator's influence rises, the speed of information diffusion in the cascade increases. Moreover, the effect by *LATENESS* on the dependent variable was negative and statistically different from zero ($p<0.01$). This means that, during a disaster event, the rate of information cascades' diffusion decreases over time as cascades are launched later during the disaster event. Proposition 3, therefore, is also confirmed. Proposition 2, however, received no support since the coefficient for *AWARENESS* was not significantly different from zero. Thus, we have no evidence to conclude that cascades carrying information that heightens situational awareness during a crisis will experience faster diffusion than cascades carrying other types of information. The lack of support for Proposition 2 is surprising based on theory and previous findings (Vieweg et al. 2010) but raises an important point that social media networks like Twitter may be limited in effectively spreading certain types of content. This is vital for HOs to understand as they create policies and strategies for managing information in a crisis.

Among the control variables, we observed that the coefficient for *BOOST* was positive and significant ($p<0.01$). Hence, boosting a cascade's original message is associated with an increase in the cascade's diffusion rate. Another result is that cascades that contain false information circulate at a faster rate than cascades that do not. This is evident from the positive and statistically significant coefficient for the control variable *FALSE* ($p<0.01$).

5. Discussion of Results and Conclusions

The planning and execution of humanitarian operations depends on a variety of resources that have very short shelf lives. Our research builds on the fact that information constitutes one of those resources. During times of crisis, it is critical to gather and share information quickly, but accomplishing this goal has been difficult for reasons that include a restricted diffusion of information relevant to humanitarian operations during the course of disasters (Day et al. 2012). While it has been theorized that social media networks built on open Internet platforms can contribute to address these restrictions (Meier 2015), there is limited work in the humanitarian operations literature that examines whether and how this can be accomplished. Moreover, while extant research in this field has focused on the development of analytical models to manage information (Özdamar and Ertem 2015), it is only recently that empirical research has begun to study these phenomena, particularly in social media settings (e.g., Korolov et al. 2015).

Our study addresses this deficit in the literature by applying Information Diffusion Theory to the context of humanitarian disasters. Our findings show that, in this context, cascades on social media networks can advance at a rate that significantly exceeds the speed at which information originates from external sources. This finding is important because, during humanitarian crises, speed is key in the diffusion of information among HOs and other stakeholders in order to plan and respond effectively to rapid changes that

occur during this type of events. Establishing that social media networks can diffuse information via connections among its users at a rate above that in which external sources of information permeate these networks during a crisis constitutes an important contribution to assessing these networks' effectiveness.

Another contribution from our results is that they show that this speed of diffusion is contingent upon the type of users that originally publish this information. When information is issued by users with high levels of influence, as measured by their number of followers, it will diffuse quickly. However, this will not be the case if the originators' influence is limited. For HOs, this implies that the development of social connections in these networks will be a valuable strategy to pursue in order to ensure fast communication with stakeholders like public donors and beneficiaries during times of crisis. Still, a question that deserves further investigation is whether information diffusion speed will experience different rates of growth as a function of the originator's number of followers once that number reaches certain thresholds. An examination of our data revealed that the rate of growth in diffusion speed as a function of the number of followers seems to increase as that number reaches a threshold of approximately 600,000 followers. Originators with an amount of followers above this threshold appear to have a significant leverage on the diffusion of information. A reason for this is that observations above this threshold sit at the head of a power law distribution across users in our dataset and, thus, can exert a significant pull on diffusion. This is in line with past research that has identified the presence of power law distributions underlying properties of social media networks like Twitter (e.g., Hodas et al. 2013).

The speed of information diffusion on social media networks during a disaster is also contingent upon the time when information is introduced in these networks. Information that is posted earlier during a disaster exhibits a significantly higher speed of

diffusion than information that is introduced later during the disaster. This is because, over time, participation in the diffusion of information cascades declines as more cascades compete for attention among users. Such a phenomenon is particularly acute in the context of a hurricane like the one in our study in which the number of new cascades increases sharply over time after hurricane effects materialize in large population areas (see Figure 1). This phenomenon also underscores a paradox in which, as a disaster progresses, there are increasingly more cascades contributed by originators, but the information in those cascades diffuses more slowly. As a result, a major challenge emerges for HOs trying to introduce urgent information and promoting its diffusion among an increasingly larger volume of new messages posted by other users. How can HOs increase the rate of diffusion of information among all this chatter? Addressing this information directly to followers or requesting explicitly that they retweet the information can augment diffusion (Huberman et al. 2008), particularly if those followers are themselves influential. Including hashtags and links in messages can influence the rate in which users spread information as well (Galuba et al. 2010).

We also observed that cascade originators may be able to increase the speed of diffusion by posting the same information repeatedly in order to raise its visibility. This practice can be justified among HOs in particular situations where information is of urgent nature, particularly during times of excessive chatter like those described above. However, it remains to be seen whether this practice carries with it diminishing marginal returns in increasing the rates of information diffusion. Moreover, we observed that cascades with fabricated information infect the network at a faster pace. Although our data demonstrate that cascades transmitting misleading information transpire rarely (only 4% of the cascades were false), this finding does raise troublesome questions about the ability by HOs and other participants in social media networks to detect and correct this type of

cascade. For instance, what attributes do cascades carrying misleading information share that could be used to identify them before they spread too far? What mechanisms can be instated in order to alert the public about these cascades and reverse their diffusion? The design of policies that address these questions and their joint implementation by a wide variety of HOs will help improve the effectiveness of social media in diffusing reliable information to other stakeholders.

It is also important to note that our research found no evidence to suggest that cascades carrying content that enhances situational awareness exhibit significantly higher diffusion rates relative to other cascades. This is surprising given that authors have previously noted that user participation is greater for cascades with information related to situational awareness (e.g., Vieweg et al. 2010). It is possible that the effects of other content-related factors, such as the use of Twitter hashtags or directional operators, on cascades' diffusion rates supersede the effect of situational awareness content. It is also possible that high diffusion rates may be observable but only for those cascades contributing new situational awareness content. That is, content that offers the most up-to-date information of how a disaster event is unfolding.

Another limitation in our research is that it does not assess the geographical implications of information flows in social media networks. We do know from our data that as information diffused on the networks, it reached a substantial amount of local individuals affected by our study's focal disaster. We found that almost 35% of all users in our data were located in geographical areas affected by the disaster. In total, users located in the areas affected by Hurricane Sandy participated in almost all (96.87%) of the cascades. In addition, in 80% of the cascades in our data, 20% or more participants were located in areas affected by the disaster. Thus, a large amount of information in these cascades did manage to reach people located in areas of need.

Prior evidence suggests that local individuals who are geographically vulnerable during a disaster share information differently in social media networks than individuals located in areas unaffected by the disaster (Starbird and Palen 2010). In particular, local individuals are more likely to contribute information during a humanitarian crisis than other individuals. Those local to a disaster are also more likely to propagate information received from other local individuals during a disaster (Kogan et al. 2015). Given this evidence, we expect that an increase in local users' participation in information cascades will improve the cascades' rate of diffusion in social media networks. Future research in the context of cascades carrying disaster-related information in social media networks could assess empirically whether local users' participation in these cascades will contribute positively to the cascades' rate of diffusion.

Lastly, this research empirically tests theoretical propositions using data from a disaster that was not completely unexpected or unpredictable. However, some disasters that HOs must respond to occur without warning (e.g., earthquakes, terrorist attacks). Future research can analyze whether the theoretical propositions presented in this paper hold in the context of sudden-onset emergencies and whether additional factors specific to this setting impact the diffusion rate of information on social media platforms.

Inspired by the emergence of social media usage during disasters, our study examined the effectiveness of information propagation on social media platforms and identified factors that affected the rate of information diffusion. Beyond this context, commercial firms have also started to leverage social media to catalyze word-of-mouth marketing and enhance brand awareness and engagement (Hoffman and Fodor 2010). However, key differences exist regarding information cascades on social media in humanitarian versus commercial settings. For instance, in an anticipated event, such as the release of a new product, firms often initiate cascades and engage with consumers to

generate buzz. HOs and other stakeholders can also use social media platforms to share preparation information as forecasted disasters draw closer and intensify. However, commercial firms are better able to control and manipulate cascade formation and diffusion in these events since information typically originates from the firm and does not involve as many stakeholders as in humanitarian settings.

Firms also utilize social media as an information tool during unexpected events involving product and service failures. For example, firms in the electronics industry frequently monitor social media to identify information about hardware and software defects reported by consumers while firms in the transportation industry routinely use social media to trace information about unexpected service failure events. Cascades with this information are more likely to originate from dispersed geographical areas unlike cascades with information from victims of unexpected, sudden-onset disaster events (e.g., earthquakes, terrorist attacks), which can largely be traced to more limited geographical areas. While these characteristics help differentiate cascades on social media in commercial and humanitarian contexts, we encourage researchers to continue investigating cascade behavior to increase our understanding of how information disseminates on social media.

Acknowledgements
We are grateful to the special issue editors, the associate editor, and two anonymous reviewers for their insightful comments and guidance. We also would like to thank Hyun-Woo (Anthony) Kim and Tommy Benning from Gnip for their invaluable help with data collection. We would also like to acknowledge Bin Gu, Fred Morstatter, and Shamanth Kumar from Arizona State University for improving our understanding of information diffusion theories and the different properties of Twitter data.

CHAPTER 2

The Interaction of Similar Content on Social Media Platforms during Disasters

Abstract

Humanitarian organizations use social media platforms to communicate information about their work and services. To ensure that information reaches the intended audience before it expires, humanitarian organizations' content must diffuse rapidly. The focus of our study is exploring the diffusion speed of social media content. Our approach is novel since we also account for the influence of content that is simultaneously disseminating on a piece of content's propagation speed. Specifically, we evaluate if social media posts that carries the same meaning and is textually similar interact positively or negatively with one another. We formulate a generalized Hawkes model and evaluate the model using Twitter data from four distinct disasters. The results from our analysis indicate that similar content generally impedes the diffusion rate of a specific piece of content. However, the interaction can sometimes be cooperative in the sense that similar content can enhance the diffusion speed of a post. This research underscores the importance of incorporating the impact of concurrent content on social media platforms when analyzing diffusion speed. In addition, our findings carry implications for humanitarian organizations on how to coordinate with one another to amplify and jointly maximize the dissemination rate of each other's social media content.

1. Introduction

When a disaster occurs, the humanitarian community mobilizes itself to respond and provide aid to those in need. In such emergency situations, responders strive to minimize human suffering and to make decisions that lead to the greatest social good (Holguín-Veras et al. 2013). The effective management of humanitarian supply chains is crucial in achieving these objectives. An overarching goal is to coordinate flows of goods, funds, and information to ensure the availability and accessibility of required resources in the right quantities and at the right time and place (Van Wassenhove 2006). In this study, we focus on the management of information flows in humanitarian supply chains. While the other flows are important, the distribution of information is essential to make educated decisions about the movement of goods and finances. Furthermore, due to the extreme uncertainty that characterizes the humanitarian context, information has been cited as the most perishable resource during times of crisis (Meier 2015). Because the rate of information perishability is exacerbated by the turbulence of the operational environment often associated with disaster events, information may lose its accuracy and relevance within very short time periods.

As a result, humanitarian organizations (HOs) must leverage information networks to disseminate information rapidly before it expires. HOs have found internet-hosted social media platforms, like Facebook, Twitter, and Instagram, to be very effective for creating content and making it available instantaneously to other users connected through networks on these platforms. HOs routinely utilize social media platforms as a communication tool to broadcast donor appeals as well as updates on their work and services. Additionally, HOs frequently use social media platforms as sources of data from those affected by a disaster since stakeholders located within disaster zones can easily post

valuable reports on these platforms related to subjects like damaged infrastructure and injuries (Gao et al. 2011).

We concentrate on developing insights for HOs' usage of social media platforms as a communication tool. Our work responds to calls for research on the implications of the use of social media platforms for HOs (Holguín-Veras et al. 2012, Swaminathan 2018) and contributes to the growing body of literature that has explored this topic (e.g., Pedraza Martinez and Yan 2016, Yoo et al. 2016). Like Yoo et al. (2016), this study is specifically concerned with expanding our understanding of how HOs can enhance the diffusion speed of their social media content. By improving the rate at which their content on social media platforms disseminates, HOs can counter the problem of information perishability and transmit to more stakeholders their content before its expiration. After investigating Twitter data from Hurricane Sandy, Yoo et al. (2016) discovered that information diffusion slows down during times of high traffic, and we directly build on this work by analyzing how the diffusion speed of social media content is affected by the contemporaneous diffusion of peripheral content.

The diffusion of user-generated content is marked by sharing through social media platforms' sharing functions. By sharing a piece of content, users forward that information to their network, and propagation continues as long as the same content is shared. A social media post and its chain of shares can be viewed as a cascade, and cascades lengthen with more shares (Lerman and Ghosh 2010). The diffusion speed of a cascade reflects the rate at which the cascade grows. A cascade's likelihood of being shared depends on a number of variables, such as the number of connections that cascade participants have and the visibility of the cascade (Bakshy et al. 2011, Lerman and Ghosh 2010). Network effects also have been shown to play a role in that weak ties are more likely to share content (Shi et al.

2014). Content attributes, like the inclusion of URLs, influence a cascade's dissemination rate as well (Boyd et al. 2010).

Beyond the factors mentioned above, we argue that a cascade's diffusion speed is also a function of interactions with other cascades. Because there are no costs to generating content, there is a tremendous volume of cascades on social media platforms. Therefore, an analysis of a cascade's diffusion rate cannot view the cascade in isolation but must account for the influence imposed by other cascades. We are particularly interested in the interactions between cascades conveying essentially the same message through content that is textually similar. A cascade's propagation may benefit from the existence of cascades communicating a similar message since the content appears legitimate (Myers and Leskovec 2012). Alternatively, the presence of similar content in other cascades may introduce a competitive dynamic and render a specific cascade as redundant. As a result, the cascade may struggle to attract attention away from its competitors and be shared. The purpose of our study is to examine how a cascade's diffusion speed is affected by other cascades expressing similar content and what determines whether the effects by other cascades are competitive or cooperative.

While others have researched interactions among cascades (e.g., Coscia 2018, Myers and Leskovec 2012, Weng et al. 2012), we are the first to evaluate this phenomenon in the humanitarian setting. Consequently, we contribute to the research stream related to the interplay of information on social media platforms by analyzing this topic during situations that require the urgent and rapid diffusion of content. Our study also carries implications for the literature on the coordination of information resources in humanitarian operations (e.g., Altay and Pal 2014, Ergun et al. 2014). It can be expected that HOs sometimes issue similar social media content as that of other HOs, especially once an emergency has occurred and HOs converge to respond. Our research supplies

guidelines on how HOs can work together to coordinate the release of similar content such that the spread of their cascades benefit, rather than compete with, one another. This will help HOs jointly maximize the diffusion speed of their content across social media platforms and spread information quickly to their combined audiences.

Our research offers a methodological contribution to the literature as well. To analyze cascades' diffusion speed, we formulated a generalized point process model that is based on the Hawkes model (Hawkes 1971). Based on the history of shares, the Hawkes model calculates the intensity of a cascade, which can be interpreted as its diffusion rate (Zhao et al. 2015). This model is also known as a self-exciting point process since the intensity for a cascade increases every time that the cascade is shared. Our model is distinctive because we not only considered the effects of the cascade's own shares but also incorporated the effects of the shares of cascades with similar content. To the best of our knowledge, ours is the first point process model to calculate a cascade's dissemination speed as a byproduct of the diffusion history of cascades other than itself. Moreover, we allowed the latter effects to be positive or negative to model the possible cooperative and competitive effects of other cascades. This extension is not commonly implemented as it makes the model difficult to estimate. Another methodological contribution from our paper is the implementation of a near-duplicate detection algorithm called the *simhash* algorithm to cluster similar content. Thus far, near-duplicate detection techniques have not been utilized in the operations management literature. Using *simhash*, we were able to successfully and efficiently identify for each cascade what other cascades were carrying similar (i.e., near-duplicate) information.

We evaluated the model using Twitter data from four distinct disasters that unfolded in different parts of the world to increase the generalizability of our findings. Twitter is a prominent social media platform that is known for microblogging since posts,

or “tweets,” on this platform are limited in length. Twitter boasts approximately 330 million users that publish more than 500 million tweets per day[5]. On Twitter, users can share, or “retweet”, a tweet and distribute that tweet to their connections. Therefore, cascades comprised a tweet and its retweets, and our sample size included almost 27,000 cascades. Beyond Twitter’s popularity, we chose to collect data from this social media platform because of its value to HOs. The United Nations Office for the Coordination of Humanitarian Affairs (OCHA), for instance, published a policy brief related to HOs’ usage of social media (Moore and Verity 2014), and in this document, OCHA singled out Twitter as the social media platform best suited for HOs.

Our findings indicate that a cascade’s diffusion speed is affected by its own history of shares as well as the history of shares for cascades conveying similar content. Therefore, the dissemination of a cascade is not immune from the influence of other cascades belonging to the same topic. We also observed that the effect of cascades with similar information on a specific cascade’s diffusion rate varies across the cascades in our sample, and the range of this effect included both negative and positive values. This provides evidence of a competitive and cooperative dynamic among cascades. On average, however, a competitive effect was imposed by cascades sharing similar content, so our study suggests that HOs should attempt to produce novel information to avoid being clustered with other cascades based on similarity of content. We also conducted an additional analysis to identify determinants of whether a competitive or cooperative dynamic emerged among cascades. Our results reveal that the diffusion speed of cascades published by producers of larger size is more likely to benefit from the spread of cascades in the same topic. HOs may thus strive to increase their size on social media platforms. We also found that the number of cascades diffusing simultaneously to a focal cascade carrying content

[5] https://blog.hootsuite.com/twitter-statistics/

under a common topic has a curvilinear relationship with the focal cascade's diffusion rate. As the count of cascades carrying similar content increases, the impact is initially positive due to the content becoming validated but then becomes negative from a crowding effect. These results shed light on how cascades interact and provide guidance for HOs on what to expect concerning the diffusion of their social media content.

We organize the rest of the paper as follows. In Section 2, we position our paper in the extant literature, discuss its contributions to these areas, and outline key factors and conditions behind the dynamics of diffusion involving multiple cascades. Then, we formulate the point process model in Section 3 and provide an overview of the data in Section 4. Subsequently, we discuss how we estimated the model and present the results in Sections 5 and 6. We conclude in Section 7 with a summary of our findings and extensions of our study for future research to consider.

2. Background

Our paper contributes to the literature on information diffusion in a humanitarian context. For HOs, the diffusion of information to stakeholders like beneficiaries, donors, and other HOs is imperative to effectively prepare for and respond to a crisis. However, there are many obstacles for sharing information in a humanitarian setting. Disasters may damage infrastructure and the physical landscape, making it difficult to access data sources and to transmit information (Holguín-Veras et al. 2012). Moreover, many HOs converge at the scene of the disaster to work for response and recovery efforts, and the lack of centralized leadership hinders an understanding of each group's capabilities as well as the systematic information exchange among the involved parties (Day et al. 2012, Van Wassenhove 2006). Other reported challenges include unreliable data from inaccurate or untimely information and inconsistent data formatting from using different measurements and systems (Altay and Labonte 2014).

To combat the challenges mentioned above, HOs have embraced a UN system that organizes HOs into clusters based on their specialties. Each cluster has an appointed leader. Altay and Pal (2014) showed that the cluster system helps to facilitate information sharing, especially if cluster leads coordinate the flows of information and filter information to the proper HOs. Furthermore, HOs have collaborated with commercial firms to develop technological solutions that have standardized data collection procedures and reduced informational delays associated with manual data entry (Ergun et al. 2014). HOs have also embraced open platforms, like Sahana, that reduces barriers against inter-organizational information sharing. On these platforms, HOs can freely view crisis maps and share information about camps and missing persons (Currion et al. 2007). Finally, many HOs have become active users on social media platforms, which are also free available and are effective at diffusing information during emergencies (Yoo et al. 2016). Social media platforms are also particularly useful because HOs can broadcast information to their stakeholders as well as gather first-hand information posted by users located within the disaster zone (Starbird et al. 2010).

Our paper extends this research by examining how the diffusion of HOs' social media content is impacted by the spread of similar content during the same timeline. To that end, our study builds on the economy of attention literature. The concept of the economy of attention was first introduced in the seminal piece by Simon (1971). According to Simon (1971), we currently live in an information-rich economy as our lives are inundated with information, especially since the rise of the internet. In such an economy, the wealth of information leads to a scarcity of what information expends: attention. Therefore, producers must develop strategies for attracting attention, and consumers must determine how to distribute their attention resources among competing pieces of information (Falkinger 2007, Simon 1971).

Economies of attention consist of producers and consumers of information, and depending on the amount of available information, they can be characterized as information-rich or information-poor. Researchers have adopted this framework to examine a variety of problems involving competition for limited attention resources across different contexts. For example, Gabaix et al. (2006) studied how information acquisition for economic decision-making is affected by limited attention, and Haas et al. (2015) investigated how individuals select which problems to pay attention to and solve on online forums. This theoretical lens was also used to examine how animation can draw attention towards online ads (Hong et al. 2004). The economy of attention framework has been applied to the area of computer science as well to argue that users cannot process (or grant attention to) all information returned by search results. Accordingly, competing search results should be prioritized based on relevance and usefulness (Huberman and Wu 2008).

Recently, social media platforms like Twitter (Weng et al. 2012) and Digg (Wu and Huberman 2007) have been analyzed under the economy of attention model. These platforms are certainly information-rich and facilitate immense amounts of traffic because each user faces minimal costs to upload user-generated content. In this context, each user that creates a post and thereby launches a cascade is viewed as a producer of information. A producer then has to compete for users' attention against other producers, and it earns greater utility as more users pay attention to its content (Iyer and Katona 2015). One way that consumers can signal their attention to a piece of content is by propagating the cascade through the platform's sharing function (Wu and Huberman 2007). Hence, Twitter cascades that receive higher amounts of attention in the form of retweets experience greater diffusion. Attention allocation on social media platforms has been studied from an individual perspective (e.g., Hodas and Lerman 2012, Weng et al. 2012)

and an aggregate perspective (e.g., Ciampaglia et al. 2015, Huberman et al. 2009). The latter examines the distribution of collective attention across cascades, and the focus of our study is at the aggregate level.

A key factor we consider in the examination of diffusion dynamics among multiple cascades is the novelty of information presented in the cascades since novel content typically experiences greater diffusion (Vosoughi et al. 2018). The content communicated in a cascade will vary in terms of its degree of novelty, or conversely similarity, relative to the content conveyed in other cascades. To the extent that a group of cascades relays content similar to that of a focal cascade while diffusing within the same timeline, this set of cascades is considered to be running in "parallel" to the focal cascade. The information embedded in the focal cascade and its parallel cascades is also said to belong to the same topic. Prior research has found that the presence of parallel cascades makes it difficult for a focal cascade to collect attention and diffuse (Coscia 2014). Other research, however, has found evidence that a focal cascade's diffusion is enhanced by the existence of parallel cascades, perhaps because the proliferation of the content in multiple cascades makes the information appear more important and valid (Myers and Leskovec 2012). In this study, we aim to deepen our understanding of the interplay between a focal cascade and its parallel cascades and identify when the effects of the latter are positive versus negative. In order to understand some of the reasons driving the possible manifestation of these contradictory results in a humanitarian setting, we follow Coscia (2018) and Dellarocas et al. (2015) and examine the diffusion dynamics between focal and parallel cascades. Our aim is to identify when the latter have positive versus negative (or cooperative versus competitive) effects on the diffusion of the former.

Several conditions can determine whether parallel cascades detract from or attract attention to a focal cascade. First, the diffusion of focal cascades supplied by large

producers may be less susceptible to competitive effects by parallel cascades. This is because large producers can generate stronger and more extensive information signals, and thus, they are capable of tapping into a greater pool of attention resources (Falkinger 2007). Because larger producers also tend to be viewed as more credible (Castillo et al. 2011), it may be easier for their content to become validated by the presence of parallel cascades. Second, the volume of parallel cascades will determine the extent to which cascades will contribute or undermine attention to a focal cascade. Haas et al. (2015) discovered a curvilinear relationship between the number of parallel cascades and a focal cascade's diffusion. As the volume of parallel cascades rises, the focal cascade's dissemination increases because its content appears more interesting and becomes legitimized from other cascades carrying similar information. At some point, however, the topic may become too crowded, and a high number cascades running in parallel will make it difficult for the focal cascade to distinguish itself and earn attention to diffuse.

Another consideration is the timing of when the focal cascade started relative to when the parallel cascades were initiated. Specifically, a cascade's diffusion may be subject to a first-mover advantage if the same cascade's producer is the first among producers of cascades in the same topic to broadcast the topic's content. A first-mover advantage over traditional news outlets has been observed for blogs that react immediately to an event. Such blogs are able to direct attention towards themselves and steer public opinion (Farrell and Drezner 2008). In contrast, Ciampaglia (2015) has demonstrated that content attracts more attention when it is issued during, rather than before, the period of peak interest in the topic that the piece of content pertains to. Since it generally requires some time to generate interest in a topic, a first-mover advantage may not exist for a focal cascade with parallel cascades.

3. Point Process Model for the Diffusion of Cascades

We consider Twitter cascades indexed by $i = 1, \dots, I$ during the observation interval $[\mathbf{0}, \boldsymbol{T}]$. Upon publishing a tweet, cascade i is launched by producer p_i, and we label the time that the cascade was initiated as t_0^i, where $t_0^i \geq \mathbf{0}$. Cascades on Twitter grow as they are retweeted by other users, or by retweeters. We only include cascades that have been retweeted at least once to guarantee minimum diffusion. Cascade i comprises $k = 1, \dots, K$ retweets, and the times that these retweets arrived are denoted as $t_1^i, \dots, t_K^i$, where $t_K^i \leq \boldsymbol{T}$. Therefore, the time that retweet k of cascade i occurred is equal to t_k^i.

In our study, we follow Zhao et al. (2015) and model a cascade's diffusion based on the occurrence of retweets as a point process. Point processes are a collection of stochastic points that represent the occurrence of an event along a finite line or space. Examples of events modeled as point processes include advertisement clicks (Xu et al. 2014), earthquakes (Ogata 1988), and crime (Mohler et al. 2011). Retweets for a cascade are described as a point process, or as a series of points along a finite and nonnegative line that represents time. A point process can also be characterized through a counting measure, $R^i(t)$, which gives the number of retweets that cascade i has accumulated by t. This means that $R^i(t_k^i) - R^i(t_{k-1}^i)$ corresponds to the number of retweets that materialized for i between $(t_{k-1}^i, t_k^i]$. We note that $R(0) = 0$. The counting measure is increasing and integer-valued, making it a step function that increases by a value of 1 at every t_k^i (Daley and Vere-Jones 2003).

The simplest type of point processes is the Poisson process. Under the Poisson process, event occurrences transpire independently at a mean rate, or intensity, equal to λ. The intensity is assumed to be constant across time, and consequently, this process is also called the homogeneous Poisson process. The homogeneous Poisson process is useful to model the arrival rates of points belonging to an event that conform to this assumption,

but there are many events for which this assumption is too restrictive. As such, the inhomogeneous Poisson process allows the intensity to vary over time, which can be expressed as $\lambda(t)$. The event realizations are still assumed to be independent under the inhomogeneous Poisson process (Daley and Vere-Jones 2003). Because event occurrences are treated as independent, Poisson processes are sometimes referred to as being "memoryless" (Gardner et al. 1995).

It may be the case, however, that the realization of an event is dependent on previous realizations. Such point processes cannot be modeled by the either of the Poisson processes mentioned previously since the assumption of independence is violated. This property of dependence among event observations has been observed within the context of social media platforms, such as Twitter (Kobayashi and Lambiotte 2016, Zhao et al. 2015) and YouTube (Crane and Sornette 2008). Accordingly, we utilize a point process model that allows the arrival of a cascade's retweets to be influenced by earlier retweets.

The self-exciting point process, also known as the Hawkes process, is able to handle dependence among event occurrences by specifying the intensity as a conditional function of time and the history of the point process (Hawkes 1971). The history of the point process until t encompasses information about all realizations prior to t as well as the times that the realizations happened, and we express this variable as $\mathcal{H}_t^i$ (Daley and Vere-Jones 2003). The conditional intensity function for cascade i is formally defined as:

$$\lambda^i\left(t\middle|\mathcal{H}_t^i\right) = \lim_{t\to 0} \frac{\Pr\left\{R^i(t+\triangle t) - R^i(t) > 0\middle|\mathcal{H}_t^i\right\}}{\triangle t}, \tag{1}$$

where $\lambda^i\left(t\middle|\mathcal{H}_t^i\right) > 0$. Within our context, the intensity represents the rate at any moment that a cascade is retweeted, conditional on the history of past retweets. The intensity can alternatively be interpreted as the diffusion rate for a cascade.

In the self-exciting point process by Hawkes (1971), every event realization increases the conditional intensity function in an additive fashion. This means that the

occurrence of one retweet heightens the cascade's diffusion speed and makes the arrival of the next retweet faster. The self-exciting point process for i is equal to:

$$\lambda^i\left(t\middle|\mathcal{H}_t^i\right) = \mu^i e^{-\gamma^i t} + \int_{-\infty}^{t} g^i(t-s)dR^i(s), \tag{2}$$

where $\mu^i > 0, \gamma^i > 0$, and $s < t$. Here, μ^i is the homogeneous Poisson process rate that represents the baseline intensity of the cascade (Hawkes and Oakes 1974). We allow μ^i to decay exponentially over time to reflect the temporal decay patterns of cascades on Twitter (Asur et al. 2011), and the decay rate is parametrized by γ^i. Furthermore, μ^i and γ^i are heterogeneous across cascades since we anticipate variation in how easily cascades disseminate and how quickly interest in them declines.

The other component of the self-exciting point process describes the impact of a retweet at time s on cascade i's diffusion speed at time t. This exciting effect is not permanent but wears off over time. As is common in extant research (e.g., Embrechts et al. 2011, Xu et al. 2014), we specify the effect of previous realizations to decay exponentially:

$$g^i(t-s) = \alpha^i e^{-\beta^i(t-s)}, \tag{3}$$

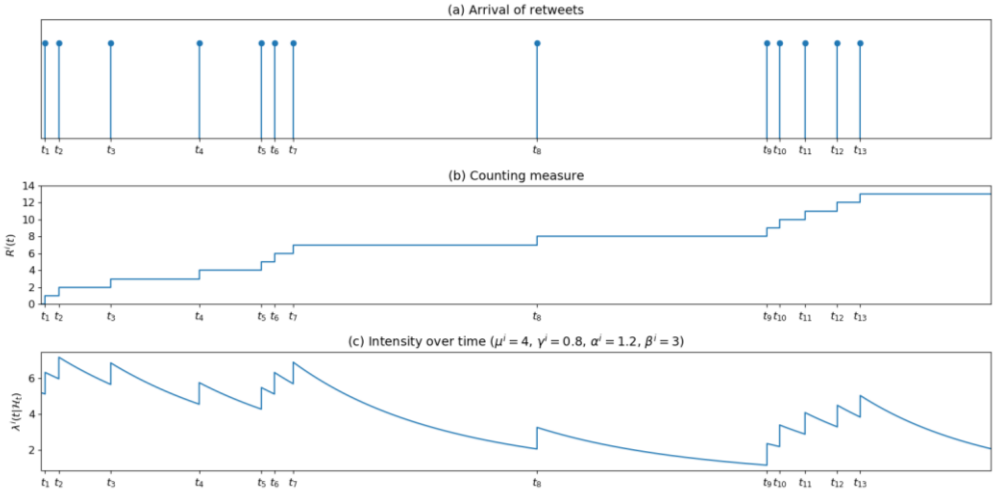
where $\alpha^i > 0$ and $\beta^i > 0$. We also enforce the restriction $\alpha^i < \beta^i$ (Hawkes 1971, Masuda et al. 2013). The parameter α^i represents the exciting effect, or the increase in intensity, attributed to retweet of i at s, and β^i reveals how quickly such an effect dissipates. Note that α^i and β^i are cascade-specific to model the heterogeneity of exciting effects across cascades. Given this information, Equation 2 can be rewritten as:

$$\lambda^i\left(t\middle|\mathcal{H}_t^i\right) = \mu^i e^{-\gamma^i t} + \int_{-\infty}^{t} \alpha^i e^{-\beta^i(t-s)}dR(s) = \mu^i e^{-\gamma^i t} + \sum_{t_k^i<t} \alpha^i e^{-\beta^i(t-t_k^i)} \tag{4}$$

From Equation 4, it is clear that a cascade's intensity at time t is a function of the sum of the exciting effects imposed by all retweets that occurred before t. Figure 2 is based on a figure by Rizoiu et al. (2017) and illustrates an example of a cascade i's self-exciting point

process in three panels. Panel (a) portrays the arrival of retweets as points at the time that they occurred. Panel (b) shows the counting measure as a step function increases as retweets arrive in Panel (a). Lastly, Panel (c) depicts the intensity over time given that $\mu^i = 4, \gamma^i = 0.8, \alpha^i = 1.2,$ and $\beta^i = 3$.

Figure 2 – Self-Exciting Point Process for a Sample Cascade



However, recall that a cascade does not diffuse in isolation, and its dissemination may be susceptible to influence from other cascades. As discussed in Section 2, parallel cascades are cascades that spread during the same timeline while carrying similar content to that of a focal cascade. These are of particular interest as they have been shown to both impede and accelerate a focal cascade's diffusion. Thus, we modify the self-exciting point process to include another point process that represents the arrival of retweets belonging to parallel cascades. This is similar to a multivariate point process in which the intensity of each process is affected by all other point processes under consideration (Hawkes 1971). Our model is not a multivariate point process since we only consider how the retweets of parallel cascades impact the intensity of a focal cascade and exclude the inverse relationship.

Under the modified model, $R^i(t) = \left[R_1^i(t), R_2^i(t)\right]$, where $R_1^i(t)$ is the counting measure for retweets belonging to cascade i and $R_2^i(t)$ is the counting measure for retweets

belonging to parallel cascades of i. The retweets of parallel cascades are indexed by $l = 1, \dots, L$, and the time that retweet l occurred is marked as t_l^i. The time that L was issued is t_L^i, and $t_L^i \leq T$. In addition, we introduce two new terms, $\phi_{t_k^i}$ and $\phi_{t_l^i}$, which respectively measure the natural logarithm of the number of followers that the retweeter of k had at t_k^i and that the retweeter of l had at t_l^i. The follower counts are logged to address skewness. Like Mishra et al. (2016) and Zhao et al. (2015), we include retweeters' follower counts to account for the change in intensity from retweeters with higher follower counts exposing a larger audience to the original piece of content. Equation 5 presents the model that includes both point processes:

$$\lambda^i\left(t\middle|\mathcal{H}_t^i\right) = \mu^i e^{-\gamma^i t} + \sum_{t_k^i<t}\left(\alpha_{11}^i * \phi_{t_k^i} * e^{-\beta_{11}^i(t-t_k^i)}\right) + \sum_{t_l^i<t}\left(\alpha_{21}^i * \phi_{t_l^i} * e^{-\beta_{21}^i(t-t_l^i)}\right) \quad (5)$$

We differentiate the exciting effects of i's own retweets and the retweets of parallel cascades by having α_{11}^i and β_{11}^i characterize the former and α_{21}^i and β_{21}^i characterize the latter. Since we incorporate $\phi_{t_k^i}$ and $\phi_{t_l^i}$, the parameters α_{11}^i and α_{21}^i represent the magnitude of the effect of retweets of the corresponding point processes while controlling for retweeters' follower counts. As before, $\alpha_{11}^i, \beta_{11}^i > 0$, and $\alpha_{11}^i < \beta_{11}^i$. Moreover, we add the constraints $\beta_{21}^i > 0$ and $\alpha_{21}^i < \beta_{21}^i$, but α_{21}^i is not restricted to be positive-valued. This allows us to model the effects of parallel cascades' retweets on a focal cascade's intensity to be both exciting and inhibitive (Bowsher 2007, Mei and Eisner 2017). In doing so, we implement a more generalized version of the Hawkes model, and we acknowledge that the diffusion of parallel cascades may compete against or cooperate with a focal cascade's diffusion rate.

4. Data

4.1. Sample

A common dimension on which to classify disasters is the amount of warning time that is possible before the events occur. Sudden-onset disasters are those that transpire instantly with no warning (e.g., earthquakes, industrial accidents) while slow-onset disasters are those with gradual and foreseeable arrivals (e.g., hurricanes, floods) (Holguín-Veras et al. 2012, Olteanu et al. 2015). We obtained Twitter data for four sudden-onset disasters from WeLink, which is a social media data services firm. We chose to concentrate on sudden-onset disasters since these have a finite starting point. The four disasters were sampled from EM-DAT[6], which is a database of disaster events hosted and maintained by the Centre for Research on the Epidemiology of Disasters (CRED) at the Université Catholique de Louvain. This database has been employed to sample disasters in previous publications (e.g., Acimovic and Goentzel 2016, Sodhi 2016). Humanitarian events were only eligible to be sampled if they occurred between 2009 and 2015 because 2009 is approximately when Twitter started experiencing rapid growth in the number of its users and when researchers began to study how Twitter can be used during emergencies. Additionally, we limited our sample to disasters from regions that spoke English or Spanish to avoid any need for translation. Table 7 provides information about the disasters that were selected for our sample, including data on the number of casualties and the size of the affected population.

(Table 7 on next page)

[6] D. Guha-Sapir, R. Below, Ph. Hoyois - *EM-DAT: International Disaster Database – www.emdat.be – Université Catholique de Louvain – Brussels – Belgium.*

Table 7 – Information on Sampled Disasters

Disaster	Event Location	Event Time (UTC)	End of Data (UTC)	Total Deaths*	Total Affected*
Joplin tornado	Joplin, MO, USA	5/22/2011 22:34	6/2/2011 23:59	176	1,150
Black Forest fire	Black Forest, CO, USA	6/11/2013 19:00	6/21/2013 23:59	2	1,617
Lac-Megantic rail disaster	Quebec, Canada	7/6/2013 05:15	7/10/2013 23:59	47	2,000
2014 Iquique earthquake	Iquique, Chile	4/1/2014 23:46	4/6/2014 23:59	6	513,837

* *Source: EM-DAT*

To collect the data, we submitted to WeLink a set of queries specific to each disaster event. These queries comprised keywords and phrases that were commonly present in hashtags and content associated with the emergencies. We also specified the date ranges that we were interested in, starting from the time the disaster materialized to approximately the end of the response period. As the selected events were sudden-onset disasters, we were able to clearly delineate if content was published before or after the disasters. The precise end of the data collection period is shown in the fourth column of Table 7. WeLink collected all tweets and retweets that were issued within the stipulated timeline and contained the keywords and phrases (not case sensitive) anywhere within the body of the text, including within hashtags. We followed Olteanu et al. (2015)'s approach to selecting the keywords and phrases. That is, we detected keywords in hashtags by searching on Google for "hashtag" in conjunction with the name of the event. We also included in our queries combinations of the location of the disaster and the event name. Table 8 lists the exact keywords and phrases that we used to collect Twitter data through WeLink.

(Table 8 on next page)

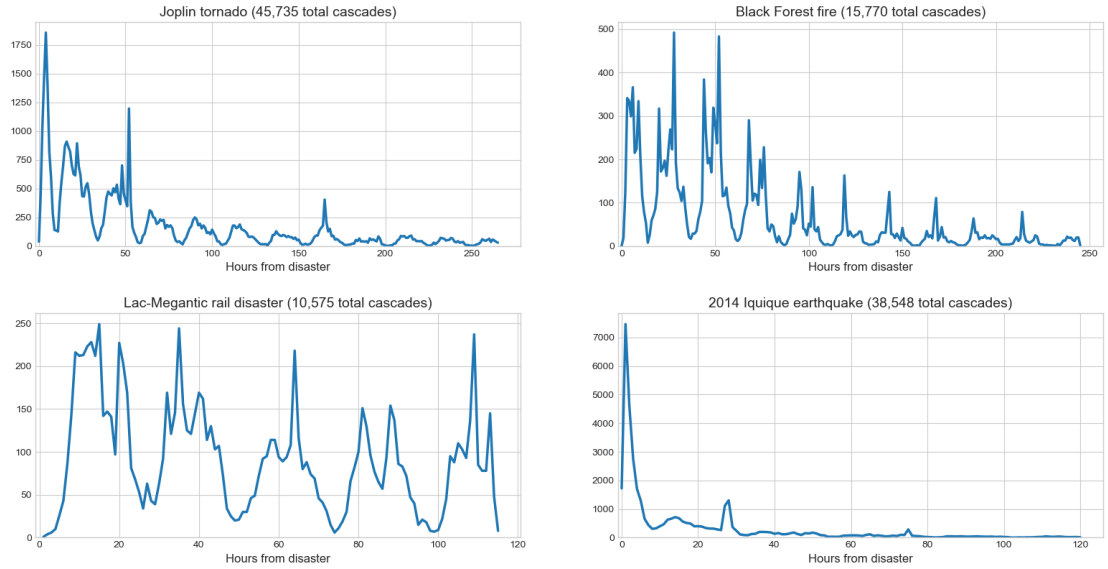
Table 8 – Keywords and Phrases in Queries

Joplin tornado	Black Forest fire	Lac-Megantic rail disaster	2014 Iquique earthquake
joplin	blackforestfire	lacmegantic	iquique temblor
prayersforjoplin	colorado fire	lacmégantic	iquique terremoto
joplintornado	black forest	megantic3rec	iquique earthquake
joplinmissouri			chile earthquake
joplinmidmo			chile temblor
			chile terremoto

The data set provides detailed information about the tweets and retweets that matched our queries, such as timestamps and profile statistics for the users that issued the tweets and retweets. We can identify which tweet was being shared by every retweet since the tweet ID is recorded for all retweets. Subsequently, we organized the tweets and retweets into cascades in accordance with previous studies (e.g., Lerman and Ghosh 2010, Vosoughi et al. 2018). A cascade is composed of a user's tweet and its chain of retweets, and we label the user that posted the tweet as the cascade's producer to be consistent with the terminology presented in Sections 2 and 3. At a minimum, a cascade was required to have gained at least one retweet. During the years when the four disasters took place, Twitter limited the amount of characters in a tweet to 140, so tweets were often short messages. In line with previous researchers' guidelines (Davidov et al. 2010, Wang et al. 2012), we eliminated any cascades with text containing less than five words because tweets with too few words are difficult to extract meaning from. Based on these considerations, we obtained a total of 110,628 cascades across all of the events in our sample. In Figure 3, we illustrate the number of cascades initiated over time, and each of the panels corresponds to a disaster.

(Figure 3 on next page)

Figure 3 – Count of Cascades Initiated over Time



4.2. Parallel Cascades

For each cascade, we detected its parallel cascades (i.e., cascades carrying similar content). We were interested in a narrow view of similarity so that each cascade and its associated parallel cascades represent fine-grained rather than broad topics. This allowed us to model the effects of parallel cascades on the diffusion of a focal cascade in a more nuanced way. To find very similar pieces of content, we applied near-duplicate detection techniques, which rely on identifying similar content based on a measure of the textual distance between cascades. The Jaccard similarity coefficient is a commonly used similarity score for a pair of cascades, and it is calculated as the intersection of two cascades' words divided by the union of two cascades' words. Thus, a Jaccard similarity coefficient of 1 means that the two cascades are textually identical while a score of 0 means that the two cascades have no words in common (Manning and Schütze 1999). This distance measure requires pairwise comparisons among all possible pairs of the cascades. Because the number of pairwise comparisons grows exponentially, calculating the Jaccard similarity coefficient for all cascades in the sample is too computationally complex.

As such, we used the *simhash* algorithm, which was developed by Charikar (2002), to more efficiently locate parallel cascades. This algorithm has been implemented by Google to ascertain whether a web page is a near-duplicate of another page while web crawling (Manku et al. 2007). We briefly describe the algorithm here. *Simhash* is a dimensionality reducing algorithm that creates one B-bit fingerprint to represent a document (i.e., in our study, a cascade's text). To implement the algorithm, we performed the following steps for each cascade. First, we maintained a vector V of length B, and each element of this vector was initialized to equal 0. The subsequent step of the algorithm was to calculate a B-bit hash for every document feature. We chose to tokenize cascades' text into words and submit tokens as features. Next, we regarded hash values equal to 1 as 1 and hash values equal to 0 as -1. For $b = 1, \ldots, B$, we summed the hash values in the bth bit across the tokens, and we set the bth element of V equal to this sum. Negative sums in V were recorded as 0 while positive sums in V were marked as 1. The fingerprint of the cascade's text is equal to V. The *simhash* algorithm's performance is fast and scales linearly with the number of cascades. Moreover, this algorithm is particularly useful for finding near-duplicates because it produces similar hashes for similar content. Hence, textual similarity can be evaluated by comparing a cascade's fingerprint with that of another cascade. The Hamming distance, which is measured as the number of differing bits between two cascade's fingerprints, is often used for this task, and a low Hamming distance is correlated with a high Jaccard similarity coefficient.

A cascade's content was represented by the text extracted from the tweet that launched the same cascade. Technically, the text in retweets only differs from the text in tweets by crediting the cascade producer at the beginning with "RT@username", where "username" equals the producer's Twitter handle. Before applying the algorithm, we preprocessed the text. First, we converted all of the cascade's text to lowercase, and any

punctuation marks were removed. We stripped the text of URLs and emojis, but we preserved hashtags as long as they did not match the queried keywords and phrases. Also, we eliminated all English and Spanish stop words, which are common, short function words like “and”, “the”, and “which”. We attempted to reduce variation in users’ spelling by modifying any words with characters repeated more than three times in a row to having the characters repeated only three times in a row (i.e., “hahaaaa” to “hahaaa”).

Once text preprocessing was complete, we ran a Python implementation of *simhash*[7] for each disaster’s collection of cascades. This implementation generated 64-bit fingerprints for cascades. A cascade was deemed to be a near-duplicate of another cascade if the Hamming distance of their fingerprints was not larger than 8 bits. Please refer to Table 9 for examples of near-duplicates identified by the *simhash* algorithm. Locating near-duplicates is critical for identifying similar content, but we must consider duplicates as well. Duplicate detection only involves searching for exact matches, and this process is much easier and does not require the application of an algorithm. Therefore, for each cascade in our sample, we found its near-duplicates (if any) and duplicates (if any), and the joint set of near-duplicate and duplicate cascades constituted the set of parallel cascades. Since our study is concerned with the interactions between a focal cascade and its parallel cascades, we only kept cascades that matched with at least one near-duplicate or duplicate cascade. Our final sample consisted of 26,896 cascades, so approximately 24.3% of the 110,655 cascades communicated content that was textually similar to content carried by at least one other cascade.

(Table 9 on next page)

[7] https://github.com/seomoz/simhash-py

Table 9 – Examples of Cascades and Their Near-Duplicates

Disaster	Sample Cascade Text	Sample Near-Duplicate Text
Joplin tornado	You can help us respond in #Joplin! Text REDCROSS to 90999 to make a $10 donation, or give online: http://ht.ly/50NRD	To help those in #joplin text REDCROSS to 90999 to make a $10 donation.
Black Forest fire	REMINDER: MANDATORY EVACUATION means you are in immediate danger. Load your family and pets , and GO NOW. #BlackForestFire	"@EPCSheriff CLARIFICATION: MANDATORY EVACUATION means you are in immediate danger. Load your family and pets and GO NOW. #BlackForestFire"
Lac-Megantic rail disaster	Train Carrying Crude Oil Derails in Quebec http://t.co/e5jiBmTKux	Crude Oil-Carrying Train Derails And Explodes in Quebec Town http://t.co/gDJ7MI7b7p via @thinkprogress | http://t.co/tIwgnl9ClX #nokxl
2014 Iquique earthquake	Major Earthquake Strikes Off Chile Coast, USGS Reports http://t.co/3wwy4gJOox	Strong earthquake strikes off coast of Chile http://t.co/916gJ3BG1d

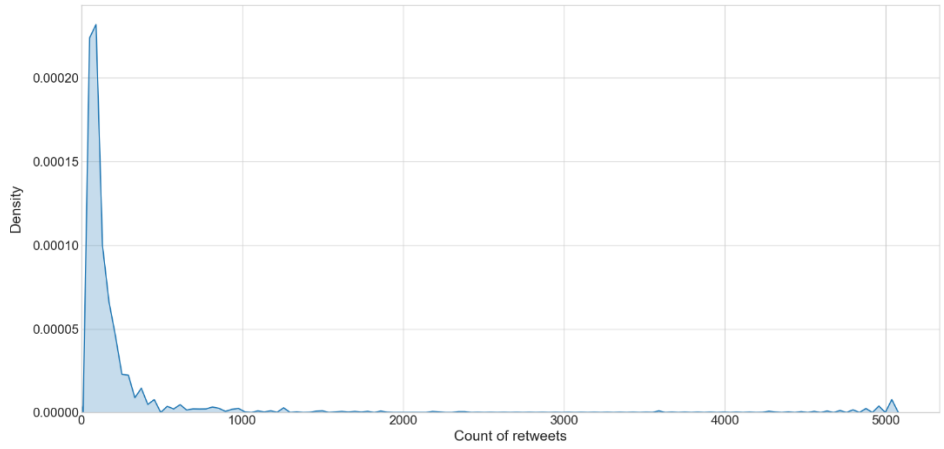
Because the unit of analysis in our study is the cascade, we organized the data for each cascade into two sets of arrivals, which listed retweets for the same cascade and retweets of the parallel cascades. As we combined all of the retweets of parallel cascades, we considered the effects of similar cascades on a focal cascade's diffusion in an aggregate form. The average number of retweets that a cascade accumulated during the data collection period was 9.779, and the mean number of retweets earned by parallel cascades over the identical time horizon was 99.799. The second statistic is higher because we aggregated the retweets across all of the parallel cascades that the focal cascade matched with. Table 10 gives a breakdown of the sample size by disaster along with the mean number of retweets in both point processes.

Table 10 – Breakdown of Sample Size and Retweets by Disaster

Disaster	Cascade Count	Mean Retweet Count for Focal Cascade	Mean Retweet Count for Parallel Cascades
Joplin tornado	9,868	6.428	64.875
Black Forest fire	2,281	4.749	14.664
Lac-Megantic rail disaster	1,951	7.112	25.336
2014 Iquique earthquake	12,796	13.666	153.261

On Twitter, the diffusion of cascades is generally small and short (Goel et al. 2016). Kwak et al. (2010), for example, found that over 90% of cascades only had been retweeted once. Our data exhibits a similar pattern. While the mean count of retweets for cascades in our sample was almost 10, the median was 2. Figure 4 portrays the kernel density plot of the number of retweets accumulated by every cascade in our study. From Figure 4, we can clearly observe that the distribution of cascades' retweet amounts is heavily right-skewed and resembles a power law distribution. The distribution of cascades' retweet counts by disasters resembled the distribution exhibited in Figure 4. The longest cascade was retweeted 5,047 times, which is more than 2.5 times more retweets than the second-longest cascade. This cascade was produced by the Spanish-language division of CNN and broadcasted information about which countries received tsunami warnings after the 2014 Iquique earthquake.

Figure 4 – Kernel Density Plot of Cascades' Retweet Counts



5. Model Estimation

We estimated the parameters for the model presented in Equation 5 using a maximum likelihood estimation procedure. The model parameters were estimated individually for every cascade in our sample (i.e., I = 26,896). Therefore, for cascade i, we estimated the vector of parameters $\theta^i = (\mu^i, \gamma^i, \alpha_{11}^i, \beta_{11}^i, \alpha_{21}^i, \beta_{21}^i)$. We created the counting

measures $R_1^i(t)$ and $R_2^i(t)$ based on the arrivals of retweets for i and i's parallel cascades respectively. The data for t_k^i and t_l^i were obtained from the timestamp information of the same set of arrivals. Across all i, the number of realizations in the first point process was equal to 263,005 and in the second point process was equal to 2,684,189. We measured t_k^i and t_l^i as the number of hours elapsed between when k and l occurred and t_0^i, where t_0^i was equal to the difference in hours from the time of i's launch to the start of the disaster. Lastly, the profile statistics of the retweeters record the number of followers that retweeters possessed at the moment that they issued any retweets in our data set. We relied on this data to evaluate $\phi_{t_k^i}$ and $\phi_{t_l^i}$.

The observation interval $[\mathbf{0}, \boldsymbol{T}]$ was the data collection period for the disaster that i belonged to. The time when the disaster transpired corresponded to $\mathbf{0}$, and $\boldsymbol{T}$ was calculated as the number of hours between $\mathbf{0}$ and the end of data collection (see Table 7 for details). Because the observation interval covered the entire data collection timeline, $R_2^i(t)$ may have included points that arrived between $\mathbf{0}$ and t_0^i or points that arrived after t_K^i. We maintained such realizations of $R_2^i(t)$ to account for the influence of parallel cascades' retweets not only during but also before and after i's lifetime. The conditional intensity function for i, however, is technically null prior to t_0^i. Consequently, we evaluated the conditional intensity function from $[t_0^i, \boldsymbol{T}]$. Time was treated as a continuous variable in this study, and this continuous-time framework enabled us to capture any time effects (Xu et al. 2014). Given the realizations of $R_1^i(t)$ and $R_2^i(t)$ during $[t_0^i, \boldsymbol{T}]$, the likelihood function for cascade i is as follows:

$$\mathcal{L}_i = \left[\sum_{k=1}^{K} \lambda^i\left(t_k^i \middle| \mathcal{H}_{t_k^i}^i\right)\right] * \exp\left(-\int_{t_0^i}^{\boldsymbol{T}} \lambda^i\left(t \middle| \mathcal{H}_t^i\right) dt\right) \quad (6)$$

Recall that we formulated a generalized point process model by permitting α_{21}^i to have an inhibitory effect on the intensity. After summing over the history of the cascade,

it is possible that the intensity at t becomes negative if α_{21}^{i} takes on a negative value. However, by definition, $\lambda^{i}(t|\mathcal{H}_t^i)$ must be positive (Daley and Vere-Jones 2003). To guarantee that the intensity is always non-negative, we executed the following nonlinear specification of our model, which was also applied in Bremaud and Massoulie (1996) and Reynaud-Bouret and Schbath (2010):

$$\tilde{\lambda}^{i}(t|\mathcal{H}_t^i) = \max(\lambda^{i}(t|\mathcal{H}_t^i), 0) \tag{7}$$

Under the nonlinear specification, the likelihood function for i now becomes:

$$\tilde{\mathcal{L}}_i = \left[\sum_{k=1}^{K} \tilde{\lambda}^{i}\left(t_k^i \middle| \mathcal{H}_{t_k^i}^i\right)\right] * \exp\left(-\int_{t_0^i}^{\boldsymbol{T}} \tilde{\lambda}^{i}(t|\mathcal{H}_t^i)\, dt\right) \tag{8}$$

The log-likelihood to estimate θ^{i} given the observed data for cascade i is presented in Equation 9.

$$\begin{aligned}\mathcal{LL}_i &= -\int_{t_0^i}^{\boldsymbol{T}} \tilde{\lambda}^{i}(t^i|\mathcal{H}_t^i)dt + \int_{t_0^i}^{\boldsymbol{T}} \log \tilde{\lambda}^{i}(t_k^i|\mathcal{H}_{t_k}^i)\, dR_1(t) \\ &= -\int_{t_0^i}^{\boldsymbol{T}} \tilde{\lambda}^{i}(t^i|\mathcal{H}_t^i)dt + \sum_{k=1}^{K} \log \tilde{\lambda}^{i}(t_k^i|\mathcal{H}_{t_k}^i)\end{aligned} \tag{9}$$

To reduce the dimensions of the functional space that the parameters can be estimated from, we used a penalized maximum likelihood function (Reynaud-Bouret and Schbath 2010, Zhou et al. 2013). We imposed the L2 regularization technique, which is also known as a ridge regression. The L2 regularization technique shrinks estimations of parameters as it penalizes the parameters based on their size. The penalty is equal to the tuning parameter, ρ, multiplied by the sum of the squared coefficients. The tuning parameter controls the amount of the penalty such that a larger tuning parameter leads to a higher penalty and more shrinkage (Hastie et al. 2009). The penalized log-likelihood function that we estimated for i is:

$$\overline{\mathcal{LL}}_i = -\int_{t_0^i}^{\boldsymbol{T}} \tilde{\lambda}^{i}(t^i|\mathcal{H}_t^i)dt + \sum_{k=1}^{K} \log \tilde{\lambda}^{i}(t_k^i|\mathcal{H}_{t_k}^i) - \rho \sum(\theta^{i})^2 \tag{10}$$

We maximized the penalized log-likelihood function each of the cascades in our sample using R. In order to make sure that we reached the global maximum, we provided

three different vectors of starting values and estimated the parameters using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. This optimization algorithm is an efficient quasi-Newton method that has been proven to reach global convergence (Fletcher 2013). Details about the integration of the first term in the penalized log-likelihood function are provided in Appendix B. Depending on the coefficients and the data, it was sometimes difficult to solve the integral analytically. In those cases, we numerically approximated the integral using a quadrature rule. We also note that the computation of the second term in Equation 10 is infeasible when $\tilde{\lambda}^i\left(t_k^i|\mathcal{H}_{t_k}^i\right)$ equals 0. Thus, we set $\tilde{\lambda}^i\left(t_k^i|\mathcal{H}_{t_k}^i\right)$ equal to ε, or the smallest positive decimal number in R, when the conditional intensity function was negative.

6. Results of Model Estimation

Using the estimation approach discussed in Section 5, we obtained parameter estimates that characterized the point process model for each cascade. The optimization algorithm was unable to converge for 58 cascades, reducing our sample size to 26,838 cascades. As this was a low percentage of the count of cascades that we attempted to optimize (58/26896 = 0.22%), the estimation procedure and results are still valid. In Table 11, we show the descriptive statistics for the parameter estimates in θ^i across all cascades. Due to space constraints, we omitted the parameter estimates for each cascade, but these are available from the authors upon request. Furthermore, we provide a breakdown of the descriptive statistics in Table 11 by disaster in Appendix C.

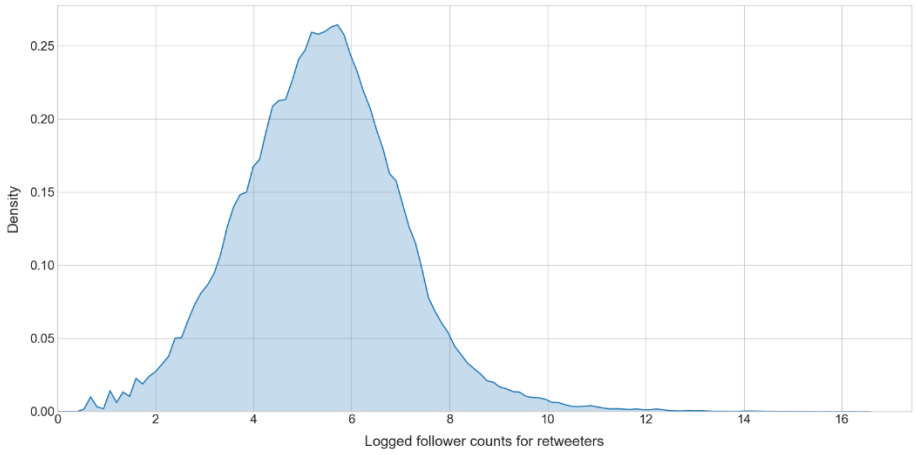
Table 11 – Descriptive Statistics for Parameter Estimates

	Mean	Median	Std. Dev.	Min.	Max.
α_{11}^i	0.080	0.001	0.212	7.27E-14	3.827
β_{11}^i	0.927	0.799	0.983	2.00E-06	18.632
α_{21}^i	-0.134	-0.009	0.320	-4.59E+00	3.356
β_{21}^i	0.571	0.355	0.641	1.79E-07	9.882
μ^i	0.576	0.575	0.505	1.74E-06	8.128
γ^i	0.356	0.335	0.298	7.83E-07	3.995

26,838 observations

According to Table 11, the mean value of α_{11}^{i} is 0.080, and the mean value of α_{21}^{i} is -0.134. These parameters respectively represent the effects of retweets of a focal cascade and of its parallel cascades on the focal cascade's intensity, after controlling for the logged count of retweeters' followers. Retweeters possessed 1,951 followers on average at the time of their retweets, and the retweeter with the highest count of followers in our sample was followed by 12,381,846 users. The extreme range of follower counts is the reason we logged the follower counts in the point process model. Figure 5 illustrates the kernel density plot of retweeters' logged follower counts.

Figure 5 – Kernel Density Plot of Logged Follower Counts for Retweeters



The mean values of α_{11}^{i} and α_{21}^{i} demonstrate that the effect of parallel cascades' retweets on the intensity of a focal cascade is negative on average. That is, on average, a focal cascade's diffusion rate is inhibited by the arrival of retweets for other cascades belonging to the same topic. We therefore find support for the existence of a competitive dynamic among cascades carrying similar content, which suggests that HOs should aim to produce novel content to avoid the negative effects imparted by parallel cascades. At the same time, however, the descriptive statistics for α_{21}^{i} indicate that the parameter is positive for some cascades. This provides evidence of the existence of a cooperative dynamic among cascades and their parallel cascades as observed in Myers and Leskovec

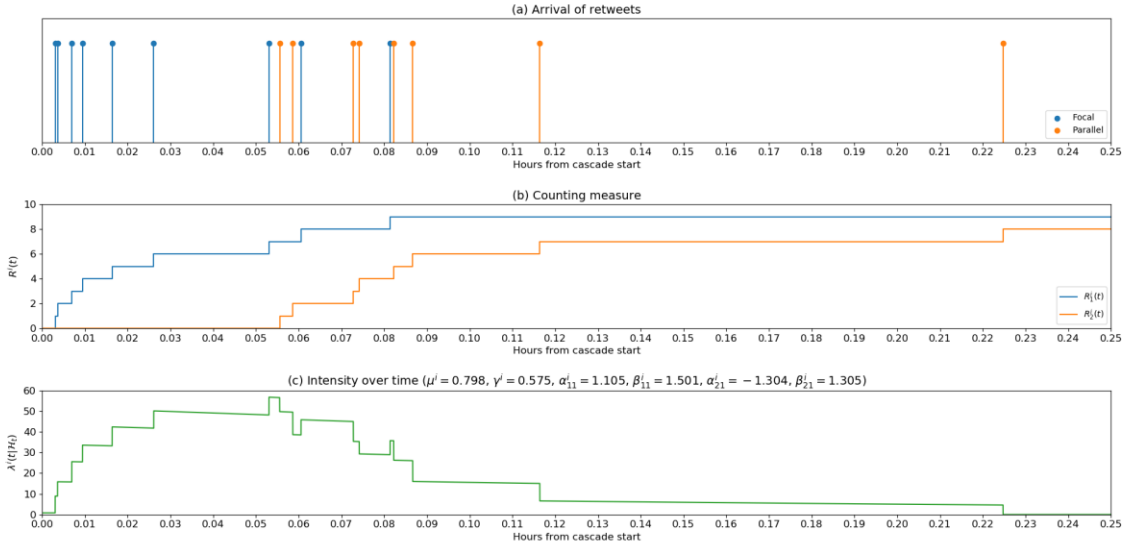
(2012). We also observed that, in absolute terms, the average value of α_{11}^{i} is smaller than that of α_{21}^{i}. Hence, the inhibitive effect of parallel cascades' retweets on the intensity of a focal cascade tends to be stronger than the self-exciting effect of the same cascade's retweets. One implication of this finding is that any impetus gained from a cascade's own diffusion history can be drowned out by the dissemination of parallel cascades.

Table 11 also gives information on the decay rates for how the two point processes influence a cascade's diffusion speed. The average value of β_{11}^{i} is 0.927 and of β_{21}^{i} is 0.571, which suggests that the self-exciting effects of a cascade's retweets wear off faster than the effects of the parallel cascades' retweets. Given this outcome in conjunction with the magnitude of α_{11}^{i} being generally smaller than that of α_{21}^{i}, we can infer that parallel cascades have a more significant and longer-lasting effect on a focal cascade's intensity. These findings underscore the drawbacks of analyzing focal cascades' diffusion speed in isolation. The analyses should integrate the influence of other cascades to obtain estimations that are more realistic. Our results also indicate that the baseline intensity for a cascade is not constant over time since the mean parameter estimate for γ^{i} is equal to 0.356. By allowing the baseline intensity to be time-varying, we were able to model the natural decay of interest in a cascade's content as time progresses.

To visualize how a cascade's intensity changes as a reaction to arrivals of retweets from two distinct point processes, we present the observed point process realizations and intensity for one cascade from the Joplin tornado data in Figure 6. Like Figure 2, Figure 6 contains three panels. Panel (a) shows the arrivals of the cascade's 9 retweets in blue and the arrivals of the parallel cascades' 8 retweets in orange. Next, Panel (b) exhibits the counting measures for both point processes, and Panel (c) graphs the intensity over time given that the estimated parameters for this cascade were $\mu^{i} = 0.798, \gamma^{i} = 0.575, \alpha_{11}^{i} = 1.105, \beta_{11}^{i} = 1.501, \alpha_{21}^{i} = -1.304$, and $\beta_{21}^{i} = 1.305$. As the graph in Panel (c) illustrates, the

initial intensity level is equal to $\mu^i = 0.798$. As retweets of the focal cascade arrive, the intensity experiences a self-exciting effect and increases by a function of $\alpha_{11}^i = 1.105$. Panel (c) depicts the inhibitory effects of parallel cascades' retweets as well. When retweets of parallel cascades occur, the intensity is lowered by a factor of $\alpha_{21}^i = -1.304$. The decay of the self-exciting and inhibitory effects is also shown through the decline in intensity between realizations of the point processes.

Figure 6 – Sample Cascade's Arrivals and Intensity Based on Estimated Parameters



7. Analysis of Competitive vs. Cooperative Effects by Parallel Cascades

The results from the model estimation procedure revealed that parallel cascades can both impede and augment a focal cascade's diffusion rate. To better understand what determines the type of effect that parallel cascades' retweets impose on the diffusion of a focal cascade, we specified a linear regression model with the estimated values of α_{21}^i as the dependent variable. The predictors in this regression correspond to the conditions that we identified in Section 2 as determining whether parallel cascades compete or cooperate with a focal cascade's diffusion.

The first independent variable of interest is the producer's size. As argued in Section 2, we expect cascades contributed by large producers to be less susceptible to competitive effects by parallel cascades. Hence, we anticipate that α_{21}^{i} is positively associated with the size of cascade's producer. We operationalized a producer's size ($size_i$) as the producer's follower count. The second determinant is the number of parallel cascades tied to a focal cascade. A higher volume of parallel cascades indicates that there are more cascades discussing the same topic as the focal cascade. We measured the volume of parallel cascades ($parallel_i$) as the number of individual cascades with realizations in the point process for retweets of parallel cascades. Based on Haas et al. (2015)'s findings, we conjectured that a curvilinear relationship exists between α_{21}^{i} and the count of parallel cascades. Accordingly, we tested the curvilinear effect of parallel cascade count by including in the regression the linear and the quadratic term for this variable.

We also analyzed whether a first-mover advantage exists for focal cascades. To that end, we used a binary variable ($firstmover_i$) that is set to 1 if the focal cascade was the first cascade in its topic and 0 otherwise. In addition, we controlled for when in relation to the disaster the cascade was launched since the timing of content release has been shown to affect cascades' diffusion speed during humanitarian events (Yoo et al. 2016). We measured this variable ($time_i$) as the number of hours between the time that the cascade was initiated and the time that the disaster materialized. Table 12 provides the descriptive statistics for the determinants in our regression.

Table 12 – Descriptive Statistics for Determinants of α_{21}^{i}

	Mean	**Median**	**Std. Dev.**	**Min.**	**Max.**
$size_i$	89,011	2,871.5	503,087.9	0	16,172,110
$parallel_i$	6.874	3	11.701	1	125
$firstmover_i$	0.351	-	-	0	1
$time_i$	34.388	18.040	45.165	0.056	264.582

26,838 observations

We estimated the coefficients of the determinants of α_{21}^{i} using the Ordinary Least Squares (OLS) method. To address nonlinearity, we logged the producer's follower counts. We also mean-centered $parallel_i$ prior to creating the quadratic term to reduce multicollinearity. Finally, we included fixed effects (ξ_1, ξ_2, ξ_3) to capture the non-time-varying unobserved heterogeneity of each disaster. The regression equation that we estimated is given in Equation 11:

$$\alpha_{21}^{i} = \delta_0 + \delta_1 \log size_i + \delta_2 parallel_i + \delta_3 parallel_i^2 + \delta_4 firstmover_i + \delta_5 time_i + \delta_6 \xi_1 + \delta_7 \xi_2 + \delta_8 \xi_3 + \varepsilon_i \quad (11)$$

The results of the OLS regression are listed in Table 13.

Table 13 – OLS Regression Results

	Coeff. (Std. Error)
δ_0 (intercept)	-9.31E-02 (1.02E-02)***
δ_1 ($\log size_i$)	6.18E-03 (8.23E-04)***
δ_2 ($parallel_i$)	1.00E-04 (1.14E-05)***
δ_3 ($parallel_i^2$)	-1.41E-08 (2.14E-09)***
δ_4 ($firstmover_i$)	-1.13E-01 (4.17E-03)***
δ_5 ($time_i$)	-6.44E-04 (4.63E-05)***
Observations	26,838
Adj. R-squared	0.043

*** $p<0.01$

Note: Fixed effects for each disaster are not reported

The coefficient for $size_i$ is positive and significant, which confirms that the effect of parallel cascades' retweets on a cascade's intensity is positively related to the size of the cascade's producer. Extant research has shown that the diffusion of social media content is augmented for producers with more followers (e.g., Hong et al. 2011, Suh et al. 2010, Yoo et al. 2016). Our finding contributes to previous work by demonstrating that larger producers are able to take advantage of parallel cascades and experience faster diffusion from parallel cascades' retweets. The results in Table 13 also lend support to a curvilinear relationship between α_{21}^{i} and the count of parallel cascades in the direction that we

expected. Specifically, the linear term for $parallel_i$ is positive and significant while the quadratic term is negative and significant. This means that HOs should not always view cascades spreading similar content as competitors but realize that participation in popular topics may enhance the diffusion of their content. Moreover, our results indicate that a cascade's diffusion speed is diminished when the cascade is the first among those in its topic to publish the topic's content. An implication of this finding is that HOs do not have to be pressured to be the first to broadcast a piece of information but can rely on parallel cascades' diffusion to improve their own cascades' propagation rate. Lastly, we found that the coefficient for $time_i$ is negative and significant. Therefore, cascades that are launched closer to the start of the disaster are less likely to face competitive effects from parallel cascades. This implies that HOs can expect retweets of parallel cascades to enhance the diffusion of their content in the immediate aftermath of a disaster when rapid information is most critical.

8. Conclusion

In this study, we assessed the diffusion speed for content posted on social media platforms using Twitter data from four disasters. We traced the propagation of content from its origin as a tweet through it being shared by other users in the form of retweets. We created cascades from this data, and each cascade was made up of a tweet and its series of retweets. Instead of calculating a cascade's rate of diffusion solely as a function of attributes of itself and of the users involved, we broadened our analysis to include the spread of parallel cascades. This allowed us to account for interactions among cascades that carry similar content, and we tested whether cascades interacted according to a competitive or cooperative dynamic. To evaluate the direction of cascades' effects on one another, we formulated a point process model based on the Hawkes model (Hawkes 1971). We extended the Hawkes model first by incorporating another point process that

represented the arrivals of retweets for parallel cascades. Secondly, we allowed the effect of the parallel cascades' retweets on a focal cascade's intensity to hold positive and negative values. This modification required us to implement a nonlinear version of the Hawkes model, which is not commonly performed due to difficulties in estimating such models.

The parameter estimates from our point process model reveal that a focal cascade's own retweets heighten the cascade's intensity, or diffusion speed. Our results also indicate that the influence of parallel cascades' retweets is negative, or competitive, on average and that the magnitude of this effect supersedes that of the focal cascade's retweets. However, we found evidence of a cooperative dynamic as well since some cascades' diffusion rate benefited from the concurrent dissemination of parallel cascades. Consequently, we conducted an additional analysis to identify what factors drive whether the effect of parallel cascades' retweets is positive or negative for a focal cascade's diffusion speed. The results of this analysis demonstrate that a focal cascade launched by a producer with more followers is more likely to experience cooperative effects, which highlights the value of producers having large follower bases for reasons beyond access to a greater audience. We also showed that the impact of parallel cascades' retweets on a focal cascade's intensity has a curvilinear association with the number of parallel cascades and that the first-mover advantage appears to be absent.

A major implication from our research is that parallel cascades typically exert a negative impact on the diffusion speed of a focal cascade. This suggests that HOs may want to spend time on curating their content to improve the novelty of their information, especially during non-emergency periods when time is less constrained. Additionally, this study provides guidelines for coordination among HOs with regards to information resources. Since our results show that a positive interaction is possible among cascades

broadcasting similar content, HOs can coordinate the publishing of their social media content to try to avoid any slowdown of their diffusion from parallel cascades produced by other HOs. For instance, we observed that a focal cascade's diffusion speed increases as the number of parallel cascades rises but only up to a certain point. Eventually, the topic becomes too crowded and the effects of parallel cascades' retweets on a focal cascade's intensity become negative. One implication, therefore, is for HOs to work together to release a limited amount of content belonging to the same topic in order to legitimize the information and take advantage of cooperative effects by parallel cascades.

While our study makes significant contributions to the literature and generates managerial implications for HOs, it is not without limitations. The sampled disasters represent only sudden-onset disasters, but there are many disasters that are not classified as sudden-onset but develop over time (e.g., hurricanes, floods). Future research may consider expanding the type of crises in our sample to include slow-onset disasters as well. Diffusion patterns may differ across such events since information is produced during both preparation and response stages. Another limitation of this work is that we lack data on users' exposure to the content, which could impact a cascade's intensity. When a cascade is retweeted, the content is forwarded to the retweeters' followers, and some of these followers may have already been exposed to the same content through another retweeter. We anticipate that a user's likelihood of retweeting is affected by the number of times they receive the same information. While our model accounts for the retweeters' follower counts, it does not measure how many times the followers have previously been exposed to the cascade's content. We call on future research to evaluate if our results hold after including a measure of the number of times retweeters' followers have been sent the same information.

Finally, our point process model can be extended and applied for predictive purposes. Like the SEISMIC model (Zhao et al. 2015), the point process model in this study may be leveraged to predict the number of retweets that a cascade will earn in its lifetime. This would be valuable for HOs that not only need to spread information quickly but also to as many users as possible. Another predictive element for future research to assess is the launch of parallel cascades. Certain variables, especially content attributes, may determine the number of other cascades that will transmit similar content. For HOs, it would be helpful to be able to anticipate the arrival of parallel cascades to better gauge if the effect of parallel cascades' diffusion will positively or negatively influence their cascades' dissemination rate.

CHAPTER 3

Expanding the Reach of Humanitarian Organizations on Social Media Platforms

Abstract

On social media platforms, all content published by a user is instantly transmitted to its set of followers. Therefore, a user's direct audience is composed of its followers. In order to reach a larger audience in real-time, humanitarian organizations that are active social media users aim to increase their follower counts. The purpose of our paper is to analyze what mechanisms motivate the growth of humanitarian organizations' social media networks during times of normalcy and emergency. We collected a unique data set from Twitter that includes dynamic network information for 47 organizations that were directly involved with relief efforts for the 2016 Ecuador earthquake. The network data encompassed over 170 million links. Our analyses indicate that the organizations in our sample collectively increased their follower counts by 275,359 followers and that a significant driver of these new connections was the exposure gained from existing network members sharing the organizations' content. In addition, we specified a structural model to investigate what determines a user's choice to become a new follower after learning about an organization from a shared piece of content. We found, for instance, that the type of content that humanitarian organizations broadcast and the frequency that this content is produced impact users' probability of starting to follow the organizations.

1. Introduction

To distribute all required goods and services to their stakeholders, humanitarian organizations (HOs) must manage a complex flow of physical resources, including food, water, and medications. Another vital flow involves information resources, especially since this flow facilitates the sourcing and delivery of physical resources to stakeholders. In fact, the effective management of information is one of the most critical factors in determining the success of humanitarian operations (Long and Wood 1995). Traditionally, the management of information flows has been a major challenge for HOs. As noted by Holguín-Veras et al. (2012), the operational environment during a disaster is volatile due to factors like a turbulent physical landscape, population migration, and disrupted economic and political states. This means that decision parameters related to the operational environment are changing constantly, and what may have been relevant or accurate information yesterday is no longer so today. Because information is highly perishable in a humanitarian context (Meier 2015), HOs require a robust information network that can quickly diffuse information among a wide array of stakeholders.

Such a network is difficult for HOs to establish, particularly when dealing with the effects of a disaster. As a result, HOs have sought to leverage social media platforms on the internet. These platforms provide a space for HOs and other users to generate, discuss, and share a wide variety of user-generated content. Moreover, because these platforms pose low entry barriers to users, they are easily accessible for not only HOs but also for their stakeholders. In addition, social media platforms have been observed to be highly reliable as a means of communication during times of crisis because users can access these platforms through internet or cellular network infrastructure. In the case that such infrastructure goes down, responders often prioritize restoring these services to facilitate communication. One of the most popular social media platforms is Twitter. As of 2018,

Twitter has 330 million monthly active users that send 500 million messages per day. Twitter users are global in that almost 80% of user accounts are from outside of the United States[8]. Numerous humanitarian organizations, such as the Red Cross and UNICEF, have accounts that represent their organizations and are active Twitter users. Moreover, the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) announced Twitter as the social media platform of choice for humanitarian organizations (Moore and Verity 2014).

A key objective when a HO uploads content to Twitter is to broadcast this information as quickly and to as many other users as possible. According to Stieglitz and Dang-Xuan (2013), Suh et al. (2010), and Yoo et al. (2016), the size of a user's direct audience, also known as its follower base, is important to fulfill this goal. When users publish content, those with larger follower bases can instantly transmit that content to a broader set of users, and consequently, the initial wave of content dissemination will be greater for users with more followers. Moreover, as the number of followers increases, it opens more avenues for content to be further distributed and shared across the platform.

Given the importance that the size of the follower base has on the diffusion of information, this paper investigates the mechanisms that drive the growth of these audiences for HOs. One way for users to become followers of a HO is for them to find the HO and establish connections out of their own initiative. To find the HO, users may look specifically for the organization on Twitter or receive Twitter recommendations to start following the organization. The first method is highly dependent on the reputation or prominence of the HO while the second requires financial resources. Most HOs, however,

[8] https://blog.hootsuite.com/twitter-statistics/

do not have the type of recognition or the financial capital required to raise their visibility among users and draw new followers via this mechanism.

A second mechanism to expand a HO's follower base is derived from users learning about the HO when they receive content contributed by the HO from other users they follow on the Twitter platform. If users find value in the type of content contributed by the HO, they may opt to form a follower relationship directly with the HO. Naturally, users will have a greater chance of learning about the HO from content transmitted through their Twitter networks as the frequency with which the HO contributes content to Twitter increases. Therefore, to the extent that the HO actively contributes content to the platform, users may have more opportunities to come across this information and establish follower links directly with the HO. Nevertheless, an HO that floods Twitter with content may not necessarily maximize its chances of gaining new followers. Increases in the frequency of the HO's contributions may have marginally decreasing returns on the likelihood of new users' following the HO due to the additional information processing costs that users anticipate incurring from receiving greater amounts of content from the HO. Finally, in addition to the HO's decisions on the type of content and the frequency of its contributions, the likelihood of users deciding to follow a HO will depend on the frictions imposed on the diffusion of content from the HO by the layers of intermediaries in the network separating the HO and the users. To the extent that these frictions will generate delays or a breakdown in the diffusion of information, the likelihood that the users will choose to bypass the network and follow the HO directly may increase.

We study these mechanisms using data collected directly from Twitter during two periods: immediately before and immediately after the start of a sudden-onset disaster event. Specifically, we utilized data related to the 2016 Ecuador earthquake. The evaluation during these two periods is important because it allows us to assess the

mechanisms that drive the growth of HOs' follower base sizes during times of normalcy and emergency. Users may experience different utilities for these mechanisms depending on whether or not a disaster has occurred. Moreover, by considering both of these periods for our study, we can account for variations in information production requirements for the HOs. When they are not facing a crisis event, HOs can strategically plan the type of content and the timing in their release of information. However, an emergency will compel HOs to become more reactive in deciding the type and frequency of information releases in order to maintain the public informed as the crisis unfolds.

Our results show that the diffusion of Twitter posts contributed by HOs not only serves to distribute information but also as a powerful and effective driver of new follower relationships, particularly during times of crisis. To explain why users would form these relationships, we formulated and estimated a two-stage structural model comprising the users' consumption of content contributed by a HO in the first stage and their decisions to follow the HO in the second stage. According to the results from our model, users that receive content contributed by a HO through their follower relationships with other users have a higher probability of forming a follower relationship with the HO after the onset of a crisis event than before the event. These users are also more likely to form follower relationships with the HO when they receive actionable information (e.g., content that contains instructions for evacuation and directions to shelters) from the HO. Moreover, we found that users are more prone to follow the HO when they anticipate they will obtain more information more completely and rapidly from doing so. Specifically, users have a greater probability of following the HO if doing so will result in an increase in the expected amount of information received from the HO and will lead to a decrease in the delay of information receipt relative to what they have experienced through their follower relationships with other users. These effects vary depending on whether the HO finds itself

attending to a crisis event or not. After a disaster, we observed that users demonstrate a greater sensitivity with regards to reducing the delay of receiving information and increasing the amount of content received when making their decisions about whether to follow HOs.

We organize the rest of the paper as follows. We expand on the related literature for our research in Section 2. We describe in detail the mechanisms for the formation of follower links in Section 3. In Section 4, we formulate the structural model for understanding when users form follower relationships with HOs after receiving content contributed by the HO through their follower relationships with other users. We describe our data in Section 5 and present our results in Sections 6 through 8. We discuss our results and conclude in Section 9.

2. Literature Review

Our paper furthers the current understanding that exists in the literature about how HOs can improve their use of social media platforms to diffuse information to their stakeholders. In so doing, our paper contributes to two different areas of literature.

2.1. Information Management in Humanitarian Operations

The management of information for an HO entails coordinating information flows within itself, with other organizations, and with individuals. To monitor information within, HOs have implemented databases for tracking and tracing the movement of inventory (Pettit and Beresford 2009) as well as for monitoring donors (Ryzhov et al. 2015). Those that have adopted such systems have improved the visibility and accessibility of data for their staff. HOs also record data about supply distribution to update inventory levels and to forecast future demand (van der Laan et al. 2016). The implementation of information technology (IT) tools, like scanners, has enhanced HOs' ability to collect and maintain data on supply allocation as well as demand (Ergun et al. 2014). Additionally,

HOs rely on their local teams for information when preparing for and responding to a crisis since these staff are already on the ground and thereby have a better understanding of demand and the environment (Tomasini and Van Wassenhove 2009).

HOs also exchange information with entities outside of their own organizational boundaries. Information sharing has been observed to be a challenge in the humanitarian setting due to factors like competition for resources and the convergence of many organizations (Balcik et al. 2010). However, HOs have started to establish interorganizational channels of information to combat this issue. The United Nations (UN) initiated the cluster approach, which groups HOs into clusters based on their area of expertise, and organizations within a cluster are encouraged to communicate. Altay and Pal (2014) found that cluster leaders play a pivotal role and should act as information hubs to achieve this goal. The UN has designed several platforms to promote information sharing and transparency. Its Joint Logistics Centre designed an online platform where logistics groups can exchange and view information about issues like weather and warehousing availability (Tomasini and Van Wassenhove 2005). The UN Humanitarian Response Depots, which house HOs' prepositioned inventory, also publish online the owners and quantities of inventory at each warehouse (Acimovic and Goentzel 2016). Finally, HOs collaborate with government and private groups to acquire information like census data, weather forecasts, and satellite images of areas affected by a disaster (Sodhi and Tang 2014).

The final set of information flows that HOs must manage are with individuals. The internet and mobile technology have made communication between HOs and individuals radically more accessible and pervasive as well as opened new opportunities for HOs to improve their responsiveness (Swaminathan 2018). One example of how HOs exchange information with individuals is through collaboration-based crowdsourcing, which occurs

when self-selected people from the crowd from work jointly to solve a problem (Afuah and Tucci 2012). In the immediate aftermath of the 2010 Haiti earthquake, a crowdsourced crisis map was set up and populated by incident reports from the crowd. This map was used by HOs to gain awareness about the operational environment and plan response efforts (Gao et al. 2011). HOs also utilize social media platforms to engage with individual stakeholders like beneficiaries and donors in real-time and at no cost (Yoo et al. 2016). Social media users upload relevant content, such as reports on injuries and damage, and many of the active users during an emergency are located within the disaster zone (Starbird et al. 2010, Vieweg et al. 2010). Consequently, HOs collect social media data as a supplemental of information on demand and the general situation. HOs not only leverage social media platforms to gather data but also to broadcast information to individuals. This may include messages expressing social support, instructions on how to find shelters, and donor appeals to drive donations (Eftekhar et al. 2017, Pedraza Martinez and Yan 2016). Our paper adds to this area of literature by being the first to study the growth of HOs' social media networks over time and during periods of normalcy and crisis.

2.2. Social Media Platforms and Operations Management

Additionally, our paper belongs to the growing literature on the applications of social media platforms in operations management. Due to the volume of users and the content they generate, social media platforms offer a trove of valuable data on consumers' preferences and behavior. As a result, the integration of social media data has been shown to raise the accuracy of sales forecasts (Cui et al. 2017). Social media platforms have also been recognized to be important in managing a firm's services. For example, firms can utilize these platforms to address instances of service failure described in online reviews by unsatisfied customers (Gu and Ye 2014). Firms can also use social media data on customers, such as their number of friends, level of engagement, and economic value, to

improve customer targeting (Allon and Zhang 2018, Momot et al. 2017). Furthermore, the adoption of social media platforms enhances firms' operational efficiency and innovativeness. This is because these platforms enable employees to easily share knowledge and interact with one another and with customers (Lam et al. 2016).

Users interact and exchange information with one another on social media platforms, which allows some users to influence others. Users conforming to other users' opinions and behavior has been observed in the case of online movie reviews (Lee et al. 2015), subscribing to a service (Bapna and Umyarov 2015), and making purchases (Lobel et al. 2016). On social media platforms, one measurement of influence is a user's number of followers since this translates into the potential pool of other users that may be influenced by the user's content. With higher counts of followers, users can broadcast content more efficiently and instantly reach a larger audience than those with smaller sets of connections (Goel et al. 2016). This jumpstarts content diffusion, and therefore, content produced by users with more followers tends to experience faster and greater contagion (Susarla et al. 2011, Yoo et al. 2016). In order to expand their follower base sizes, users have to invest time and effort to develop strategies to attract new followers (Iyer and Katona 2015). Caro and Martinez-de-Albeniz (2018) found, for instance, that the timing of content production affects follower base growth. In our paper, we consider the frequency of information supplied by HOs and identify other factors that drive the expansion of HOs' followers. Therefore, we contribute to this stream of literature by investigating not only the strategies that HOs can adopt to increase their follower counts but also how these strategies evolve when HOs are operating under normal versus emergency conditions.

3. Mechanisms for the Formation of Follower Links

On Twitter, a user can post short messages called "tweets" that may contain text as well as URLs and multimedia content. Through Twitter's sharing function (known as "retweeting"), users can forward another user's original tweet to their own network. Retweets sent by these users (commonly referred to as "retweeters") preserve the original tweet's content and timestamp and also assign credit for the content to the original tweet's author. Like other social media platforms, Twitter operates on an underlying user network. A user can have a list of other users that are its "followers" or that it is "following", which we refer to as "friends". Twitter feeds display messages (both tweets and retweets) by a user's friends in reverse chronological order and without any delay from when the content is posted onto the Twitter platform. This feature is key because it gives Twitter users the ability to instantly diffuse this information to their followers. Moreover, because followers that consume this information can retweet it to their own followers, they can quickly diffuse it to a broader set of users potentially beyond those that follow the author, or the "supplier", of the original tweet. The dissemination of a supplier's tweet can be extended even further, such as by followers of a retweeter, because Twitter allows users that are not following the supplier to still retweet its content.

As a byproduct of this diffusion, recipients of a retweet can learn who the content supplier is, which may prompt these users to form new connections and start following the supplier. We label this mechanism of generating connections as *internal* because it produces new follower relationships out of the diffusion of information within the underlying social networks. Follower relationships can also originate *externally* because of stimuli outside the network of retweeters. For instance, users may simply search and start following other users because of their reputation. This is most likely to occur for celebrities or for users that represent large entities, such as news organizations. Moreover,

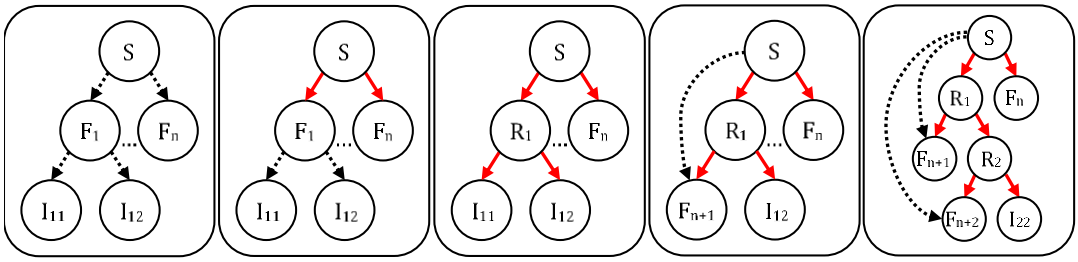
Twitter, like other social media platforms, provides recommendations on who to follow, and this may motivate some users to start following the suggested users. Twitter may provide these recommendations as part of campaigns in which users pay a fee for each new follower referred to them by Twitter.

The vast majority of HOs do not have the resources to generate the different types of external stimuli that can lead to the formation of new follower relationships. An internal mechanism like the one we described above constitutes a more cost-efficient option to foster the development of these links for these organizations. The internal mechanism is founded on the principle of triadic closure, which states that two individuals are more likely to be connected if they are both associated with a third individual (Granovetter 1973). For instance, if A and B do not know one another, but each of them is linked with C, the property of triadic closure implies that a connection between A and B is likely to transpire due to the fact that both A and B know C. On Twitter, this mechanism would essentially involve a user (C) issuing a retweet of a supplier (B) to its followers (including A). Since retweeting credits the supplier (in this case, B) of the message being shared, the follower (A) that receives and consumes the retweet will learn about the identity of the supplier and may choose to follow the supplier directly.

More formally, we label the author of a tweet as the information supplier, S. Assume that S is followed by n users that make up its follower base (i.e., $\boldsymbol{F} = [F_1, F_2,\ldots,F_n]$). Once S supplies a tweet, its set of followers, $\boldsymbol{F}$, immediately receives the content. The users in $\boldsymbol{F}$ can distribute information by retweeting S's content to their own followers. These retweeters belong to the set $\boldsymbol{R} = [R_1, R_2,\ldots,R_m]$, where m is the total number of retweeters. Continuing the example, assume that F_1 becomes R_1 by retweeting a tweet of S. This generates an "exposure" of S's tweet in each of the R_1 followers' feeds. Assuming R_1's followers consume these exposures, they will face the decision of whether or not to follow

S. We refer to R_1's followers as "candidates" (i.e., $\boldsymbol{I_{R_1}} = [I_{11}, I_{12},..., I_{1k}]$, where k is the number of users following R_1) since they now have the opportunity to start following S. Therefore, candidate I_{11}'s decision to follow S and join $\boldsymbol{F}$ as F_{n+1} marks the establishment of a link based on the internal mechanism. Figure 7 illustrates this example. Moreover, the internal mechanism can be present in a generalized version of triadic closure where users with greater distance between them, or more degrees of separation, become connected (Kossinets and Watts 2006). This is due to Twitter's property that allows users to retweet a supplier's content without following the supplier. As a result, candidates (that may or may not have converted to new followers) in $\boldsymbol{I_{R_1}}$ can join the set of retweeters in $\boldsymbol{R}$. To exemplify this phenomenon, suppose that I_{12} does not choose to connect with S but does convert to R_2 by also sharing the retweet initially distributed by R_1. The internal mechanism is complete when one of R_2's followers (labeled as I_{21} to be consistent) consumes the exposure from R_2 and subsequently starts following S, thereby joining $\boldsymbol{F}$ as F_{n+2}. A candidate can thus connect with a supplier on account of a retweet issued by a user that is not necessarily linked with the supplier. This scenario will become increasingly common as the diffusion of a tweet broadens. Please refer to panel (e) in Figure 7 for an illustration.

Figure 7 – The Internal Mechanism



(a) Initial state (dashed arrows denote follower relationships).

(b) S posts content, which is sent to its followers (as denoted by the red arrows).

(c) Follower F_1 becomes R_1 by retweeting S's content. Candidates I_{11} and I_{12} receive exposures of S.

(d) Candidate I_{11} starts following S and becomes F_{n+1}.

(e) Candidate I_{12} does not follow S but retweets to become R_2. I_{21} and I_{22} (R_2's followers) receive exposures. Candidate I_{21} then follows S and becomes F_{n+2}.

4. Structural Model

One of our objectives is to understand how the internal mechanism described above drives link formation. In particular, we want to know how HOs can benefit from this type of mechanism in converting candidates to new followers. For ease of exposition, we refer in the remainder of this paper to these new follower relationships as "internal links". Conversely, we refer to links formed from an external mechanism as "external links".

In broad terms, we aim to trace the distribution and consumption of tweets through the Twitter network leading to candidates' decisions to form internal links with the suppliers of those messages. To that end, we formulate a two-stage structural model to specify attributes of social media platforms like Twitter and candidates' decision-making processes to form these links. This modeling approach is consistent with those employed by Huang et al. (2015), Shi et al. (2014), and Tang et al. (2012) in the literature on social media platforms. The first stage assesses a candidate's consumption of an exposure in its feed of a supplier's tweet, and the second stage models the candidate's decision to follow the supplier. We elaborate on the two stages of our structural model below.

4.1. Stage 1: Consumption

The first stage models whether a candidate, *i*, consumes, or reads, an exposure of tweet *t* contributed by supplier *s*. Candidate *i*'s Twitter feed will not only contain the exposure of *t* by *s* but also other content posted by *i*'s friends in reverse chronological order. Whether or not *i* actually consumes the exposure of *t* depends on several factors, such as how frequently *i* logs in, *i*'s attention span, and how much information *i* receives from its friends on Twitter. Since *i* is unlikely to constantly monitor its Twitter feed nor read all the activity published between its current and last login, some exposures may go unseen. In such cases, it is not possible for *i* to legitimately learn about *s* from the exposure

of *t* and form an internal link with *s*. This means that *i* must consume the exposure of *t* in the first stage to advance to the second stage of our structural model.

To evaluate whether a candidate consumes an exposure, we apply a modified version of Shi et al. (2014)'s consumption model, which was originally designed to test whether content is consumed by potential retweeters. Like Shi et al. (2014), we assume that, upon login, candidates consume a limited number of their friends' activity starting at the top of their Twitter feeds and that candidates do not favor consumption of certain friends' content over others'. The amount consumed depends on *i*'s attention span (i.e., α_i), which is directly unobserved by the researcher. Each candidate will read its friends' tweets and retweets that are within the index [1, α_i] on its Twitter feed. As long as the place of *t*'s exposure lies within this index, the candidate will consume the exposure.

The condition for consumption by *i* of the exposure of *t* authored by *s* is:

$$\frac{1}{b_{sti}^{\beta_1}} > L_{sti}, \tag{1}$$

where b_{sti} represents *i*'s number of friends, L_{sti} stands for *i*'s unobserved inverse login frequency, and β_1 is the effect of b_{sti} on consumption. We assume that b_{sti} and L_{sti} are uncorrelated (Shi et al. 2014) and that b_{sti} is linearly associated with the volume of activity in Twitter feeds (Gomez Rodriguez et al. 2014). The left side of the inequality signifies the scaled proportion of activity in *i*'s Twitter feed that is the exposure of *t*. Candidates that login more frequently will have a lower value of L_{sti}, making it more likely that this condition will be satisfied. At the same time, candidates will have a higher amount of activity in their feeds when they have more friends and, thus, will be less likely to consume an exposure. Following Shi et al. (2014), we assume that the unobserved α_i is absorbed into L_{sti} and that L_{sti} is log-normally distributed with mean L and variance σ_L^2. Based on this, Equation 1 can be rewritten as:

$$-\frac{L}{\sigma_L^2} - \frac{\beta_1}{\sigma_L^2}\log b_{sti} > \frac{\log L_{sti} - L}{\sigma_L^2},$$

where log(·) means taking the natural logarithm in Equation 1 and throughout the rest of the paper. The following equation is the probability that *i* consumes the exposure of *t*:

$$P1 = P\left(-\frac{L}{\sigma_L^2} - \frac{\beta_1}{\sigma_L^2}\log b_{sti} > \frac{\log L_{sti} - L}{\sigma_L^2}\right) \quad (2)$$

4.2. Stage 2: Follow Decision

Given *i*'s consumption of *t*'s exposure in the first stage, *i* must decide whether or not to follow *s*. We model this decision in the second stage of our model as a function of the utility and the cost that *i* will incur after following *s*.

The utility that *i* derives from following *s* after the consumption of *t*'s exposure (U_{sti}) depends on the value attached to the type of content in *t* consumed by *i* (a_{sti}). It is also contingent on the value *i* can earn by potentially becoming a first-hand distributor of *s*'s content to its own network of followers. By gaining immediate access to *s* and retweeting *s*'s content, *i* may disseminate information that was not previously available to its followers and subsequently improve its standing as an information distributor (Boyd et al. 2010). The utility for *i* of becoming a retweeter of *s* after following *s* is a function of *i*'s audience size (p_{sti}), or its number of followers, and how active *i* is on Twitter, particularly in its commitment to sharing content with its followers (q_{sti}).

The topology of the network separating *i* from *s* also influences the utility that *i* will obtain from following *s*. For *i*, its utility will depend on the size of the follower base for *s* (r_{sti}). This is because suppliers with larger counts of followers tend to be viewed as more credible and as producers of higher-quality information (Ringel Morris et al. 2012). The calculation of utility also includes the distance, or the degree of separation, between *i* and *s* when *i* consumes *t*'s exposure (g_{sti}). Since information typically does not propagate far on Twitter (Goel et al. 2016), *i* will be able to access content that is more innovative relative to what is available through its local network as the distance between *i* and *s* grows (Aral

and Van Alstyne 2011). As such, we expect that *i* will find more value in following *s* when g_{sti} is large.

Finally, the utility that *i* will obtain from following *s* depends on the performance of the retweeters that distribute *s*'s content to *i*. One aspect of performance is the amount of *s*'s content that *i* receives through its network. When the volume of *s*'s information that *i* receives via retweets falls short of the total volume of information *s* contributes on Twitter, *i* will find utility in following *s*. That is, *i* will gain utility from following *s* directly when the entire quantity of content that *s* contributes on Twitter does not diffuse completely down to *i*. This may occur, for instance, because users in the network between *i* and *s*, including *s*'s followers and *i*'s friends, do not retweet a lot of the content posted by *s* or retweet very infrequently. Let f_{sti} be the number of tweets published by *s* over a fixed amount of time prior to *i*'s consumption of the exposure of *t*, and let z_{sti} be the count of these tweets that *i* ultimately receives through retweets from friends in the same amount of time. Thus, the expected increase in coverage of *s*'s activity for *i* after following *s* will equal $f_{sti} - z_{sti}$, and as this difference increases, *i* will obtain greater utility from following *s*. Another facet of the performance of the retweeter network is the speed at which information is circulated. Consequently, *i*'s utility will be contingent on the lag between the time *s* posts content on Twitter and the time retweets of this content reach *i* (w_{sti}). The longer this delay, the greater the utility that *i* will obtain from following *s*.

Equation (3) formally presents the utility function and includes coefficients $(\gamma_1, \ldots, \gamma_9)$ that measure the change in utility from their associated variables. This function assumes a Cobb-Douglas functional form (Arrow et al. 2011). Please note that the function allows the utility from following *s* as w_{sti} expands to increase exponentially in order to account for the exponential decay in the value of information over time, as has been observed for information on social media (e.g., Wu and Huberman 2007) and assumed for

other perishable resources (e.g., Blackburn and Scudder, 2009). In addition, we interact w_{sti} with a binary indicator (d_{sti}) to take into consideration differences in utility as w_{sti} increases depending on whether *t*'s exposure in *i*'s feed occurs during a crisis event (d_{sti}= 1) or not (d_{sti}= 0). This is important to evaluate because the extreme uncertainty and volatility in emergencies causes information to expire more quickly (Meier 2015), which may impact how candidates assign value to the speed at which information is distributed to them.

$$U_{sti} = a_{sti}^{\gamma_1} * p_{sti}^{\gamma_2} * q_{sti}^{\gamma_3} * r_{sti}^{\gamma_4} * g_{sti}^{\gamma_5} * (f_{sti} - z_{sti})^{\gamma_6} * e^{(\gamma_7 w_{sti} + \gamma_8 d_{sti} + \gamma_9 w_{sti} * d_{sti})}, \quad (3)$$

The cost for *i* of following *s* is primarily driven by the information processing cost of the expected increase in contents that *i* will receive from *s* after becoming a follower. Equation (4) presents the cost function. In this function, we assume that information processing cost is a strictly convex function of the quantity of information (Anderson and de Palma 2009). Moreover, because we know that during times of crisis people actively seek out information to cope with stress and to improve their responses (Sutton et al. 2008), we conjecture that the additional effort required to process more information may be lower during times of crisis. We test this by moderating the information processing cost component with d_{sti} in our cost function. In addition, we let ε_{sti} represent the unobserved cost component and be log-normally distributed with mean ε and variance σ_{ε}^2.

$$C_{sti} = \left(e^{f_{sti}} - e^{z_{sti}}\right)^{\gamma_{10}} * \left(e^{f_{sti}} - e^{z_{sti}}\right)^{\gamma_{11} * d_{sti}} * \varepsilon_{sti}, \quad (4)$$

where γ_{10} marks the change in cost from the anticipated increase in the quantity of information received by a candidate and γ_{11} represents the change in cost from the same variable if the exposure occurred after a disaster. The cost function, like the utility function, adopts the Cobb-Douglas functional form.

For *i* to follow *s*, utility must be greater than cost. That is,

$$U_{sti} > C_{sti}. \quad (5)$$

We can rewrite Equation 5 as follows:

$$-\frac{\varepsilon}{\sigma_\varepsilon^2}+\frac{\gamma_1}{\sigma_\varepsilon^2}\log a_{sti}+\frac{\gamma_2}{\sigma_\varepsilon^2}\log p_{sti}+\frac{\gamma_3}{\sigma_\varepsilon^2}\log q_{sti}+\frac{\gamma_4}{\sigma_\varepsilon^2}\log r_{sti}+\frac{\gamma_5}{\sigma_\varepsilon^2}g_{sti}+\frac{\gamma_6}{\sigma_\varepsilon^2}\log(f_{sti}-z_{sti})$$
$$+\frac{\gamma_7}{\sigma_\varepsilon^2}w_{sti}+\frac{\gamma_8}{\sigma_\varepsilon^2}d_{sti}+\frac{\gamma_9}{\sigma_\varepsilon^2}w_{sti}*d_{sti}-\frac{\gamma_{10}}{\sigma_\varepsilon^2}\log\left(e^{f_{sti}}-e^{z_{sti}}\right)$$
$$-\frac{\gamma_{11}}{\sigma_\varepsilon^2}\log\left(e^{f_{sti}}-e^{z_{sti}}\right)*d_{sti}>\frac{\log\varepsilon_{sti}-\varepsilon}{\sigma_\varepsilon^2}$$

Subsequently, the probability of becoming a new follower conditional on consumption is

$$P2=P\Big(-\frac{\varepsilon}{\sigma_\varepsilon^2}+\frac{\gamma_1}{\sigma_\varepsilon^2}\log a_{sti}+\frac{\gamma_2}{\sigma_\varepsilon^2}\log p_{sti}+\frac{\gamma_3}{\sigma_\varepsilon^2}\log q_{sti}+\frac{\gamma_4}{\sigma_\varepsilon^2}\log r_{sti}+\frac{\gamma_5}{\sigma_\varepsilon^2}g_{sti}+\frac{\gamma_6}{\sigma_\varepsilon^2}\log(f_{sti}-$$
$$z_{sti})+\frac{\gamma_7}{\sigma_\varepsilon^2}w_{sti}+\frac{\gamma_8}{\sigma_\varepsilon^2}d_{sti}+\frac{\gamma_9}{\sigma_\varepsilon^2}w_{sti}*d_{sti}-\frac{\gamma_{10}}{\sigma_\varepsilon^2}\log\left(e^{f_{sti}}-e^{z_{sti}}\right)-\frac{\gamma_{11}}{\sigma_\varepsilon^2}\log\left(e^{f_{sti}}-e^{z_{sti}}\right)*d_{sti}>$$
$$\frac{\log\varepsilon_{sti}-\varepsilon}{\sigma_\varepsilon^2}\Big|-\frac{L}{\sigma_L^2}-\frac{\beta_1}{\sigma_L^2}\log b_{sti}>\frac{\log L_{sti}-L}{\sigma_L^2}\Big) \quad (6)$$

5. Data

For this study, we obtained Twitter data generated one week before and one week after a 7.8 magnitude earthquake that occurred in Ecuador on April 16, 2016. The earthquake devastated Ecuador's coastal provinces, caused over 650 casualties, and injured approximately 16,600 people (Symmes Cobb and Ore 2016). Since the earthquake occurred in Ecuador, the language that was predominantly represented in our data was Spanish. We selected this event as the setting for our research because it represented a sudden, unexpected incident; as of now, earthquakes cannot be reliably predicted. Thus, we were able to cleanly compare effects before versus after the earthquake in our structural model. Another reason for our selection of this crisis for our research is that Ecuador is a small country, which helped guarantee that the national level of attention was focused on the crisis and minimized the possibility that another event happened around the same time, which could have interfered with our analysis. Ecuador is also a country where the internet and cellular network infrastructures are well-developed. These networks also proved to be robust enough to withstand the effects of the earthquake and provided the

support necessary to facilitate the communication of information among the population (CNN Español 2016).

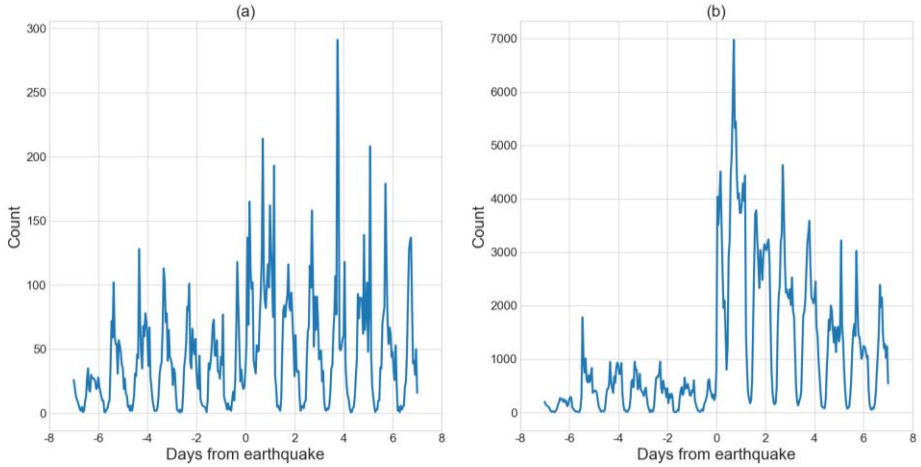
Because our goal is to investigate the growth of follower bases for HOs, we sampled Twitter users that represented Ecuadorean organizations involved with disaster relief. We found the users by locating those in the days after the earthquake that were contributing information under nine commonly used hashtags in tweets related to the crisis. These hashtags included "#TerremotoEcuador", "#EcuadorEarthquake", and "#EcuadorListoySolidario". To control for unobserved effects of content, only those organizations that tweeted exclusively about the earthquake in the week following the disaster were included in our sample. This process resulted in a sample of 55 organizational users. After filtering out users with privacy issues or that had not published Twitter activity both before and after the disaster, our final sample was made up of 47 organizations, or suppliers. These suppliers represented four categories of organizations directly involved with relief efforts: (1) humanitarian; (2) government; (3) medical; and (4) emergency services. Table 14 lists the categories and the suppliers' Twitter handles in each category.

Table 14 – Categorization of Suppliers Listed by Twitter Handles

Emergency Svcs.	Government		Humanitarian	Medical
BOMBEROSGIRECAN	AdmPublicaEc	InclusionEc	ANEPPCE	HGuayaquil
BomberosGYE	AgriculturaEc	IndustriasEc	aldeasosecuador	HVCMCuenca
BomberosQuito	alcaldiagye	MFAEcuador	cruzrojaecuador	IESSHCAM
ECU911Esmeralda	ANT_ECUADOR	MinInteriorEc	cruzrojaguayas	IESSHJCA
Ecu911Macas	CancilleriaEc	MunicipioQuito	CRUZROJAZUAY	
ECU911PVO	CancilleriaEcZ8	ObrasPublicasEc	OPSECU	
ecu911Riobamba	ComunicacionEc	Riesgos_Ec	PNUDEcuador	
ecu911sambo	Ecuador_OEA	Salud_CZ6	worldvisionEC	
PoliciaEcuador	eerssaoficial	Salud_CZ7		
	goberazuay	Salud_Ec		
	GoberdelGuayas	Seguridad_Ec		
	GoberLoja	SENAE_Aduana		
	gobermorona_s	SocialEc		

The data in our study is compiled from multiple sources as we obtained data from Gnip (a Twitter subsidiary) and scraped additional data using Twitter's application programming interface (API). The Gnip data provide information on the tweets published by the sampled suppliers along with all of the retweets of those tweets from the week before and after the earthquake. In total, the 47 suppliers issued 15,399 tweets across the two weeks, which were retweeted 376,732 times in the same amount of time. These retweets were posted by 66,308 retweeters, meaning that each retweeter in our sample contributed 5.68 retweets on average. Nearly 65% of the tweets and 85% of the retweets occurred after the earthquake, and this highlights the surge in Twitter activity in the post-earthquake scenario. In Figure 8, we show the amount of tweet and retweet activity over the two weeks of our study.

Figure 8 – Count of (a) Tweets and (b) Retweets



For the suppliers and retweeters, the Gnip data set also incorporates information from their Twitter profiles, such as the account creation dates as well as the counts of followers, friends, and cumulative number of tweets that they have posted. The profile data are longitudinal since the data were captured for every supplier each time it tweeted or was retweeted and for every retweeter each time it retweeted. The number of followers across all of the suppliers totaled 3.6 million while the count of candidates was 168 million.

Clearly, the amount of candidates dwarfs the number of suppliers' followers. This exemplifies how Twitter, through the retweet function, enables suppliers to expand the reach of their content far beyond their immediate networks and communicate their content to a large audience. We portray the magnitude of the audiences that received content from the suppliers directly as well as through retweets in our sample in Figure 9. Additionally, the median number of followers for each supplier was 13,115 and for each retweeter was 156. Figure 10 illustrates the cumulative distribution functions of the follower counts (logged due to extreme skewness) for the suppliers and retweeters in this study.

Figure 9 – Magnitude of Audiences for Suppliers' Content

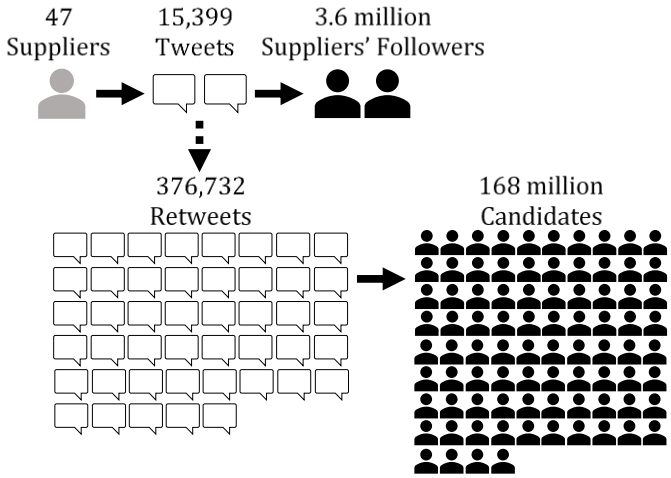
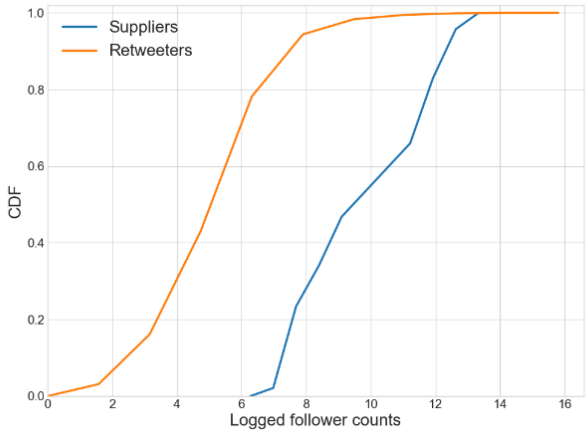


Figure 10 – Cumulative Distribution Functions of Logged Follower Counts for Suppliers and Retweeters



The data we obtained using our second source (the Twitter API) provides more detailed information on the suppliers' followers and the retweeters' followers (that is, the candidates). Specifically, this data include the follower lists for every supplier and retweeter in our sample. Through these lists, we obtained the identities of the suppliers' followers and the candidates. We could not download this data for 411 retweeters due to their profiles being set to private or being deleted, so we dropped these retweeters and their 1,034 retweets. Given that the dropped retweeters represented a minimal fraction of the entire set of retweeters (411/66,308=0.62%) and of retweets (1,034/376,732=0.27%), we do not expect any changes in our results due to their removal from our sample. In addition, we used Twitter's API to scrape the candidates' profile data in order to collect the same statistics available in the Gnip data set for suppliers and retweeters (e.g., account creation dates, follower counts). Except for those with deleted profiles (1% of the total count of candidates), we were able to acquire this data successfully[9].

6. Internal and External Link Analysis

From our data, we were able to identify whether new follower relationships materialized as a result of internal or external mechanisms. The first step was to identify which new followers each supplier gained during the week before and the week after the earthquake. We employed the scraped follower lists for the suppliers for this task. Each list provided a supplier's follower identities in reverse chronological order according to the time they started following the supplier. The exact times that followers started following were not accessible and, to the best of our knowledge, this information is not available to

[9] We gathered candidates' profile statistics one year after the earthquake, and it is to be expected that the statistics evolved during the elapsed time. To test the consistency of candidates' scraped profile data, we compared the profile information obtained from Gnip and from Twitter's API for 1,000 randomly sampled candidates that we had both sets of information for. Because the correlation between both types of measurements was greater than 90%, the measurements in the scraped data constitute a valid proxy for the candidates' profile statistics at the time of the earthquake.

scrape or to purchase, even from Twitter. One method of approximating following times is to download each supplier's list of followers at regular intervals (e.g., hourly or daily) and see what followers were added. Due to the size of the suppliers in the sample and limits imposed by Twitter's API, it was infeasible to frequently and repeatedly download the suppliers' follower lists. As such, we estimated which followers from each list were new followers by leveraging the Gnip data that capture the suppliers' follower counts at the time that suppliers tweeted or were retweeted. Using this data, we deduced the suppliers' follower counts at the time of the supplier's first record (i.e., **b**) and the last record (i.e., **e**) of the two weeks of interest. We also counted the total number of followers (i.e., **n**) from the scraped lists of suppliers' followers[10]. The suppliers' new followers corresponded to the followers that matched with the following index on the suppliers' follower lists: [**n-e**+1, **n-b**] (see Figure 11). We assumed that users did not unfollow, or dissolve their connection with the supplier during the two-week period of analysis, which would have altered the index of each follower. Research shows that unfollowing rates tend to be negligible, particularly during short periods of time (Antoniades and Dovrolis 2015, Xu et al. 2013). We confirmed the low unfollowing rate by tracking the follower lists of 40 randomly selected Twitter users every day for a month, and we found that the average daily unfollowing rate was minimal (approximately 0.02% of the total follower count across all 40 users). Furthermore, of the new followers added in the two weeks of this study, suppliers retained on average 94% one year later.

(Figure 11 on next page)

[10] The suppliers' follower lists were scraped immediately once the week after the earthquake concluded. However, during the time it took to download these lists, suppliers could have gained more followers, which would have been reflected in the scraped data. As a result, **n** may be slightly different than **e**.

Figure 11 – Locating New Followers in Scraped Follower Lists

Sample follower list for a supplier where **n**=10, **b**=3, and **e**=8.

	Index	Follower's Twitter ID	
Most recent →	1	4100064	Followers from after period of interest
	2	3511860	
	3	4188869	New followers gained in period of interest (within the index [**n**-**e**+1, **n**-**b**] = [3,7])
	4	4518973	
	5	3429531	
	6	4360721	
	7	4949372	
	8	3948740	Followers from prior to period of interest
	9	4262104	
Oldest →	10	3898730	

The second step was to determine if each supplier's new followers were candidates. That is, we verified if, for every supplier, new followers received exposures from their friends of content posted by the supplier. This process involved attempting to match each new follower also as a follower of one of the supplier's retweeters at the time the exposure was sent. If a match was successful, we inferred that the new follower decided to follow after learning about the supplier through an exposure and classified that follower relationship as an internal link. If not, we classified the new follower relationship as an external link. Appendix D provides the technical details into the process of determining whether new follower relationships were internal or external links. Based on this analysis, the mean lag time between a candidate's receiving an exposure of a supplier's tweet and the candidate's decision to follow the same supplier was 9.28 hours, and 89.6% of candidates made this decision within 24 hours of receiving an exposure. As noted previously, retweeters may or may not have been following a supplier when they retweeted the supplier's content. A single retweeter could also have exposed its followers to multiple suppliers by retweeting more than one supplier's content. In fact, 41.2% of the retweeters in our sample issued retweets of more than one supplier. We also observed that on average

approximately 30% of each supplier's retweeters distributed the same individual supplier's content multiple times. This means that candidates following such retweeters were exposed repeatedly to a supplier by the same retweeter.

In total, the 47 suppliers' follower bases grew by 275,359 followers during the week before and after the earthquake. Figure 12 displays the cumulative number of followers gained across all of the suppliers in the studied two weeks. A little over 93% of the new followers connected with the suppliers after the earthquake, demonstrating that not only was tweeting and retweeting up after the disaster (see Figure 8) but network activity too. This finding also implies that the demand for information provided by the suppliers in our sample increased post-earthquake, which seems appropriate given that the suppliers provided information relevant to relief efforts.

Figure 12 – Cumulative Count of New Followers

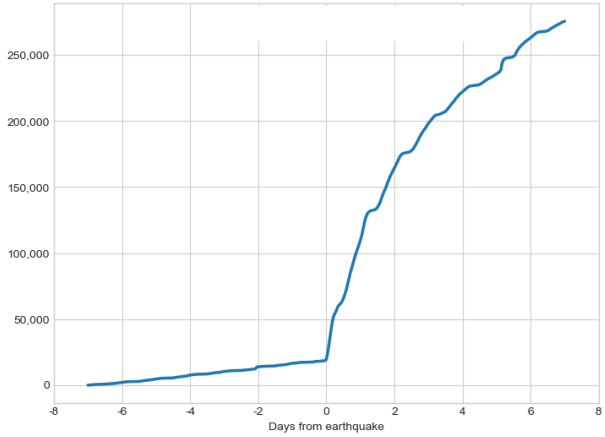


Table 15 breaks down the number of new followers into the frequencies of those that were classified as internal and external links. The table compares these numbers before versus after the earthquake across all of the suppliers as well as by supplier type. The results underscore the value of the internal mechanism as a means of gaining new followers since the percentage of internal links was substantial, especially once the emergency occurred. Prior to the earthquake, the percentage of total new follower

relationships classified as internal links was 35.4%, and this percentage climbed to 78.2% in the subsequent week. We observed that the number of retweets rose dramatically after the earthquake too, which could have driven the escalation of internal link formation in that period. However, our data reveal that the mean number of internal links per retweet before the earthquake was 0.117 but was 0.628 after the event. Therefore, the sharing of Twitter posts not only serves as a method to distribute information but also as a powerful and effective driver of new follower relationships, particularly during times of crisis.

Table 15 – External and Internal Links

	Pre-earthquake		Post-earthquake	
	External	Internal	External	Internal
Overall	11,972	6,561	56,017	200,809
Emergency Services	4,243	1,582	18,148	46,374
Government	7,524	4,782	36,042	145,987
Humanitarian	158	167	1,626	8,027
Medical	47	30	201	421

7. Structural Model Analysis

To estimate our structural model, we allowed each exposure to count as one observation since each exposure represented an opportunity for a candidate to consume and then start following a supplier in our sample. Once candidates established a new follower link with a supplier, they ceased to receive exposures of that supplier's tweets to motivate their decision to follow the supplier regardless of the amount of retweets they continued to receive from the supplier's retweeters. The total number of observations in our data was 2,042,306,645, and these exposures were generated by 65,897 retweeters that transmitted 375,698 retweets during the weeks before and after the disaster. Table 16 offers a summary of the notation used and an explanation of the operationalization of the variables in our model. The dependent variable y_{sti} represents the joint outcome of the two stages of our model and is binary. That is, y_{sti} is equal to 1 if i started to follow s after consuming an exposure of t published by s, and y_{sti} is equal to 0 otherwise. From the

analysis of internal and external mechanisms in Section 6, we were able to trace 207,370 internal links, and this translates into 207,370 observations where $y_{sti} = 1$.

Table 16 – Summary of Notation and Variable Operationalization

y_{sti}	Binary variable equal to 1 if the candidate followed the supplier from an exposure
b_{sti}	Candidate's number of friends
L_{sti}	Candidate's unobserved inverse login frequency
a_{sti}	Type of content (categorical variable distinguishing Actionable, Informative, and Other)
p_{sti}	Candidate's number of followers
q_{sti}	Candidate's retweeting frequency (measured as average daily tweeting rate)
r_{sti}	Supplier's count of followers
g_{sti}	Degrees of separation between candidate and supplier
f_{sti}	Number of tweets published by the supplier within 24 hours of the exposure
z_{sti}	Number of f_{sti} received by the candidate within 24 hours of the exposure
w_{sti}	Expected decrease in delay of information receipt (measured in hours)
d_{sti}	Binary variable equal to 1 if the exposure happened after the earthquake
ε_{sti}	Candidate's unobserved cost component

In the first stage of our model, we included b_{sti} and measured this variable as the number of friends that was scraped from *i*'s profile. The second stage of our model introduced a_{sti}, which was operationalized as a categorical variable that indicates if the content in *t* is Actionable, Informative, or Other. We based this classification on previous research regarding the types of information issued during humanitarian events, (Altay and Pal 2014, Moore and Verity 2014, Pedraza Martinez and Yan 2016, Qu et al. 2011). Actionable content attempts to motivate behavior through directions or suggestions, and Informative content contains factual reports, descriptions, and updates about the state of the operating environment and relief efforts. Finally, tweets belonging to the Other category convey messages that could not be defined as actionable or informative. Typically, they involved content related to opinions or emotional support. Due to the large number of tweets, we utilized text-mining techniques, specifically a supervised learning approach, to categorize the tweets. Appendix E provides the technical details related to this analysis.

As shown in Table 17, the Informative category had the highest amounts of tweets followed by the Actionable category. The Other category had the lowest number of tweets,

but the share of tweets belonging to this class experienced the most change by increasing from 4.15% before the earthquake to 10.87% after the earthquake. We compared the most frequently used words in each category during the pre and post-earthquake scenarios to understand how content evolved within each class. Before the earthquake, Actionable tweets were mainly concerned with instructing drivers where to drive based on accidents and road closures, whereas after the earthquake, the most common words for Actionable tweets were related to calls for donations and specific instructions for where and what to donate. Informative tweets in the pre-earthquake period were related to general news or updates about organizations' services and in the post-earthquake period presented information regarding emergency zones, rescue efforts, casualties, and updates on domestic and international humanitarian aid. Lastly, Other tweets discussed opinions and ideals of the country of Ecuador before the disaster. Following the earthquake, the most common words for tweets in the Other category pertained to uplifting and encouraging messages, such as solidarity, support, and unity.

Table 17 – Classification of Tweets and their Content

	Pre-earthquake		Post-earthquake	
Category	**Count**	**Example (translated from Spanish)**	**Count**	**Example (translated from Spanish)**
Actionable	1431	Road Macas-##SanJoséDeMorona is open for driving. Drive within the speed limits.	2487	When donating, prioritize bottled and non-perishable food. #EcuadorListoYSolidario #SismoEcuador https://t.co/JWaI5PPhOJ
Informative	3667	For the first time in history, Ecuador is a country that exports electrical energy #CocaCodoSinclair #InicioCocaCodo https://t.co/GACbz7S8lG	6184	A state of emergency has been declared in 6 provinces: Esmeraldas, Los Ríos, Manabí, Santa Elena, Guayas y Santo Domingo @JorgeGlas #SismoEcuador
Other	221	#Ecuador is considered one of the best destinations for retirees. #AllYouNeedIsEcuador https://t.co/tdzY8sAAVU	1058	We thank the security forces, doctors, and workers that have mobilized themselves with the patriotism that this emergency requires

To measure p_{sti}, we utilized the number of followers listed in *i*'s profile data. While Twitter profiles do not provide aggregate statistics on a user's retweeting behavior, the number of tweets posted by a user in its lifetime along with its account creation date is available. From this information, we can calculate a user's average daily tweeting rate as the total number of tweets divided by its tenure on Twitter. We posit that there exists a positive correlation between a user's tweeting and retweeting (Yang et al. 2010), so we employed *i*'s average daily tweeting rate as a proxy to measure q_{sti}. Because profile data for candidates was scraped once, the values for p_{sti} and q_{sti} (as well as for b_{sti}) vary across but not within the candidates. In contrast, r_{sti} was operationalized as the count of *s*'s followers at the time of each observation, which was available from the Gnip data.

The next variable in our model is g_{sti}, or the distance in the network between candidates and suppliers. We closely followed Goel et al. (2016)'s tree construction method for retweets. A tree represents the diffusion path for a tweet by marking each retweeter of that tweet as a node and drawing a link between nodes and their inferred parent, and a parent is the user that distributed the tweet to the retweeter. A retweeter's parent can be determined as the supplier of the original tweet or another retweeter, but it is also possible that a parent cannot be located. In such cases, the node is marked as a "root". Following Goel et al. (2016), we identified the parents of every retweeter in our data by first finding the set of potential parents. We then, if possible, connected each retweeter to the parent that most recently passed on the content. After constructing trees for all of the tweets in our data set, we were able to trace the degrees of separation between the retweeter of tweet *t* and *s,* and we used this value to measure g_{sti}[11]. We assigned missing values to nodes that were designated as roots since we could not completely trace how

[11] Technically, g_{sti} should be equal to the degrees of separation between the retweeter of *t* and *s* plus the value of 1 since the candidate is one more degree separated from *s*. Both measurements of g_{sti} are perfectly correlated and should yield the same results.

information reached these users, and this affected 41,228 retweets (or 10.9% of the total number of retweets).

We measured f_{sti} as the number of tweets published by s during a period of time leading up to t's exposure by i and measured z_{sti} as the count of f_{sti} received in i's feed during the same amount of time. Due to the rapid decay of information diffusion on social media platforms (Leskovec et al. 2009, Yang and Leskovec 2011), we focused on tweeting and retweeting activity during the 24 hours prior to t's exposure by i. Finally, we calculated w_{sti} as the time elapsed in hours between the time s published tweet t and the time i received t's exposure. Also, recall that w_{sti}, f_{sti}, and z_{sti} are moderated with d_{sti}, which is a binary variable that is established as 1 if the exposure occurred after the earthquake and 0 otherwise.

7.1. Model Estimation

The two stages described earlier together form the full model that analyzes the likelihood of a candidate beginning to follow a supplier after consuming an exposure. The outcome of the first stage is binary and unobserved, but success here is necessary to progress to the second stage. This means that y_{sti}=1 implies success at both stages of our model; however, if y_{sti}=0, we cannot distinguish in which stage there was a failure. Because of these aspects, we used a bivariate probit model with partial observability (Poirier 1980). We allowed the unobserved variables at each stage to be correlated, and the vector θ contains the model parameters to be estimated. We note that the estimation process, for example, cannot distinctly identify β_1 and σ_L^2 but can identify $\frac{\beta_1}{\sigma_L^2}$. The signs for β_1 and $\frac{\beta_1}{\sigma_L^2}$ are identical, and determining the direction of β_1 without the exact parameter estimate still allows us to gauge the partial effect of the associated variable on the dependent variable. Thus, estimating the value of the ratio $\frac{\beta_1}{\sigma_L^2}$ is sufficient for our study.

We rewrite Equations 2 and 6 in a simpler form, and the full specification of the model for estimation is provided in Equation 7.

$$P1 = P\left(\beta_0 - \beta_1 \log b_{sti} > \frac{\log L_{sti} - L}{\sigma_L^2}\right)$$

$$P2 = P\Big(\gamma_0 + \gamma_1 \log a_{sti} + \gamma_2 \log p_{sti} + \gamma_3 \log q_{sti} + \gamma_4 \log r_{sti} + \gamma_5 g_{sti} + \gamma_6 \log(f_{sti} - z_{sti}) + \gamma_7 w_{sti} + \gamma_8 d_{sti} + \gamma_9 w_{sti} * d_{sti} - \gamma_{10} \log\left(e^{f_{sti}} - e^{z_{sti}}\right) - \gamma_{11} \log\left(e^{f_{sti}} - e^{z_{sti}}\right) * d_{sti} > \frac{\log \varepsilon_{sti} - \varepsilon}{\sigma_\varepsilon^2} \Big| \beta_0 - \beta_1 \log b_{sti} > \frac{\log L_{sti} - L}{\sigma_L^2}\Big)$$

$$P(y_{sti} = 1) = P1 * P2$$

$$P(y_{sti} = 0) = 1 - P1 * P2$$

$$L_{sti}, e_{sti} \sim N(0,1)$$

$$cor(L_{sti}, e_{sti}) = \rho$$

$$\theta = \{\rho, \beta_0, \beta_1, \gamma_0, \dots, \gamma_{11}\} \quad (7)$$

The two stages must be estimated jointly, so the log-likelihood function is

$$\mathcal{L}(\beta, \gamma, \rho) = \sum[y_{sti} \log(\mathbf{\Phi}(\mathrm{X}_1\beta, \mathrm{X}_2\gamma, \rho)) + (1 - y_{sti}) \log(1 - \mathbf{\Phi}(\mathrm{X}_1\beta, \mathrm{X}_2\gamma, \rho))], \quad (8)$$

where β represents the vector of parameters in the first stage and γ represents the vector of parameters in the second stage. Note that $\mathbf{\Phi}(\cdot)$ represents the bivariate standard normal distribution.

Recall that the total number of observations in our data is roughly 2.042 billion and that the count of observations where y_{sti}=1 is 207,370. Thus, the percentage of successful events is very small (approximately 0.01%), and our sample can be considered to include rare event data. Estimating models using samples with rare event data can be problematic since coefficients are biased. A strategy to address this bias involves the use of response-based or choice-based sampling (King and Zeng 2001). Suppose that in a sample of rare event data, the percentage of successful events (i.e., Y=1) is μ and the percentage of unsuccessful events (i.e., Y=0) is 1-μ. Response-based sampling involves

creating a new sample composed of two sub-samples: (1) all or a random sample of observations where Y=1 and (2) a random sample of observations where Y=0. In this new sample, the proportion of observations where Y=1 is now $\bar{y}$ and $\bar{y} > \mu$. While response-based sampling helps ensure there is a sufficient number of positive events, it yields inconsistent and asymptotically biased estimates since observations are selected on the dependent variable, but this can be statistically corrected for using Manski and Lerman (1977)'s weighted maximum likelihood estimator (WMLE). For the bivariate probit model with partial observability, the WMLE can be obtained by maximizing the weighted log-likelihood function presented in Equation 9.

$$\mathcal{L}_w(\beta, \gamma, \rho) = \sum \left[\frac{\mu}{\bar{y}} y_{sti} \log\left(\mathbf{\Phi}(\mathrm{X}_1\beta, \mathrm{X}_2\gamma, \rho)\right) + \frac{1-\mu}{1-\bar{y}} (1 - y_{sti}) \log(1 - \mathbf{\Phi}(\mathrm{X}_1\beta, \mathrm{X}_2\gamma, \rho)) \right] \quad (9)$$

We applied the response-based sampling technique and formed a new sample. In line with Singh (2005), we included all of the observations where y_{sti}=1, and we selected a stratified random sample across the suppliers for an equivalent number of observations where y_{sti}=0. The percentage of positive events therefore was 50%. Because we dropped observations with missing values for g_{sti}, the size of the sample from response-based sampling was 371,420. We mean-centered w_{sti} and $\log\left(e^{f_{sti}} - e^{z_{sti}}\right)$ since these variables are moderated with d_{sti}, and we estimated the model using WMLE with robust standard errors. Additionally, we verified that our choice of how the response-based sampling method was adopted did not drive our results by creating other samples. These alternates included samples that maintained the same ratio of positive to negative events as well as samples that varied the ratio of positive to negative events. We estimated our model with the alternate samples, and the results were consistent and robust to changes in how the response-based sampling method was applied. The results are available from the authors upon request.

7.2. Results

We list the descriptive statistics of the key variables for the sample used to estimate the model in Table 18. The table presents the binary and categorical variables along with their means first, followed by the descriptive statistics for continuous variables before any transformation is applied. We then present the results attained from the WMLE method in Table 19. The table displays the first stage results in the top set of coefficients. In our model, we assumed that the amount of incoming information into a user's Twitter feed is linearly associated with the user's count of friends, so, we conjectured that a candidate is less likely to consume a certain exposure as its number of friends increases. The value of the coefficient (β_1) for b_{sti} is negative and statistically significant, which implies a negative association between a candidate's friend count and the probability of consumption for an exposure. This result not only aligns with our expectations but with what researchers have previously found (e.g., Shi et al. 2014).

Table 18 – Descriptive Statistics for Key Variables

	Mean	**Std. Dev.**	**Min**	**Max**
y_{sti}	0.514			
a_{sti} (Actionable)	0.265			
a_{sti} (Other)	0.052			
a_{sti} (Informative)	0.683			
d_{sti}	0.902			
b_{sti}	824.960	11,505.240	1	1,548,099
p_{sti}	1,298.447	31,985.390	0	8,091,149
q_{sti}	1.032	8.298	0	1,580.021
r_{sti}	230,527.600	164,278	532	648,749
g_{sti}	1.210	0.653	1	29
f_{sti}	90.257	55.165	0	361
z_{sti}	10.941	14.276	0	192
w_{sti}	1.726	5.350	0.001	302.214
371,420 observations				

(Table 19 on next page)

Table 19 – Results of the Weighted Maximum Likelihood Estimation

	Coeff.	(Robust Std. Err.)
Stage 1: Consumption		
β_0 (Intercept)	-2.074***	(0.092)
β_1 (b_{sti})	-0.240***	(0.011)
Stage 2: Follow Decision		
γ_0 (Intercept)	-4.860***	(0.029)
$\gamma_{1_{action}}$ (a_{sti} = Actionable)	0.031***	(0.003)
$\gamma_{1_{other}}$ (a_{sti} = Other)	0.062***	(0.006)
γ_2 ($\log p_{sti}$)	0.042***	(0.002)
γ_3 ($\log q_{sti}$)	0.127***	(0.011)
γ_4 ($\log r_{sti}$)	0.016***	(0.002)
γ_5 ($\log g_{sti}$)	-0.087***	(0.004)
γ_6 ($\log(f_{sti} - z_{sti})$)	0.149***	(0.003)
γ_7 (w_{sti})	0.016***	(0.002)
γ_8 (d_{sti})	0.318***	(0.007)
γ_9 ($w_{sti} * d_{sti}$)	-0.018***	(0.002)
γ_{10} ($\log(e^{f_{sti}} - e^{z_{sti}})$)	-0.004***	(2E-04)
γ_{11} ($\log(e^{f_{sti}} - e^{z_{sti}}) * d_{sti}$)	0.002***	(2E-04)
rho	0.954***	(0.006)
Observations	371,420	
Pseudo log-likelihood	-383.259	

** p<0.1, ** p<0.05, *** p<0.01*

The second set of coefficients presented in Table 6 corresponds to the estimated parameters for the second stage of our structural model. To measure the type of content in *t*, we included a categorical variable, a_{sti}, which distinguishes the content in *t* as Actionable, Informative, or Other. Because the frequency of messages classified as Informative was highest, we set the Informative category as the baseline category. The results demonstrate that a candidate is more likely to begin following a supplier when the content in *t* belongs to the Actionable or Other classes as compared to the Informative class (see the positive and significant values for the $\gamma_{1_{action}}$ and $\gamma_{1_{other}}$ coefficients). This finding is important for HOs that often send crucial information in Actionable tweets (e.g.,

evacuation instructions, directions to shelters) to learn that Actionable content may also spur online action in the form of initiating new follower links.

Furthermore, we observed that the future benefit of becoming a distributor of a supplier's content affects a candidate's decision to follow the supplier. In particular, a candidate that has a larger audience of followers and retweets more frequently is more likely to begin following a supplier, as shown by the positive and significant coefficient estimates for γ_2 and γ_3 respectively. We also evaluated the relationship between a supplier's number of followers and a candidate's likelihood of following the supplier since we argued that candidates will prefer to follow larger suppliers. We found support for this since the estimated value of the coefficient (γ_4) for r_{sti} was positive and statistically significant. The outcome that candidates have a higher probability of following suppliers that have already accumulated a substantial amount of followers also provides evidence of preferential attachment in social media networks (Barabási and Albert 1999).

The value of the coefficient (γ_5) for g_{sti} was negative and significant, implying that candidates farther away in the network are less probable to follow a supplier. This result diverges from our expectation that a candidate's probability of becoming a supplier's new follower is positively associated with network distance since the candidate is more likely to obtain a greater utility from information that is locally scarce and novel (Aral and Van Alstyne 2011). One explanation for our finding of a negative coefficient for g_{sti} is that as the degrees of separation grow between a candidate and supplier, the supplier's content becomes too novel such that there is no overlap with the interests of the candidate. Therefore, a candidate at a significant distance from the supplier may anticipate not earning much utility from receiving the supplier's tweets in the future, which will lessen its propensity to start following the supplier.

According to the table, the coefficients (γ_6 and γ_7) for the variables $\log(f_{sti} - z_{sti})$ and w_{sti} are positive and statistically significant. These findings mean that a candidate's conditional probability of starting to follow a supplier increases if doing so will result in an increase in the expected amount of coverage of the supplier's activity in addition to a decrease in the delay of information receipt. That is, candidates are more prone to follow a supplier when they anticipate they will obtain information more completely and rapidly from doing so. As such, a candidate's decision to follow a supplier is partly contingent upon the performance of the retweeters with regards to how fully and quickly information is distributed. We also tested how the effect of the decrease in the delay of information receipt on the probability of following is moderated by whether or not the exposure happened after the disaster. Our results indicate that a candidate's conditional probability of becoming a supplier's new follower is higher after a disaster since the estimated value of the coefficient (γ_8) for d_{sti} is positive and significant. In addition, the coefficient estimate for the interaction of w_{sti} and d_{sti} (γ_9) is negative and significant, which suggests that candidates that receive exposures after an emergency tend to convert to new followers when the expected reduction in the time lag for information receipt is smaller than the expected reduction under no emergency conditions. Our finding demonstrates that, after a disaster, a candidate is prompted to follow a supplier even when the improvement in how quickly it can obtain the supplier's information is not as large, and this behavior may be driven by the urgent atmosphere and by information perishing at a faster rate.

Finally, we tested how the change in information processing cost, driven by the change in the amount of information received from a supplier after following, affects a candidate's choice to follow a supplier. The value of the coefficient (γ_{10}) for $\log\left(e^{f_{sti}} - e^{z_{sti}}\right)$ is negative and statistically significant, so as the marginal increase in the volume of information received by a candidate upon following rises, the candidate is less

likely to connect with the supplier. This behavior is not surprising given that candidates do not want to incur a higher information processing cost. However, we also found that the parameter estimate for the interaction between $\log\left(e^{f_{sti}} - e^{z_{sti}}\right)$ and d_{sti} is positive and significant (see value for γ_{11}). This means that, under a crisis event, a candidate's conditional probability of following a supplier increases even though the amount of information to be received and thereby the cost to process this information escalates also. An implication from our finding is that candidates may perceive the cost of information processing to be lower under a crisis scenario. The change in the calculation of information processing cost may be attributed to users feeling the need to obtain as much information as possible to alleviate the uncertainty that is typically rampant once a disaster materializes.

8. Robustness Checks

To validate the robustness of our findings, we conducted several robustness checks. First, we accounted for our data being potentially right-censored. Candidates that received exposures towards the end of the week after the earthquake may have consumed an exposure of a supplier and started following the supplier, but these decisions may have been made after data collection was complete. We ensured that censoring did not affect our results by eliminating any observations where the time of exposure occurred within the last 24 hours of the period of interest. By the termination of the week after the earthquake, candidates that received exposures during the final day may not have had enough time to complete the stages of consumption and deciding whether to follow the supplier. We chose to drop observations within the last 24 hours since nearly all of the candidates in our data that followed a supplier after consuming an exposure of the supplier did so within 24 hours of receiving the exposure. This reduced our sample by 20,146 observations (0.001% of the sample). Using this data, we re-estimated the model, and the

results were consistent with those presented in Table 6, demonstrating that our findings are robust to potential censoring effects. The results are available in Appendix F.

Second, it is possible that some candidates are more likely to follow a specific supplier because they follow other suppliers in our sample. While the sampled suppliers belong to four different categories (see Table 1), all represent legitimate organizations that are involved with disaster relief and public services. Candidates that follow multiple suppliers from our data demonstrate an interest in these organizations' content and thus may be more inclined to follow another supplier after consuming an exposure of that supplier. Of the 126,576 unique candidates that established a new follower relationship with a supplier during the weeks before and after the earthquake, 62,048 candidates followed more than one supplier in the same time interval. We controlled for this type of behavior for candidate i by counting how many other suppliers i followed at the time that i received an exposure of s's tweet t (v_{sti}). On average, candidates already followed 0.275 suppliers at the time of an exposure. We included this variable in the second stage of our model and estimated the model again. The results indicated that the parameter estimates and significance levels were robust and consistent with those listed in Table 6. Furthermore, the coefficient for v_{sti} was positive and significant (p-value <0.001), confirming that candidates are indeed more likely to follow a supplier when they have previously connected with other suppliers. The results from this analysis are also available in Appendix F.

While we account for a candidate's number of followers in our model, another important characteristic of a candidate is its ratio of its counts of followers-to-friends count. Typically, Twitter accounts that represent organizations or celebrities have high ratios of followers-to-friends. As a result, it is generally perceived to be advantageous for a user's reputation to have a larger followers-to-friends ratio, and users with a high

followers-to-friends ratio may be more reluctant to follow to maintain this ratio. Research on spam detection also indicates that spammers and bots tend to follow many other users and therefore possess low followers-to-friends ratios (Yardi et al. 2009). We controlled for users' preferences to sustain their followers-to-friends ratios in addition to the possible presence of bots among the candidates in our study by including a measure of candidates' followers-to-friends ratios (ϕ_{sti}). This variable was logged to account for possible nonlinearity. We re-estimated the parameters after including candidates' followers-to-friends ratios in the second stage of the model. The parameters and significance levels are again robust and consistent with those listed in Table 6. We also observed that, as expected, a candidate's followers-to-friends ratio was negatively associated with the probability of following a supplier ($p<0.001$). Please refer to Appendix F for the results from the third robustness check.

The final robustness check we conducted controlled for the expectation that retweets of popular tweets have a higher probability of being consumed by the candidate. To accomplish this, we identified how many retweets a tweet *t* had accumulated at the time that the candidate received the exposure of *t* (δ_{sti}). A tweet's popularity rises as it earns more retweets. We logged the count of retweets for possible nonlinearity and inserted this variable into the consumption stage of our structural model. The results of the model with $\log \delta_{sti}$ are consistent with the outcomes presented in Table 6, and we found support for our argument that a candidate's probability of consuming an exposure of *t* is positively related to the popularity of *t*. The results of the fourth robustness check are also shown in Appendix F.

9. Conclusion

During humanitarian crises, HOs need to relay important and potentially life-saving information rapidly and to as many of their stakeholders as possible. HOs have

started to leverage social media platforms because information is shared instantaneously to their followers through this technology. Furthermore, these platforms typically have a sharing function that allows users to distribute another user's content to their own networks, which further accelerates the diffusion of social media content. One method for HOs of guaranteeing the diffusion of their social media content is to have a larger set of followers, which translates into a larger audience size for HOs' content. This study examines the mechanisms that drive the growth of HOs' follower bases. Specifically, the external mechanism relies on stimuli outside of the network of users involved in sharing HOs' content while the internal mechanism depends on users learning about HOs through content distribution. We specified a two-stage structural model to analyze what influences the probability that an individual user becomes an internal link, or starts to follow a HO after learning about the organization through the sharing of content authored by the HO. To estimate the model, we collected a unique data set from Twitter with dynamic network data for HOs and other organizations directly involved with disaster relief during the 2016 Ecuador Earthquake.

The results from our study indicate that, especially in the post-disaster scenario, the internal mechanism is a significant driver of the expansion of HOs' follower bases. This means that the sharing of content is not only valuable for disseminating HOs' content but also to catalyze the formation of new follower relationships. Our finding carries important implications for HOs. First, HOS may not be able to spend the time or financial capital required to build follower links through the external mechanism since they are often constrained by limited resources. However, our study shows that HOs can rely on their networks to help expand follower bases at no cost. Another implication is that HOs should develop policies towards mobilizing and encouraging users to distribute their content. For

example, the American Red Cross initiated the Digital Volunteer Program in 2013[12], and volunteers in this program help monitor online conversations during disasters and answer questions from social media users. Based on this study's result of the prominence of internal links, Digital Volunteers should also play an active role in disseminating content to spread awareness about the American Red Cross and motivate users to start following this organization.

Moreover, this study provides guidance towards differentiating what HOs can do and what HOs must rely on their network of information distributors to do in order to gain internal links. HOs can adjust the type of content and the frequency of publishing new content to attract new followers. In particular, we found that users prefer to not follow HOs that publish social media content too frequently before the disaster, but this preference reverses once a disaster has materialized, likely to reduce the uncertainty from the emergency. Hence, under non-emergency conditions, HOs should concentrate on determining the optimal timing of social media content release as in Caro et al. (2018). After a disaster, HOs should attempt to keep their audience well-informed and produce information frequently.

This study and its investigation of the drivers of new follower links for HOs can be extended by future research. Because of data limitations, we do not include the geographic location of candidates in the structural model. However, we anticipate that users located within the disaster zone are more likely to start following HOs due to being personally impacted by the disaster. HOs may also value earning new followers that are local to the disaster to ensure that information about resources and services are received by beneficiaries. Therefore, future research can evaluate how the physical location of users influences their decision to start following an HO. An alternative avenue of future research

[12] https://redcrosschat.org/digitalvolunteer/

is to more deeply explore the behavior of new followers once the time of crisis has passed. We found that, on average, the 47 organizations in our data retained 94% of their new followers one year after the end of data collection. Future research can further study the retention rate of new followers in addition to their level of engagement as information distributors. Finally, future research can assess the economic value of internal links given that followers can be purchased. Twitter, as an example, allows firms to purchase followers through their "followers campaigns" product for approximately $3 per new follower. Using this value as a benchmark, future research can assign monetary value to internal links overall as well as to the individual variables that affect the probability of internal link conversion from our structural model.

Acknowledgements
We are grateful to the social media team at Red Cross Ecuador for valuable insights into how Twitter is being utilized at this organization. We would also like to acknowledge Zhan (Michael) Shi from Arizona State University for helpful feedback regarding the structural model.

REFERENCES

Acimovic J, Goentzel J (2016) Models and metrics to assess humanitarian response capacity. *J. Oper. Manag.* 45:11–29.

Afuah A, Tucci CL (2012) Crowdsourcing as a solution to distant search. *Acad. Manage. Rev.* 37(3):355–375.

Allon G, Zhang DJ (2018) Managing Service Systems in the Presence of Social Networks. *Work. Pap.*

Altay N, Labonte M (2014) Challenges in humanitarian information management and exchange: evidence from Haiti. *Disasters* 38(s1):S50–S72.

Altay N, Pal R (2014) Information Diffusion among Agents: Implications for Humanitarian Operations. *Prod. Oper. Manag.* 23(6):1015–1027.

Anderson SP, de Palma A (2009) Information Congestion. *RAND J. Econ.* 40(4):688–709.

Antoniades D, Dovrolis C (2015) Co-evolutionary dynamics in social networks: a case study of Twitter. *Comput. Soc. Netw.* 2(1):14.

Aral S, Van Alstyne M (2011) The Diversity-Bandwidth Trade-off. *Am. J. Sociol.* 117(1):90–171.

Arrow KJ, Bernheim BD, Feldstein MS, McFadden DL, Poterba JM, Solow RM (2011) 100 Years of the American Economic Review: The Top 20 Articles. *Am. Econ. Rev.* 101(1):1–8.

Asur S, Huberman BA, Szabo G, Wang C (2011) Trends in Social Media: Persistence and Decay. *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media*. ICWSM '11. (Barcelona, Spain), 434–437.

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. *Proc. Fourth ACM Int. Conf. Web Search Data Min.* WSDM '11. (ACM, Kowloon, Hong Kong), 65–74.

Balcik B, Beamon BM, Krejci CC, Muramatsu KM, Ramirez M (2010) Coordination in humanitarian relief chains: Practices, challenges and opportunities. *Int. J. Prod. Econ.* 126(1):22–34.

Bapna R, Umyarov A (2015) Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks. *Manag. Sci.* 61(8):1902–1920.

Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. *Science* 286(5439):509–512.

Bhattacharya D, Ram S (2012) Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. *Proc. 2012 Int. Conf. Adv. Soc. Netw. Anal. Min.* ASONAM '12. (IEEE, Istanbul, Turkey), 966–971.

Bonacich P (1972) Technique for Analyzing Overlapping Memberships. *Sociol. Methodol.* 4,:176–185.

Bowsher CG (2007) Modelling security market events in continuous time: Intensity based, multivariate point process models. *J. Econom.* 141(2):876–912.

Boyd D, Golder S, Lotan G (2010) Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *Proc. 2010 43rd Hawaii Int. Conf. Syst. Sci.* HICSS '10. (IEEE, Honolulu, HI, USA), 1–10.

Bremaud P, Massoulie L (1996) Stability of Nonlinear Hawkes Processes. *Ann. Probab.* 24(3):1563–1588.

Caro F, Martinez-de-Albeniz V (2018) Managing Online Content to Build a Follower Base: Model and Applications. *Work. Pap.*

Castillo C, Mendoza M, Poblete B (2011) Information Credibility on Twitter. *Proc. 20th Int. Conf. World Wide Web*. WWW '11. (ACM, Hyderabad, India), 675–684.

Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring User Influence in Twitter: The Million Follower Fallacy. *Proc. Fourth Int. Conf. Weblogs Soc. Media*. ICWSM '10. (Washington, DC, USA), 10–17.

Chandrasekaran D, Tellis GJ (2007) A Critical Review of Marketing Research on Diffusion of New Products. Malhotra NK, ed. *Rev. Mark. Res. Vol. 3*. (Emerald Group Publishing Limited), 39–80.

Charikar MS (2002) Similarity Estimation Techniques from Rounding Algorithms. *Proc. Thiry-Fourth Annu. ACM Symp. Theory Comput.* STOC '02. (ACM, Montreal, Canada), 380–388.

Ciampaglia GL, Flammini A, Menczer F (2015) The production of information in the attention economy. *Sci. Rep.* 5:9452.

CNN Español (2016) #DesaparecidosEC: Ecuador recurre a las redes para encontrar a desaparecidos del terremoto. *CNN* http://cnnespanol.cnn.com/2016/04/18/ecuador-recurre-a-las-redes-sociales-para-encontrar-a-los-desaparecidos-del-terremoto/.

Coscia M (2014) Average is Boring: How Similarity Kills a Meme's Success. *Sci. Rep.* 4:6477.

Coscia M (2018) Popularity Spikes Hurt Future Chances For Viral Propagation of Protomemes. *Commun. ACM* 61(1):70–77.

Coxe S, West SG, Aiken LS (2013) Generalized Linear Models. Little TD, ed. *Oxf. Handb. Quant. Methods Psychol. Vol 2 Stat. Anal.* (Oxford University Press, New York, NY), 26–51.

Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.* 105(41):15649–15653.

Cui R, Gallino S, Moreno A, Zhang DJ (2017) The Operational Value of Social Media Information. *Prod. Oper. Manag.* Forthcoming.

Currion P, Silva C de, Van de Walle B (2007) Open Source Software for Disaster Management. *Commun. ACM* 50(3):61–65.

Daley DJ, Vere-Jones D (2003) *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods* 2nd ed. (Springer-Verlag, New York).

Davidov D, Tsur O, Rappoport A (2010) Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. *Proc. 23rd Int. Conf. Comput. Linguist. Posters*. (Association for Computational Linguistics, Stroudsburg, PA), 241–249.

Day JM, Melnyk SA, Larson PD, Davis EW, Whybark DC (2012) Humanitarian and Disaster Relief Supply Chains: A Matter of Life and Death. *J. Supply Chain Manag.* 48(2):21–36.

Dellarocas C, Sutanto J, Calin M, Palme E (2015) Attention Allocation in Information-Rich Environments: The Case of News Aggregators. *Manag. Sci.* 62(9):2543–2562.

Dobson AJ, Barnett AG (2008) *An Introduction to Generalized Linear Models* 3rd ed. (Chapman and Hall/CRC, Boca Raton, FL).

Dodds PS, Watts DJ (2005) A generalized model of social and biological contagion. *J. Theor. Biol.* 232(4):587–604.

Eftekhar M, Li H, Van Wassenhove LN, Webster S (2017) The Role of Media Exposure on Coordination in the Humanitarian Setting. *Prod. Oper. Manag.* 26(5):802–816.

Ellison NB, Steinfield C, Lampe C (2007) The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *J. Comput.-Mediat. Commun.* 12(4):1143–1168.

Embrechts P, Liniger T, Lin L (2011) Multivariate Hawkes processes: an application to financial data. *J. Appl. Probab.* 48(A):367–378.

Ergun Ö, Gui L, Heier Stamm JL, Keskinocak P, Swann J (2014) Improving Humanitarian Operations through Technology-Enabled Collaboration. *Prod. Oper. Manag.* 23(6):1002–1014.

Falkinger J (2007) Attention economies. *J. Econ. Theory* 133(1):266–294.

Farrell H, Drezner DW (2008) The power and politics of blogs. *Public Choice* 134(1–2):15.

Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76(5):378–382.

Fletcher R (2013) *Practical Methods of Optimization* 2nd ed. (John Wiley & Sons, Chichester, England).

Fowler JH, Christakis NA (2010) Cooperative behavior cascades in human social networks. *Proc. Natl. Acad. Sci.* 107(12):5334–5338.

Gabaix X, Laibson D, Moloche G, Weinberg S (2006) Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *Am. Econ. Rev.* 96(4):1043–1068.

Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W (2010) Outtweeting the twitterers-predicting information cascades in microblogs. *Proc. 3rd Wonference Online Soc. Netw.* WOSN'10. (USENIX Association, Boston, MA, USA), 3–3.

Gao H, Barbier G, Goolsby R (2011) Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. *IEEE Intell. Syst.* 26(3):10–14.

Gardner W, Mulvey EP, Shaw EC (1995) Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol. Bull.* 118(3):392–404.

Goel S, Anderson A, Hofman J, Watts DJ (2016) The Structural Virality of Online Diffusion. *Manag. Sci.* 62(1):180–196.

Goldenberg J, Han S, Lehmann DR, Hong JW (2009) The Role of Hubs in the Adoption Process. *J. Mark.* 73(2):1–13.

Goldenberg J, Libai B, Muller E (2001) Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Mark. Lett.* 12(3):211–223.

Gomez Rodriguez M, Gummadi K, Scholkopf B (2014) Quantifying Information Overload in Social Media and its Impact on Social Contagions. *Proc. Eighth Int. Conf. Weblogs Soc. Media*. ICWSM '14. (Ann Arbor, MI, USA), 170–179.

Granovetter M (1978) Threshold Models of Collective Behavior. *Am. J. Sociol.* 83(6):1420–1443.

Granovetter MS (1973) The Strength of Weak Ties. *Am. J. Sociol.* 78(6):1360–1380.

Gu B, Ye Q (2014) First Step in Social Media: Measuring the Influence of Online Management Responses on Customer Satisfaction. *Prod. Oper. Manag.* 23(4):570–582.

Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: a survey. *ACM SIGMOD Rec.* 42(2):17–28.

Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. (International World Wide Web Conferences Steering Committee, 2488033), 729–736.

Haas MR, Criscuolo P, George G (2015) Which Problems to Solve? Online Knowledge Sharing and Attention Allocation in Organizations. *Acad. Manage. J.* 58(3):680–711.

Hardin JW, Hilbe JM (2007) *Generalized Linear Models and Extensions* 2nd ed. (Stata Press, College Station, TX).

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. (Springer Science & Business Media, New York).

Hawkes AG (1971) Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika* 58(1):83–90.

Hawkes AG, Oakes D (1974) A Cluster Process Representation of a Self-Exciting Process. *J. Appl. Probab.* 11(3):493–503.

Hodas NO, Kooti F, Lerman K (2013) Friendship Paradox Redux: Your Friends Are More Interesting Than You. *Proc. Seventh Int. AAAI Conf. Weblogs Soc. Media*. ICWSM '13. (Cambridge, MA, USA), 225–233.

Hodas NO, Lerman K (2012) How Visibility and Divided Attention Constrain Social Contagion. *Proc. 2012 ASEIEEE Int. Conf. Soc. Comput. 2012 ASEIEEE Int. Conf. Priv. Secur. Risk Trust*. SOCIALCOM-PASSAT '12. (IEEE, Amsterdam, Netherlands), 249–257.

Hoffman DL, Fodor M (2010) Can You Measure the ROI of Your Social Media Marketing? *MIT Sloan Manag. Rev.* 52(1):41–49.

Holguín-Veras J, Jaller M, Van Wassenhove LN, Pérez N, Wachtendorf T (2012) On the unique features of post-disaster humanitarian logistics. *J. Oper. Manag.* 30(7–8):494–506.

Holguín-Veras J, Pérez N, Jaller M, Van Wassenhove LN, Aros-Vera F (2013) On the appropriate objective function for post-disaster humanitarian logistics models. *J. Oper. Manag.* 31(5):262–280.

Hong L, Dan O, Davison BD (2011) Predicting Popular Messages in Twitter. *Proc. 20th Int. Conf. Companion World Wide Web*. WWW '11. (ACM, Hyderabad, India), 57–58.

Hong W, Thong JYL, Tam KY (2004) Does Animation Attract Online Users' Attention? The Effects of Flash on Information Search Performance and Perceptions. *Inf. Syst. Res.* 15(1):60–86.

Huang Y, Singh PV, Ghose A (2015) A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media. *Manag. Sci.* 61(12):2825–2844.

Huberman B, Romero DM, Wu F (2008) Social networks that matter: Twitter under the microscope. *First Monday* 14(1–5).

Huberman BA, Romero DM, Wu F (2009) Crowdsourcing, attention and productivity. *J. Inf. Sci.* 35(6):758–765.

Huberman BA, Wu F (2008) The economics of attention: maximizing user value in information-rich environments. *Adv. Complex Syst.* 11(04):487–496.

Iyer G, Katona Z (2015) Competing for Attention in Social Communication Markets. *Manag. Sci.* 62(8):2304–2320.

Kaigo M (2012) Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Commun. Rev.* 34,:19–35.

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. *Proc. Ninth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* KDD '13. (ACM, Washington, DC, USA), 137–146.

King G, Zeng L (2001) Logistic Regression in Rare Events Data. *Polit. Anal.* 9(2):137–163.

Kobayashi R, Lambiotte R (2016) TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. *Proc. Tenth Int. AAAI Conf. Web Soc. Media.* ICWSM '16. (Cologne, Germany), 191–200.

Kogan M, Palen L, Anderson KM (2015) Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable During Hurricane Sandy. *Proc. 18th ACM Conf. Comput. Support. Coop. Work Soc. Comput.* CSCW '15. (ACM, Vancouver, Canada), 981–993.

Korolov R, Peabody J, Lavoie A, Das S, Magdon-Ismail M, Wallace W (2015) Actions Are Louder Than Words in Social Media. *Proc. 2015 IEEEACM Int. Conf. Adv. Soc. Netw. Anal. Min. 2015.* ASONAM '15. (ACM, New York, NY, USA), 292–297.

Kossinets G, Watts DJ (2006) Empirical Analysis of an Evolving Social Network. *Science* 311(5757):88–90.

Kovács G, Spens KM (2007) Humanitarian logistics in disaster relief operations. *Int. J. Phys. Distrib. Logist. Manag.* 37(2):99–114.

Kumar S, Havey T (2013) Before and after disaster strikes: A relief supply chain decision support framework. *Int. J. Prod. Econ.* 145(2):613–629.

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a Social Network or a News Media? *Proc. 19th Int. Conf. World Wide Web*. WWW '10. (ACM, Raleigh, NC, USA), 591–600.

van der Laan E, van Dalen J, Rohrmoser M, Simpson R (2016) Demand forecasting and order planning for humanitarian logistics: An empirical assessment. *J. Oper. Manag*. 45:114–122.

Lam HKS, Yeung ACL, Cheng TCE (2016) The impact of firms' social media initiatives on operational efficiency and innovativeness. *J. Oper. Manag*. 47–48:28–43.

Landis JR, Koch GG (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

Lee YJ, Hosanagar K, Tan Y (2015) Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. *Manag. Sci.* 61(9):2241–2258.

Lerman K, Ghosh R (2010) Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. *Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media*. ICWSM '10. (Washington, DC, USA), 90–97.

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the Dynamics of the News Cycle. *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* KDD '09. (ACM, Paris, France), 497–506.

Lobel I, Sadler E, Varshney LR (2016) Customer Referral Incentives and Social Media. *Manag. Sci.* 63(10):3514–3529.

Long DC, Wood DF (1995) The logistics of famine relief. *J. Bus. Logist.* 16(1):213.

Manku GS, Jain A, Das Sarma A (2007) Detecting Near-duplicates for Web Crawling. *Proc. 16th Int. Conf. World Wide Web*. WWW '07. (ACM, Banff, Canada), 141–150.

Manning CD, Schütze H (1999) *Foundations of Statistical Natural Language Processing* (The MIT Press, Cambridge, MA).

Manski CF, Lerman SR (1977) The Estimation of Choice Probabilities from Choice Based Samples. *Econometrica* 45(8):1977–1988.

Masuda N, Takaguchi T, Sato N, Yano K (2013) Self-Exciting Point Process Modeling of Conversation Event Sequences. *Temporal Netw.* Understanding Complex Systems. (Springer, Berlin, Heidelberg), 245–264.

McCullagh P, Nelder JA (1989) *Generalized Linear Models* 2nd ed. (Chapman and Hall/CRC, London).

Mei H, Eisner JM (2017) The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. *Adv. Neural Inf. Process. Syst. 30*. (Long Beach, CA), 6754–6764.

Meier P (2012) How the UN Used Social Media in Response to Typhoon Pablo. *iRevolution*. http://irevolution.net/2012/12/08/digital-response-typhoon-pablo/. Last accessed: 17 Jul. 2014.

Meier P (2015) *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response* (CRC Press, Boca Raton, FL).

Miller JH (1998) Active Nonlinear Tests (ANTs) of Complex Simulation Models. *Manag. Sci.* 44(6):820–830.

Mishra S, Rizoiu MA, Xie L (2016) Feature Driven and Point Process Approaches for Popularity Prediction. *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag.* CIKM '16. (ACM, Indianapolis, Indiana), 1069–1078.

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-Exciting Point Process Modeling of Crime. *J. Am. Stat. Assoc.* 106(493):100–108.

Momot R, Belavina E, Girotra K (2017) The Use and Value of Social Information in Selective Selling of Exclusive Products. *Work. Pap.*

Moore R, Verity A (2014) *Hashtag Standards for Emergencies* (OCHA).

Myers SA, Leskovec J (2012) Clash of the Contagions: Cooperation and Competition in Information Diffusion. *2012 IEEE 12th Int. Conf. Data Min.* ICDM '12. (Brussels, Belgium), 539–548.

Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. *18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (ACM, Beijing, China), 33–41.

Ogata Y (1988) Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *J. Am. Stat. Assoc.* 83(401):9–27.

Oloruntoba R, Gray R (2006) Humanitarian aid: an agile supply chain? *Supply Chain Manag. Int. J.* 11(2):115–120.

Olteanu A, Castillo C, Diaz F, Vieweg S (2014) CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. *Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media.* ICWSM '14. (Ann Arbor, MI, USA).

Olteanu A, Vieweg S, Castillo C (2015) What to Expect When the Unexpected Happens: Social Media Communications Across Crises. *Proc. 18th ACM Conf. Comput. Support. Coop. Work Soc. Comput.* CSCW '15. (ACM, Vancouver, Canada), 994–1009.

Özdamar L, Ertem MA (2015) Models, solutions and enabling technologies in humanitarian logistics. *Eur. J. Oper. Res.* 244(1):55–65.

Pedraza Martinez AJ, Yan L (2016) Actionable Information for the Disaster Management Cycle through Social Media. *Work. Pap.*

Pettit S, Beresford A (2009) Critical success factors in the context of humanitarian aid supply chains. *Int. J. Phys. Distrib. Logist. Manag.* 39(6):450–468.

Poirier DJ (1980) Partial observability in bivariate probit models. *J. Econom.* 12(2):209–217.

Qu Y, Huang C, Zhang P, Zhang J (2011) Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. *Proc. ACM 2011 Conf. Comput. Support. Coop. Work.* CSCW '11. (ACM, New York, NY, USA), 25–34.

Rand W, Herrmann J, Schein B, Vodopivec N (2015) An Agent-Based Model of Urgent Diffusion in Social Media. *J. Artif. Soc. Soc. Simul.* 18(2):1–24.

Rand W, Rust RT (2011) Agent-based modeling in marketing: Guidelines for rigor. *Int. J. Res. Mark.* 28(3):181–193.

Reynaud-Bouret P, Schbath S (2010) Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Stat.* 38(5):2781–2822.

Ringel Morris M, Counts S, Roseway A, Hoff A, Schwarz J (2012) Tweeting is believing?: understanding microblog credibility perceptions. *Proc. ACM 2012Conference Comput.-Support. Coop. Work.* CSCW '12. (ACM, Seattle, WA, USA), 441–450.

Rizoiu MA, Lee Y, Mishra S, Xie L (2017) A Tutorial on Hawkes Processes for Events in Social Media. *ArXiv Prepr. ArXiv170806401.*

Rogers EM (1995) *Diffusion of Innovations* 4th ed. (Free Press, New York).

Rogers-Pettite C, Herrmann J (2015) Information Diffusion: A Study of Twitter During Large Scale Events. Cetinkaya S, Ryan JK, eds. *Proc. 2015 Ind. Syst. Eng. Res. Conf.* (Nashville, TN, USA), 1591–1600.

Ryzhov IO, Han B, Bradić J (2015) Cultivating Disaster Donors Using Data Analytics. *Manag. Sci.* 62(3):849–866.

Sarcevic A, Palen L, White J, Starbird K, Bagdouri M, Anderson K (2012) "Beacons of Hope" in Decentralized Coordination: Learning from On-the-ground Medical Twitterers During the 2010 Haiti Earthquake. *Proc. ACM 2012 Conf. Comput. Support. Coop. Work.* CSCW '12. (ACM, Seattle, WA, USA), 47–56.

Sarter NB, Woods DD (1991) Situation Awareness: A Critical But Ill-Defined Phenomenon. *Int. J. Aviat. Psychol.* 1(1):45–57.

Shi Z, Rui H, Whinston AB (2014) Content Sharing in a Social Broadcasting Environment: Evidence from Twitter. *MIS Q.* 38(1):123–142.

Simon H (1971) Designing Organizations for an Information-Rich World. Greenberger M, ed. *Comput. Commun. Public Interest*. (Johns Hopkins Press, Baltimore, MD), 37–72.

Singh J (2005) Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Manag. Sci.* 51(5):756–770.

Sodhi MS (2016) Natural disasters, the economy and population vulnerability as a vicious cycle with exogenous hazards. *J. Oper. Manag.* 45:101–113.

Sodhi MS, Tang CS (2014) Buttressing Supply Chains against Floods in Asia for Humanitarian Relief and Economic Recovery. *Prod. Oper. Manag.* 23(6):938–950.

Starbird K, Palen L (2010) Pass it on?: Retweeting in mass emergency. *Proc. 7th Int. ISCRAM Conf.* ISCRAM '10. (Seattle, WA, USA), 1–10.

Starbird K, Palen L, Hughes AL, Vieweg S (2010) Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information. *Proc. 2010 ACM Conf. Comput. Support. Coop. Work*. CSCW '10. (ACM, Savannah, GA, USA), 241–250.

Stieglitz S, Dang-Xuan L (2013) Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *J. Manag. Inf. Syst.* 29(4):217–248.

Stonedahl F, Rand W, Wilensky U (2010) Evolving viral marketing strategies. *Proc. 12th Annu. Conf. Genet. Evol. Comput.* GECCO '10. (ACM, Portland, OR, USA), 1195–1202.

Stonedahl F, Wilensky U (2010a) *BehaviorSearch [computer software]* (Center for Connected Learning and Computer Based Modeling, Northwestern University, Evanston, IL. http://www.behaviorsearch.org.).

Stonedahl F, Wilensky U (2010b) Finding Forms of Flocking: Evolutionary Search in ABM Parameter-Spaces. Bosse T, Geller A, Jonker CM, eds. *Multi-Agent-Based Simul. XI*. Lecture Notes in Computer Science. (Springer Berlin Heidelberg), 61–75.

Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *Proc. 2010 IEEE Second Int. Conf. Soc. Comput.* SOCIALCOM '10. (Minneapolis, MN, USA), 177–184.

Susarla A, Oh JH, Tan Y (2011) Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. *Inf. Syst. Res.* 23(1):23–41.

Sutton J, Palen L, Shklovski I (2008) Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires. *Proc. 5th Int. ISCRAM Conf.* –. ISCRAM '08. (Washington, DC, USA).

Swaminathan JM (2018) Big Data Analytics for Rapid, Impactful, Sustained, and Efficient (RISE) Humanitarian Operations. *Prod. Oper. Manag*. Forthcoming.

Symmes Cobb J, Ore (2016) Death toll from Ecuador earthquake surpasses 650. *Reuters*. Retrieved http://www.reuters.com/article/us-ecuador-quake-idUSKCN0XK0GQ.

Tang Q, Gu B, Whinston AB (2012) Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model. *J. Manag. Inf. Syst.* 29(2):41–76.

Tomasini RM, Van Wassenhove LN (2005) *Managing Information in Humanitarian Crisis: UNJLC Website* (INSEAD, France).

Tomasini RM, Van Wassenhove LN (2009) From preparedness to partnerships: case study research on humanitarian logistics. *Int. Trans. Oper. Res.* 16(5):549–559.

Van Wassenhove LN (2006) Humanitarian Aid Logistics: Supply Chain Management in High Gear. *J. Oper. Res. Soc.* 57(5):475–489.

Van Wassenhove LN, Pedraza Martinez AJ (2012) Using OR to adapt supply chain management best practices to humanitarian logistics. *Int. Trans. Oper. Res.* 19(1–2):307–322.

Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.* CHI '10. (ACM, Atlanta, GA, USA), 1079–1088.

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151.

Wang Y, Agichtein E, Benzi M (2012) TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* KDD '12. (ACM, Beijing, China), 123–131.

Watts DJ (2002) A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci.* 99(9):5766–5771.

Watts DJ, Dodds PS (2007) Influentials, Networks, and Public Opinion Formation. *J. Consum. Res.* 34(4):441–458.

Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. *Sci. Rep.* 2.

Wilensky U (1999) NetLogo. Retrieved https://ccl.northwestern.edu/netlogo/.

Wilensky U, Rand W (2015) *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo* (The MIT Press, Cambridge, MA).

Wu F, Huberman BA (2007) Novelty and collective attention. *Proc. Natl. Acad. Sci.* 104(45):17599–17601.

Xu B, Huang Y, Kwak H, Contractor N (2013) Structures of Broken Ties: Exploring Unfollow Behavior on Twitter. *Proc. 2013 Conf. Comput. Support. Coop. Work*. CSCW '13. (ACM, New York, NY, USA), 871–876.

Xu L, Duan JA, Whinston A (2014) Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion. *Manag. Sci.* 60(6):1392–1412.

Yang J, Leskovec J (2011) Patterns of Temporal Variation in Online Media. *Proc. Fourth ACM Int. Conf. Web Search Data Min.* WSDM '11. (ACM, Hong Kong, China), 177–186.

Yang Z, Guo J, Cai K, Tang J, Li J, Zhang L, Su Z (2010) Understanding Retweeting Behaviors in Social Networks. *Proc. 19th ACM Int. Conf. Inf. Knowl. Manag.* CIKM '10. (ACM, Toroto, ON, Canada), 1633–1636.

Yardi S, Romero D, Schoenebeck G, Boyd D (2009) Detecting spam in a Twitter network. *First Monday* 15(1).

Yoo E, Rand W, Eftekhar M, Rabinovich E (2016) Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. *J. Oper. Manag.* 45:123–133.

Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* KDD '15. (ACM, Sydney, Australia), 1513–1522.

Zhou K, Zha H, Song L (2013) Learning Triggering Kernels for Multi-dimensional Hawkes Processes. *Proc. 30th Int. Conf. Mach. Learn.* ICML '13. (Atlanta, GA, USA), 1301–1309.

APPENDIX A

DESCRIPTION OF THE AGENT-BASED MODEL

This appendix describes the agent-based model (ABM) used in this paper, which is based on Rand et al. (2015). This model and this documentation were created using the guidelines for building ABMs recommended in Rand and Rust (2011). We begin by explaining why ABM is appropriate for the present application. We then describe the model along with the two major variants discussed in the paper: Independent Cascade (IC) and Linear Threshold (LT). Next, we discuss the verification and validation of the model. Finally, we describe the parameter optimization approach that we used and present the pseudocode for the underlying model, as well as the parameters for the ABM used in the search and the parameters for the search itself.

I. Appropriateness of ABM

Rand and Rust (2011) lay out six conditions for determining whether or not ABM is appropriate for a given problem. As they describe, the more of these conditions are met, the more useful ABM will be. The conditions are:

1. A Medium Number of Agents – Is there a medium number of agents as opposed to a very small or a very large number of agents? In this case, we are investigating a medium number of agents since we are not interested in how one or two agents process and share information during a disaster nor are we interested in billions of agents. Rather, we seek to model how, at most, around three thousand individuals on Twitter find and distribute information in order to form a cascade.
2. Local and Potentially Complex Interactions among Agents – Do the agents interact only among their local neighborhood and potentially maintain memories about those interactions? In our case of information diffusion on Twitter during a disaster, both of these conditions are met. The model as proposed has the agents mainly paying attention to their local neighborhoods for information. Moreover, the agents do not just directly respond to each piece of information but, rather, judge based on the IC and LT rules if they should adopt the information.

3. Agents' Heterogeneity – Are the agents different from each other in substantial ways? In the system we are examining, agents have one important source of heterogeneity, which is that they differ substantially based on their exact location within the overall social network. This is an important source of heterogeneity, and, in fact, it is this network position that creates the diffusion dynamics that we observe.
4. Rich Environments – Does the environment enable a rich set of interactions? The environment of information diffusion on Twitter in a crisis is defined by the agents' social connections. As the social connections are quite varied between individuals, the environment can be considered rich. Moreover, the network is not just defined by user connections but also by the agents at the other end of those links, which further enhances the environment's richness.
5. Temporal Aspects – Is the phenomenon of interest something that evolves over time or is it static? In this case, we are interested in how quickly information spreads through the network, so this requirement is clearly met.
6. Agents' Adaptability – Do agents change their actions based on previous experience? It should be noted that Rand and Rust (2011) denote that this is not a common element of ABM. In our model, agents are not adaptive, though this could be explored in future research.

Because these six conditions were generally fulfilled, ABM is clearly an appropriate methodology for understanding the phenomenon at hand.

II. Model Construction

We constructed this model in the popular ABM language NetLogo (Wilensky 1999). There are seven design choices that we needed to consider for the model (Rand and Rust 2011):

1. Scope – The scope of our model is a local Twitter network of communication in the context of a particular tweet and its retweets. We do not seek to replicate the whole Twitter network or to study effects beyond simple information diffusion processes.
2. Agents – There is essentially only one type of agent in the model. This agent is an information diffuser on Twitter.
3. Properties – Agents have four properties: [i] probability of external influence (p); [ii] parameter of internal influence (q); [iii] whether or not they have adopted the product (*adopt?*); and [iv] their local social network. Both p and q are set exogenously by the optimization algorithm, and they are not modified during the runs. *adopt?* is initially set to FALSE for all agents and then updated based on either the IC or LT adoption rule. Finally, agents' social networks include the links among agents. These are drawn from an empirical network of the largest cascade in our data. If the cascade we are examining is smaller than the largest cascade, then we trim the network by eliminating any nodes (and their accompanying links) that would expand the network beyond the size of the current cascade.
4. Behaviors – Agents in this model have essentially one behavior: decide whether or not to adopt new information. We examine two forms of this behavior governed by either the LT model (Granovetter 1978) or the IC model (Goldenberg et al. 2001).
    a. The LT Model
        i. External Influence - Agents first decide whether to adopt based on external influence. To do this, they draw a random number, x, from the uniform distribution of [0,1). If $x < p$, they adopt that information. Agents keep their state hidden until the end of the turn. Thus, if an agent adopts during this phase of the model, it is still counted as *not* having adopted during the internal influence stage. This is known as synchronous updating and is standard practice (Wilensky and Rand 2015).

ii. Internal Influence - Each agent then counts up the number of neighbors that have adopted the information, *n_adopt*, and divides by the total number of neighbors, *n*. They then compare this number to $\phi = 1-q$, and if $(n\_adopt / n) > \phi$, they adopt the information. It should be noted that this is a directed network based on the following / follower relationship in Twitter, so users only consider their neighbors to be those people they are following, not the neighbors that are following them. Moreover, agents do not reveal again if they have adopted during this turn, so if a neighbor has just adopted, it is counted as not having adopted during this time step.

b. The IC Model

i. External Influence – Agents first decide whether to adopt based on external influence. To do this, they draw a random number, x, from a uniform distribution [0,1). If $x < p$, they adopt that information. Agents hide their state until the end of the turn.

ii. Internal Influence – Each agent who adopted the information in the most recent time step (a record is kept of which time step the agent adopted in to facilitate this) transmits the information to all of its neighbors who have not adopted via the "following" relationship, i.e., neighbors who are following the focal user. These uninformed agents draw a random number, x, from the distribution of [0,1), and if $x < q$, then they adopt the information. Agents who just adopted in this time step or who adopted more than one time step before do not influence adoption[13].

[13] A note of clarification: The LT model uses ϕ, while the IC model uses q directly. Since ϕ is a threshold that must be exceeded before diffusion occurs in the LT model, a lower value of ϕ indicates a higher level of internal influence, while a higher level of ϕ indicates a lower level of internal influence. In the IC model, q is a probability of internal diffusion: a high q value indicates a high rate of diffusion whereas a low q value indicates a low rate of diffusion. To compare ϕ and q, therefore, we measure internal influence in the LT model using $q = 1- \phi$.

5. Environment – The main environment of the model is defined by the empirically grounded Twitter network of the largest cascade, which consisted of 3,315 users. The network was trimmed when appropriate to fit smaller cascades.
6. Input and Output – Three parameters control the basic model, and the results are examined through one output variable. The three parameters are: (a) p, (b) q, and (c) the cascade number to examine. Both p and q are set homogenously for all agents in the network. The cascade number loads in the appropriate network structure by trimming the network of 3,315 users to the size of the current cascade. It also loads in the actual time series of retweets / adoptions in the empirical data at one minute resolutions, i.e., the cumulative new retweets at each minute. This time series is called *Empirical(t)*. Once all the data is loaded, the model is run until all nodes have adopted. A time series, *Y(t)*, is recorded, which corresponds to the cumulative number of adoptions in each time step. The output is a Mean Absolute Percentage Error (MAPE), described in Section 3.3.1 of the paper. *Y(t)* may be longer or shorter than *Empirical(t)*. If *Y(t)* is shorter, then it is padded with 0's to reach the same length. If it is longer, it is trimmed from the end to reach the same length.
7. Time Step – Almost all ABMs have two phases: an initialization phase and an iterative phase. In our model's initialization phase, agents are created and given their initial properties (p and q) to then be embedded in the social network. In the iterative step, agents decide whether to adopt according to the behaviors in (4). In the first time step, no one has adopted, so only external influence affects adoption. After this, all statistics, *Y(t)*, are recorded.

III. Verification of the Model

There are three standards in place to ensure that our implemented model corresponds to the conceptual model as described, i.e., the process of verification (Rand and Rust 2011).

1. Documentation – The model was well-documented both within the code and within lab notes. This documentation and the code will be published on OpenABM.org, a repository that maintains such information. This appendix serves as another source of documentation.
2. Programmatic testing – To examine the model, we used a combination of unit testing and code walk-throughs. In unit testing, as each additional level of complexity was added to the model, we ran the model to see if prior results could still be created. Then, a code walk-through was carried out as a coauthor reviewed the program with another coauthor.
3. Test Cases – Corner cases and sampled cases were examined to see if the model was creating any aberrant behavior.

IV. Validation of the Model

Validation involves the comparison of the implemented model to the real world in some meaningful way. Rand and Rust (2011) describe four standards for validating a model: (1) Micro-face validation, (2) Macro-face validation, (3) Empirical input validation, and (4) Empirical output validation. Most of our model's validation is documented in the main body of the paper.

Micro-face validation involves determining that the agents at the micro-level behave the way real agents do. The IC and LT models are drawn from literature that claims they are reasonable models of actual behavior at the individual level (Goldenberg et al. 2001, Granovetter 1978). Macro-face validation involves determining whether the processes at the macro-level reflect real-world macro-processes. Given that our model

shows the standard s-shaped diffusion curves found in many empirical settings (Rogers 1995), the model is valid from a macro-face perspective.

Empirical input validation and empirical output validation relate to comparing the model's input and outputs to real data. For empirical input validation, we used an empirically derived network from the actual Twitter following network. This is an accurate representation for the largest cascade in our network. Due to computational constraints it was not feasible to pull down the networks of all cascades, but by using a trimmed network version, we can represent the same topological constraints and properties observed by the Twitter network in general. As to p and q, we constrained these values to ranges that have been empirically observed in similar diffusion models (Chandrasekaran and Tellis 2007). It should be noted that in our context, empirical input validation is tied to empirical output validation. Therefore, we searched over the space of all reasonable input parameters to find parameters that produced empirical output data, which is explained in the next section. Thus, our model also has the best possible fit given the computational power expended to the empirical data.

V. Parameter Optimization

To identify the parameters for the ABM that best created output patterns matching the real data, we used a method known as parameter optimization. Through this method, we identified a set of input parameters and an output measure, often called a fitness function, and then applied an optimization procedure to select the best possible parameters to minimize or maximize the fitness function (Miller 1998, Stonedahl et al. 2010). In our context, we identified the parameters p and q that offered the best match between the output of the model and the empirical adoption patterns that we observed. As our fitness function, we chose to minimize the MAPE between our model data and the empirical data in line with previous work (Rand et al. 2015).

To robustly test each model, we needed to examine all 5,434 cascades in our data with multiple runs per model setting due to the stochastic nature of the models. This precluded a full sweep of the parameter space. Consequently, we turned to machine learning methods to intelligently search the parameter space. We used BehaviorSearch, which is an add-on to NetLogo that carries out parameter optimization automatically on NetLogo models (Stonedahl and Wilensky 2010a). BehaviorSearch provides three standard parameter optimization methods: simulated annealing, genetic algorithms, and mutation hill climbing (Stonedahl and Wilensky 2010b). During robustness checks on a smaller number of cascades, we found that using the simulated annealing (SA) algorithm provided quick convergence while at the same time identifying parameters with a low overall error compared to the other two methods. For these reasons, we utilized the SA approach for optimization of the parameters and results presented in this paper. We note that for all of our searches, we restricted the search space to previously empirically observed values for p and q in similar models (p range= [0.0007, 0.03], q range= [0.38, 0.53]) (Chandrasekaran and Tellis 2007).

For each cascade, BehaviorSearch carried out the SA algorithm with 150 evaluations, i.e., 150 different p and q values. For each of these values, BehaviorSearch executed the model ten times and then averaged the results since the model runs are stochastic. This yielded an average idea of the underlying fitness. Anytime we encountered a best solution, we re-ran the model 25 times to determine a more precise value for that solution. We then executed the overall SA algorithm three times to make sure that we had the best possible fit, and we kept the parameter values that gave us the lowest MAPEs overall. Thus, for each cascade we evaluated up to 450 p and q values with at least 10 runs per value for a total of 4,500 runs per cascade. It should be noted that since it is possible to generate through the SA algorithm the same value twice, we enabled caching of fitness values so that if we returned to the same value, we did not re-run the model. We repeated

this process for every cascade and for both models (LT and IC), resulting in 48,906,000 runs at most[14] (not including the additional runs for checking best results).

VI. Pseudo-code of Models

In this section, we describe each of the model variants using pseudo-code, which is a natural language version of the code used to create the models. The full code of the model as well as documentation will be available from OpenABM.org.

```
Base Model
to setup
        read in Empirical(t) from data
        read in Network from data
        trim Network  so that the number of nodes in Empirical(t) is equivalent to the number of
        nodes in the Network
        set p and q for all nodes
        set adopted? to false for all nodes
        set adopt-time -1
end
to go
        for all agents that have not adopted
                x = U[0,1)
                if x < p then adopt
                if model = threshold then
                        n = inbound neighbors
                        n_adopt = inbound neighbors with [adopted? = true]
                        ϕ = 1 - q
                        if n_adopt / n > ϕ then adopt
                if model = cascade then
                        for all agents with adopt-time = current-time – 1
                                for all outbound neighbors with [ adopted? = false ]
                                        x = U[0,1)
                                        if x < q then adopt
        Y(current_time) = count agents with [ adopted? = true ]
        MAPE = calc_MAPE(Y(t), Empirical(t))
end
to adopt
 set adopted? true
 set adopt-time current-time
end
```

[14] Since ten runs provided a reasonable estimation of the MAPE, we allowed the SA to reuse MAPE values for points in the parameter space that it revisited for a given network. This saved time in the runs and did not alter significantly the results. The calculation on the runs corresponded to an upper limit since points that were revisited were not rerun. For the runs, we used three different machines in parallel with 68 cores between all of them. It took about 38 hours to carry out all of the model runs. Based on the number of each machine's cores and the time it took to execute the runs per machine, it would have taken over 99 days to run all of these model iterations on one CPU.

VII. Parameters of Paper

The table below details the exact parameters used to create and run the models.

Base ABM parameters		*Behavior Search Parameters*		*Simulated Annealing Parameters*	
Parameter	*Value*	*Parameter*	*Value*	*Parameter*	*Value*
Diffusion Model	Linear Threshold or Independent Cascade	Number of Searches per Cascade	3	Mutation Rate	0.5
p	[0.0007,0.03]	Fitness Caching	True	Temperature Change Factor	0.99
q/ϕ	[0.38, 0.53]	Fitness Function	Minimizing MAPE	Initial Temperature	1.0
cascade_num	[0, 5433]	Function for Combining Replications	Mean		
Number of Agents	size of cascade	Model Runs per Parameter Setting	10		
Length of Run	2000 time steps or full adoption				

APPENDIX B

INTEGRATION OF THE CONDITIONAL INTENSITY FUNCTION

In this appendix, we provide the integration of the initial term in the penalized log-likelihood function expressed in Equation 10. This term can be expanded into the following system of equations:

$$\int_{t_0^i}^{\boldsymbol{T}} \tilde{\lambda}^i\left(t^i \middle| \mathcal{H}_t^i\right) dt = \int_{t_0^i}^{\boldsymbol{T}} \max\left(\lambda^i\left(t^i \middle| \mathcal{H}_t^i\right), 0\right) = \begin{cases} \int_{t_0^i}^{\boldsymbol{T}} \lambda^i\left(t^i \middle| \mathcal{H}_t^i\right) dt & \text{when } \lambda^i\left(t^i \middle| \mathcal{H}_t^i\right) \geq 0 \\ \int_{t_0^i}^{\boldsymbol{T}} 0 dt & \text{when } \lambda^i\left(t^i \middle| \mathcal{H}_t^i\right) < 0 \end{cases}$$

We now present the analytical integration of the first equation within the system of equations. To simplify some of the integrals, we set t_0^i equal to 0. As such, we evaluated the integrals over $[0, \boldsymbol{T} - t_0^i]$. This change did not affect the model or the interpretation of the results.

$$\int_0^{\boldsymbol{T}-t_0^i} \lambda^i\left(t^i \middle| \mathcal{H}_t^i\right) dt$$

$$= \int_0^{\boldsymbol{T}-t_0^i} \mu^i e^{-\gamma^i t} + \sum_{t_k^i < t} \left(\alpha_{11}^i * \phi_{t_k^i} * e^{-\beta_{11}^i (t - t_k^i)}\right)$$

$$+ \sum_{t_l^i < t} \left(\alpha_{21}^i * \phi_{t_l^i} * e^{-\beta_{21}^i (t - t_l^i)}\right) dt$$

$$= -\frac{\mu_n^i}{\gamma_n}\left(e^{-\gamma(\boldsymbol{T}-t_0^i)} - 1\right) + \sum_{t_k^i < t} \left[\frac{\alpha_{11}^i * \phi_{t_k^i}}{\beta_{11}^i} * \left(1 - e^{-\beta_{11}^i\left((\boldsymbol{T}-t_0^i)-t_k^i\right)}\right)\right]$$

$$+ \sum_{t_l^i < t} \left[\frac{\alpha_{21}^i * \phi_{t_l^i}}{\beta_{21}^i} * \left(1 - e^{-\beta_{21}^i\left((\boldsymbol{T}-t_0^i)-t_l^i\right)}\right)\right]$$

APPENDIX C

DESCRIPTIVE STATISTICS OF PARAMETER ESTIMATES BY DISASTER

Joplin tornado

	Mean	Median	Std. Dev.	Min.	Max.
α_{11}^i	0.057	0.001	0.157	7.27E-14	2.264
β_{11}^i	0.883	0.826	0.774	2.62E-05	10.615
α_{21}^i	-0.133	-0.018	0.299	-4.45E+00	2.006
β_{21}^i	0.576	0.418	0.584	2.62E-05	6.518
μ^i	0.577	0.569	0.490	1.74E-06	8.128
γ^i	0.373	0.360	0.291	2.62E-05	2.886

9,849 observations

Black Forest fire

	Mean	Median	Std. Dev.	Min.	Max.
α_{11}^i	0.074	0.001	0.175	1.12E-12	1.901
β_{11}^i	0.925	0.901	0.660	1.31E-04	4.726
α_{21}^i	-0.109	-0.012	0.295	-3.05E+00	1.120
β_{21}^i	0.654	0.499	0.629	1.31E-04	3.527
μ^i	0.605	0.615	0.438	6.08E-05	3.923
γ^i	0.410	0.408	0.290	1.31E-04	2.983

2,280 observations

Lac-Megantic rail disaster

	Mean	Median	Std. Dev.	Min.	Max.
α_{11}^i	0.066	0.001	0.156	4.11E-12	1.390
β_{11}^i	0.925	0.856	0.783	2.00E-06	5.458
α_{21}^i	-0.121	-0.014	0.284	-2.17E+00	1.090
β_{21}^i	0.594	0.403	0.633	1.65E-06	5.245
μ^i	0.631	0.645	0.478	3.78E-05	4.999
γ^i	0.370	0.347	0.297	7.83E-07	2.309

1,947 observations

2014 Iquique earthquake

	Mean	Median	Std. Dev.	Min.	Max.
α_{11}^i	0.101	0.001	0.256	6.66E-13	3.827
β_{11}^i	0.961	0.739	1.181	5.36E-05	18.632
α_{21}^i	-0.142	-0.005	0.345	-4.59E+00	3.356
β_{21}^i	0.549	0.260	0.683	1.79E-07	9.882
μ^i	0.562	0.558	0.529	5.18E-06	7.253
γ^i	0.332	0.300	0.304	1.57E-05	3.995

12,762 observations

APPENDIX D

DETERMINING NEW FOLLOWERS AS INTERNAL OR EXTERNAL LINKS

In Section 5, we explained how we used the Gnip data and the scraped follower lists to identify every supplier's set of new followers. The new followers were the users within the index [**n-e**+1**, n-b**] on the suppliers' follower lists in reverse chronological order. Next, we determined if the new followers were internal or external links. This process required knowledge on when new followers started following the suppliers, but unfortunately, such data are not available. For each supplier, we estimated the times that users became new followers by again relying on Gnip's records of the supplier's follower count at the time that the supplier tweeted or was retweeted. For each record, we noted the time as τ and the follower count of the supplier as $\mathbf{r}_{\tau}$, and we also located the immediately preceding record and logged its time as τ-*1* and the associated follower count as $\mathbf{r}_{\tau-1}$. We then estimated that the users within the index [$\mathbf{n-r}_{\tau-1}$, $\mathbf{r}_{\tau}-\mathbf{r}_{\tau-1}+1$] on the supplier's scraped follower list started following the supplier at τ. We performed this analysis for every record of the suppliers' follower counts to approximate the following times of new followers. While this method is not exact, it is highly precise since the supplier's follower counts were logged frequently due to the large amount of activity by suppliers and retweeters, especially after the earthquake.

The internal mechanism requires that candidates start following a supplier after being exposed to the supplier via a retweet of the supplier's content. Because we already knew who the new followers were, we worked backwards to verify if they had formerly been candidates using the following method. We conducted this process for every supplier and for each of the supplier's new followers (i.e., *nf*):

1. Check if *nf* was following any of the supplier's retweeters. We accomplished this by matching the *nf*'s Twitter ID in the scraped list of followers for every one of the supplier's retweeters. Multiple matches meant that *nf* was following multiple retweeters of the supplier.

2. For each match, make sure that *nf* was following the retweeter prior to the time that *nf* started to follow the supplier (or "*s-follow time*" for brevity). We relied on the same method that we adopted to ascertain the following times of suppliers' followers for retweeters' followers as well. The information for this process came from the follower counts for retweeters logged in the Gnip data set. We then removed from consideration any retweeters for whom this condition did not hold since such retweeters could not have distributed any of the supplier's tweets to *nf* before *s-follow time*.
3. For the remaining matched retweeters, pull all of their retweets of the supplier. Retain only the retweets that occurred after the time that *nf* started following the retweeter and before *s-follow time*. This guarantees that the retained retweets were sent as valid exposures of the supplier to *nf* by the retweeters.
4. Sort the retained retweets from most recent to oldest. Assign the *nf* as an internal link, and in line with Antoniades and Dovrolis (2015), assign the most recent retweet as the exposure that motivated *nf* to follow the supplier.

If this method failed at any point for a new follower, this implied that we could not trace the user to the diffusion path of a supplier's tweet, so we categorized that user as an external link. In other words, we were unable to match the new follower as a legitimate candidate of any of the supplier's retweeters. We note that the final step rests on the assumption that new followers actually consumed, or read, the most recent retweet. We argue that our assumption is valid for several reasons, the first being that new followers appear to be active since we observed their decision to start following a supplier (which implies that these users logged into Twitter). This raises the likelihood that new followers saw the retweet. Furthermore, nearly all of the new followers identified as internal links (89.6%) started following the supplier within 24 hours of the retweet they were assigned to. We anticipate little delay between a candidate consuming a retweet

and following a supplier. Even so, this means that the content is relatively new and should be near the top of the candidates' Twitter feeds, again increasing the likelihood that the assigned retweet was read.

APPENDIX E

TEXT CLASSIFICATION

To analyze the type of content presented in the suppliers' tweets, we categorized the tweets as belonging to one of the following three categories: (1) Actionable; (2) Informative; or (3) Other. In total, the suppliers published 15,399 tweets during the weeks before and after the earthquake. To classify the text in these tweets, we adopted a supervised learning approach, which involves training a classifier based on a labeled training data set (Manning and Schütze 1999).

First, we preprocessed the tweets according to the following standard natural language processing procedures:

1. All text was converted to lowercase.
2. Any punctuation and emojis were removed.
3. All links and hashtags outside of those used to query the data for this study were retained.
4. Tweets were tokenized, or split up into tokens that consisted of one word each.

To maintain the consistency of our data and improve the classifier's accuracy, we removed 215 tweets that were not written in Spanish and 367 tweets that contained less than five words. We coded these tweets manually.

Next, we randomly selected 1,500 tweets and manually coded each tweet as Actionable, Informative, or Other. We divided the manually classified tweets into a training data set (80%) and test data set (20%). Using the training data set, we extracted several features to develop the classifier. First, we calculated the term frequency-inverse document frequency (TF-IDF) scores, which measure the frequency of a token (i.e., a word) in a tweet while also accounting for how common the term is across all of the tweets (Manning and Schütze 1999). We also added as features part-of-speech tags that were obtained using the Spanish module of the Stanford POS Tagger. Lastly, we included the supplier's Twitter handle as another feature since suppliers may tend to post certain types of content.

We applied the Naïve Bayes and Support Vector Machine (SVM) algorithms to the training data set. Using 10-fold cross-validation on the training data, we found that the classification accuracy for Naïve Bayes was 70.6% and for SVM was 74.3%. Thus, we primarily relied on SVM for text classification in our study. We trained the classifier using the features described above with and without stop words, which are commonly used terms (e.g., "the", "and"). Furthermore, we used the grid search approach to tune the SVM parameters. The trained classifier was then applied to the test data set, and accuracy was measured as the percentage of tweets that were categorized correctly by the classifier. The most accurate SVM classifier (86%) utilized all of the features in conjunction with stop words. Since we were able to achieve high accuracy, we applied the trained classifier on the remaining tweets in our data.

APPENDIX F

RESULTS FROM ROBUSTNESS CHECKS

	Robustness Check #1		Robustness Check #2		Robustness Check #3		Robustness Check #4	
	Coeff.	**(Robust Std. Err.)**	**Coeff.**	**(Robust Std. Err.)**	**Coeff.**	**(Robust Std. Err.)**	**Coeff.**	**(Robust Std. Err.)**
				Stage 1: Consumption				
β_0 (Intercept)	-2.056***	(-0.100)	-2.331***	(0.047)	-0.577*	(0.358)	-2.160***	(0.092)
β_1 (b_{sti})	-0.241***	(0.012)	-0.119***	(0.008)	-0.408***	(0.036)	-0.233***	(0.011)
β_2 (δ_{sti})							0.009***	(0.001)
				Stage 2: Follow Decision				
γ_0 (Intercept)	-4.784***	(0.030)	-4.718***	(-0.039)	-4.833***	(0.047)	-4.858***	(0.030)
$\gamma_{1_{action}}$ (a_{sti} = Actionable)	0.023***	(0.003)	0.038***	(0.004)	0.030***	(0.003)	0.030***	(0.003)
$\gamma_{1_{other}}$ (a_{sti} = Other)	0.0173***	(0.006)	0.056***	(0.007)	0.055***	(0.006)	0.065***	(0.006)
γ_2 ($\log p_{sti}$)	0.043***	(0.002)	0.017***	(0.001)	0.088***	(0.002)	0.043***	(0.002)
γ_3 ($\log q_{sti}$)	0.131***	(0.012)	0.052***	(0.004)	0.139***	(0.010)	0.134***	(0.013)
γ_4 ($\log r_{sti}$)	0.011***	(0.002)	0.011***	(0.002)	0.008**	(0.002)	0.015***	(0.002)
γ_5 ($\log g_{sti}$)	-0.093***	(0.004)	-0.065***	(0.005)	-0.083***	(0.005)	-0.092***	(0.004)
γ_6 ($\log(f_{sti} - z_{sti})$)	0.142***	(0.004)	0.199***	(0.006)	0.161***	(0.004)	0.150***	(0.003)
γ_7 (w_{sti})	0.016***	(0.002)	0.017***	(0.002)	0.014***	(0.001)	0.017***	(0.002)
γ_8 (d_{sti})	0.335***	(0.007)	0.339***	(0.012)	0.327***	(0.008)	0.316***	(0.007)
γ_9 ($w_{sti} * d_{sti}$)	-0.017***	(0.002)	-0.020***	(0.002)	-0.016***	(0.002)	-0.019***	(0.002)
γ_{10} ($\log(e^{f_{sti}} - e^{z_{sti}})$)	-0.004***	(2E-04)	-0.005***	(2E-04)	-0.004***	(2E-04)	-0.004***	(2E-04)
γ_{11} ($\log(e^{f_{sti}} - e^{z_{sti}}) * d_{sti}$)	0.002***	(2E-04)	0.003***	(2E-04)	0.003***	(2E-04)	0.002***	(2E-04)
γ_{12} (v_{sti})			11.121***	(0.087)				
γ_{13} (ϕ_{sti})					-0.255***	(0.009)		
rho	0.953***	(0.006)	0.660***	(0.027)	0.759***	(0.042)	0.956***	0.005
Observations	352,288		371,420		371,420		371,420	
Pseudo log-likelihood	-368.913		-371.059		-382.798		-383.250	

** p<0.1, ** p<0.05, *** p<0.01*