International Journal of Research in Marketing 34 (2017) 100–119

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Modeling the role of message content and influencers in social media rebroadcasting

Yuchi Zhang [a,*], Wendy W. Moe [b], David A. Schweidel [c]

[a] Leavey School of Business, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053, United States
[b] Robert H. Smith School of Business, University of Maryland, 7621 Mowatt Ln, College Park, MD 20740, United States
[c] Goizueta Business School, Emory University, 1300 Clifton Rd NE, Atlanta, GA 30322, United States

ARTICLE INFO

Article history:
First received on September 3, 2014 and was under review for 7 months
Available online 30 August 2016

Guest Editor: Steven M. Shugan

Keywords:
Social media
Social influence
Text mining
Twitter
Bayesian estimation

ABSTRACT

We develop a model that examines the role of content, content-user fit, and influence on social media rebroadcasting behavior. While previous research has studied the role of content or the role of influence in the spread of social media content separately, none has simultaneously examined both in an effort to assess the relative effects of each. Our modeling approach also accounts for a message's "fit" with users, based on the content of the message and the content of messages typically shared by users.

As an empirical application, we examine how Twitter posts originating from top business schools are subsequently rebroadcasted (or retweeted) by other users. We employ an individual-level split hazard model that accounts for variation in rebroadcasting decisions related to (1) content, (2) the content-user fit and (3) the influence of other users. We find that the rebroadcasting a message depends not only on message content but also on the message's fit with a user. Our analysis also yields measures of influence and susceptibility to influence for each user, which can be used to identify influential social media users. We demonstrate how our approach can be used to evaluate different types of seeding strategies designed to increase the reach of social media messages.

1. Introduction

In 2013, consumers posted 139 billion messages to Twitter, a popular social media messaging platform. Of these messages, about 24% were rebroadcasted (i.e., retweeted) by other Twitter users (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013). Rebroadcasting, or "sharing" in other settings, is an extremely important activity for marketers, as they attempt to disseminate marketing communications about their brands on social media. Recently, there has also been great interest in studying rebroadcasting behavior, specifically in regards to who is more likely to retweet messages on Twitter (Lambrecht, Tucker, & Wiertz, 2015; Suh, Hong, Pirolli, & Chi, 2010; Zaman, Fox, & Bradlow, 2014). By understanding the drivers of rebroadcasting activities, managers could more effectively control and disseminate communications on social media. That is, by generating greater rebroadcasting activity, managers can reach potential consumers who are not actively following the content posted by the firm. However, beyond identifying characteristics of individuals who are more likely to rebroadcast, little is known about why those 24% of tweets are subsequently rebroadcasted. Nor is it known what the factors were that affected users' decisions to rebroadcast these messages while the remaining 76% were not rebroadcast.

* Corresponding author.
E-mail addresses: yzhang6@scu.edu (Y. Zhang), wmoe@rhsmith.umd.edu (W.W. Moe), dschweidel@emory.edu (D.A. Schweidel).

http://dx.doi.org/10.1016/j.ijresmar.2016.07.003
0167-8116/

One potential explanation is that the message content was key to users' rebroadcasting decisions. A number of researchers have examined how message content affects the spread of word-of-mouth (Berger, 2011; Berger & Milkman, 2012; Berger & Schwartz, 2011; Heath, Bell, & Sternberg, 2001). They find that certain types of content are more likely to be rebroadcast than others.

Alternatively, influential users could have played a role in affecting the rebroadcasting decisions of other users. Researchers have studied the role of influence in both offline and online contexts (e.g., Watts & Dodds, 2007). As more data become available through online platforms, such work has also focused on measuring the effects of influence as a function of a user's ability to influence the behavior of others and a user's susceptibility to social influence (Aral & Walker, 2012; Trusov, Bodapati, & Bucklin, 2010). If firms can entice rebroadcasting by influential users, others may subsequently follow and amplify the reach of the original marketing communication.

While the above-mentioned streams of research have contributed significantly to our understanding of social media behavior, these streams have been developing in parallel with very little integration. Our goal in this research is to propose a model of social media rebroadcasting behavior that integrates the various factors shown to influence rebroadcasting behavior. That is, can we model both the role of message content and influence simultaneously, allowing us to assess the relative impact of each? Furthermore, an integrated model allows us to investigate the impact of content-user fit, a measure that considers the interaction between the message content and user preferences. The modeling integration of content, influence, and content-user fit allows us to potentially uncover scenarios where disseminating the most viral content, on average, may not generate the most rebroadcasting activity for the firm due to the unique preferences of the audience and influentials who follow the firm's messages. Thus, this research provides managers with a method to properly determine a social media rebroadcasting strategy that is specific to their context and audience base.

To represent rebroadcasting activity over time, we employ a split hazard model to model an individual's decision of whether or not to rebroadcast a social media message and, if so, when. In addition to incorporating content- and user-specific variables, we develop a new measure that captures the fit between content and user. The fit measure is especially important because it allows us to investigate the extent to which managers should tailor the message content for individual users in their audience. We also account for social influence in our modeling framework to capture both a user's ability to influence the behavior of others and a user's susceptibility to the influence of other users.

One challenge with modeling social media data is the unstructured nature of the text that constitutes the data. Consistent with prior research, we employ a text analysis procedure to incorporate unstructured textual data into our model. We apply Latent Dirichlet Allocation (LDA) methodology to analyze the text and identify the topics featured in the content of each message (e.g., Blei, Ng, & Jordan, 2003; Tirunillai & Tellis, 2014). We then integrate the LDA results into our proposed model structure, thereby allowing us to empirically identify the number of topics or key themes in the posted comments. The identification of key themes is important in our ability to characterize individuals' interests and their fit with posted content, which marketers can use to develop targeted messages.

Our results suggest – consistent with the literature – that the rebroadcasting of social media messages varies with the content of the message. However, more importantly, our analysis provides evidence that the fit between the message content and the audience's interest (as characterized by the content of their previous tweets) is a significant driver of rebroadcasting behavior. Our analysis reveals that rebroadcasts by influentials affect the rebroadcasting behavior of other users who are susceptible to influence. However, we observe considerable variation across users both in terms of their ability to influence other users and their susceptibility to the influence of others' rebroadcasting.

To illustrate the managerial relevance of our model, we simulate how tweets are retweeted over time by varying (1) the content of the original tweet, (2) the audience composition, and (3) the seeding strategies in which we induce different users to retweet soon after the original post. These simulations provide managers with insight into whether to focus their efforts on viral content, seeding strategies, or tailoring content to fit their audience's preferences. For example, firms could identify the appropriate content to disseminate to their unique audience base and develop a message by using words that have a high probability of being associated with that content topic (as determined by LDA). They could also identify influentials and disseminate messages with words that are associated with the topics generally broadcasted by the influentials. Our results suggest that tailored messages designed to match the preferences of the audience are most effective when there is less heterogeneity in the audience's interests. Our analysis also provides evidence that while seeding influential users can increase the rebroadcasting of a message, there may be a limit to its effectiveness because the interests of influential users may differ from those of the broader user base. Our findings highlight the importance of modeling content-user fit in a user's rebroadcasting decisions and consequently in a firm's social media marketing strategies.

The remainder of this paper proceeds as follows. In the next section, we discuss the related literature and show our conceptual framework of social media rebroadcasting. Next, we describe the data we employ in our empirical analysis and detail our approach for converting the unstructured text to quantitative metrics. In the Model section, we develop our modeling framework and describe how we incorporate user-specific differences, content effects, content-user fit and the role of social influence into a unified model. We then describe the results of our empirical analysis and conduct a series of simulations in which we evaluate the effect of alternative message design and seeding strategies on rebroadcasting activity. We conclude with a discussion of the implications of our research.

2. What drives rebroadcasting?

Social media rebroadcasting is the act of sharing content found online with social peers. Academics have a keen interest in understanding this activity, as marketers seek to leverage customers to propagate the firm's content in order to reach a larger

audience base that did not previously follow the firm (Lambrecht et al., 2015; Suh et al., 2010; Zaman et al., 2014). Previous research has identified two key drivers of rebroadcasting: content and influence. Thus, in this section, we first provide a brief review of the literature that discusses the importance of content. We then discuss a second stream of research that focuses on the role of social influence.

2.1. The role of content

Several researchers have examined how the content of social media messages affects the likelihood with which users rebroadcast or otherwise share messages with others. For example, Berger and Milkman (2012) examined the sharing of New York Times articles to better understand why some articles are commonly forwarded to their friends by readers while others are not. They found that articles that induced certain emotions (e.g., awe) are more likely to be shared than other articles. Likewise, Heath et al. (2001) found that emotional content was more likely to be shared. Specifically, people are more willing to pass along stories that contained the negative emotion of disgust.

Berger and Schwartz (2011) also recognized the importance of content in encouraging rebroadcasting activity; but rather than focusing on the dimensions of the content itself, they identified how the content interacted with the word-of-mouth context. Specifically, they found that content related to environmental cues were more likely to be the focus of word-of-mouth activity. This suggests that while the nature of the content can be important in understanding rebroadcasting behavior, it is also important to examine how content interacts with other factors.

Besides environmental factors, some users may be more likely to rebroadcast popular content. For example, Toubia and Stephen (2013) examine motivations driving rebroadcasting behavior and differentiate between intrinsic-related and image-related motivations. In their study, they conclude that some users share content on social media in an effort to manage their image or personal brand. This suggests that those with image-related motivations would selectively share content that would be consistent with the user's desired image.

While the aforementioned studies have looked at how different types of content are more or less likely to be shared, few (if any) have examined how content interacts with the user's preferences (i.e. content-user fit). One possible mechanism of the effects of fit on rebroadcasting is establishing a social image. If some social media users rebroadcast messages to build their personal brands (Toubia & Stephen, 2013), it seems logical that they would be more likely to rebroadcast content that is consistent with the image they have developed (e.g., Kirmani, 2009). Thus, in this research, we will examine both the effects of content on users' rebroadcasting activities as well as the effect of how that content fits with a given user's preferences.

2.2. The role of influence

Social influence is another important driver of rebroadcasting activity. We know from offline research that social contagion and influence can significantly affect the diffusion process (Bass, 1969; Bell & Song, 2007; Bhatia & Wang, 2011; Haenlein, 2013; Iyengar, Van den Bulte, & Valente, 2011; Nair, Manchanda, & Bhatia, 2010; Shriver, Nair, & Hofstetter, 2013; Van den Bulte & Joshi, 2007). In addition, within a population, there exist users who are inherently more influential than other individuals. These users can potentially affect others' decisions to rebroadcast. One possible explanation could be the effects of early adopters, whose adoption activity exerts social pressure and thus increases the adoption probability of other imitators (Bass, 1969). Certain individuals could also inherently be more influential due to their social status or character traits (Van den Bulte & Joshi, 2007). However, identifying the role of influence has been a challenge.

In the context of social media, a number of researchers have developed models that explicitly capture the effects of influence. For example, De Bruyn and Lilien (2008) develop a model to examine how word-of-mouth affects multi-stage decisions. Trusov et al. (2010) examine the dyadic relationships between users of an online social network and model how one's usage of an online service affects the usage behavior of others in his/her social network. Using a similar methodology, Aral and Walker (2012) conduct a randomized field experiment and measure how much an individual user's purchase activity influences the purchase behavior of others. Their measure of influence was then linked to the demographic characteristics of users (i.e. age, gender, and relationship status), allowing them to build a demographic profile of influential users. However, Watts and Dodds (2007) suggested that influence has two dimensions: (1) the ability of a user to influence others and (2) the susceptibility of a user to the influence of others. In both Trusov et al. (2010) and Aral and Walker (2012), the researchers separately modeled the ability to influence and the susceptibility to influence at the individual level.

While the previous research has developed tools to identify influence effects, there is little work on whether managers can leverage this in an organic manner (i.e. designing marketing communications that entices rebroadcasting activity from influentials without monetary incentives). In this research, we assess the extent to which tailoring content to match the preferences of influential users can increase the rebroadcasting activity of other users and identify a potential limitation of such a strategy that has not previously been documented.

3. Modeling framework and contribution

Our approach to modeling rebroadcasting behavior integrates the two separate research streams discussed above. Fig. 1 provides an overview of our modeling framework. We examine how content affects whether a message is rebroadcasted while also accommodating user-specific effects to allow for heterogeneity across individuals. Unique to our research, we examine how

the message content interacts with the user's preferences. We incorporate this into our modeling framework by developing a measure of fit between users and message content and allow this fit measure to impact a user's rebroadcasting decision. In other words, our rebroadcasting model examines user-specific and content-related "main effects" as well as the "interaction" between message content and the users. We also explicitly model the role of influence in the rebroadcasting decision. Using methodology similar to that of Trusov et al. (2010), we model each user's ability to influence as well as his/her susceptibility to influence.

While past researchers have identified content and influence effects, our contribution is two-fold. First, while the literature places emphasis on assessing the impact of either content (e.g., Berger, 2011) or influence (e.g., Aral & Walker, 2012; Trusov et al., 2010) in isolation, we propose an integrated model that considers content effects, influence effects, and content-user fit simultaneously. This allows us to measure the relative impact of content versus influence and their interplay with content-user fit on rebroadcasting behavior. An integrated approach is critical because investigating either content effects or influence effects in isolation can result in misleading inferences. For example, if an influential social media user is more prone toward sharing one type of content compared to another, we may mistakenly conclude that the content is more popular when it is actually the impact of the influential user that is contributing to increased rebroadcasting activity. Conversely, we may erroneously conclude that a user who rebroadcasts a message is influential because others subsequently rebroadcast when it may simply be that this user is inclined to rebroadcast highly shared content.

Second, we contribute to the rebroadcasting literature (Lambrecht et al., 2015; Suh et al., 2010; Zaman et al., 2014) by providing managerial guidance on how to disseminate messages effectively across specific audience bases. That is, we show that the effectiveness of certain content depends on the audience's preferences. To the best of our knowledge, this interplay between message content and individual preferences has not been documented previously in research on users' social media behavior. This is crucial for managers to understand, as many firms have different audience composition. As a result, a message content that successfully generated rebroadcasting activity for one firm may not be as effective for another firm. For example, firms may erroneously believe that certain viral content (i.e. content that is more likely to be shared on average) should be disseminated whereas in reality the content may not fit the preferences of their audience base and will subsequently fail to be rebroadcasted. Our novel approach to operationalizing content-user fit could be used to identify the optimal message content for a given audience base. However, we also demonstrate how a strategy that customizes messages for various audience members can possibly have limitations due to certain audience compositions.

4. Data

Given the relative nascence of social media data in academic research, we describe our data and some of the unique challenges involved in working with social media data before developing our model.

Our data is collected from Twitter, a popular social media platform. Many organizations use Twitter to disseminate messages to their followers. That is, an organization can broadcast, or "tweet," a short message that can be seen by any Twitter user who has "followed" that organization's Twitter feed. These users then have the option to rebroadcast, or "retweet," the organization's original message to their own followers, thereby amplifying the reach of the original message.

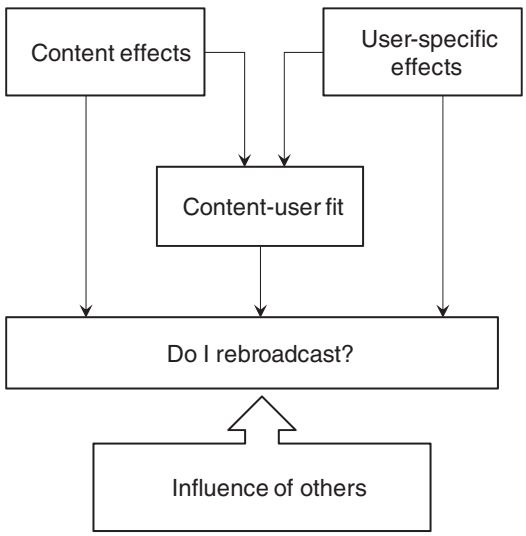

Fig. 1. Factors that drive rebroadcasting behavior.

For the purposes of this study, we focus on the retweets of messages posted by the top 10 business schools, ranked according to Business Week in 2011 (Top Business School Rankings: MBA, Undergrad, Executive and Online MBA – Businessweek, 2011).[1] We choose this set of "firms" because information dissemination is critical for business schools to attract students and connect with alumni (Symonds, 2013). Schools generate revenue and raise funds through enrollment and alumni support, both of which can be influenced through school generated social media communications. Our data, purchased from PeopleBrowsr.com, span an eight month period from April 2011 to November 2011 and include the full text and time of the original tweets posted by these business schools as well as all subsequent retweets by other Twitter users.

Table 1 provides some descriptive statistics of the retweet behavior and shows significant variation across schools. On average, each school posted 302 tweets during our data observation period. Of these tweets, almost half were retweeted (145). This is more than the average retweet rate on Twitter, at 6%, most likely due to the size of the follower bases and popularity of business schools (Sysmosos, 2010).[2] We also present the relative number of followers for each school, with Michigan indexed at 100. In all, our data sample includes a total of 3019 original tweets, of which 1454 were retweeted by 1760 users for a total of 3074 retweets. Thus, our data contains, for each user, an observation (either retweet or no retweet) for each original tweet posted by any business school she follows.

In addition to the tweet and retweet data, we expand our dataset to include a sample of non-retweets. That is, for each school, we collect a random sample of followers (50 from each school) who did not retweet any of the schools' messages during our observation period. This sample serves as a baseline that allows us to identify the differences between those who rebroadcast versus those who do not. As a result, our final data set includes the 1760 unique rebroadcasters found in the PeopleBrowsr dataset and the 500 additional users randomly sampled from the follower population, providing a total of 732,060 observations (3074 retweets and 728,986 non-retweets).

For each of the 3074 rebroadcast observations, the time elapsed, measured in minutes, since the original message broadcast is calculated. The median time to rebroadcast across all rebroadcast observations is 322 min, ranging from 1 min to 167 days. In the case of non-rebroadcasts (728,986 observations), the time elapsed between the original broadcast and the close of our data period is recorded as a survival time. We will describe how we accommodate the right censored nature of our data later in the Model Development section.

4.1. Text analysis and variable specification

One of the most significant challenges in social media research is characterizing the message content (i.e. unstructured text data) for use in a quantitative model (Lee & Bradlow, 2011; Netzer, Ronen, Jacob, & Moshe, 2012). While some research has focused on the valence of the message (e.g. East, Hammond, & Lomax, 2008; Schweidel & Moe, 2014), we differ by using the unstructured text to uncover the underlying topics of the message content. In this research, we need to develop measures of message content and content-user fit for use in our rebroadcasting model. Thus, in this section, we describe how we convert the text that appears in social media posts into a quantitative measure of content and fit.

Our text analysis follows a four-step process. First, we create a dataset of tweets that include (1) all original messages for the 10 schools in our data and (2) the most recent 200 messages (excluding retweets) from each user in our sample. This results in a total of 268,609 messages.[3]

Second, we preprocess the textual data by applying part-of-speech tagging to retain only nouns present in the school messages. We focus on nouns as our research objective is to identify focal topics that are more (or less) likely to be rebroadcasted.[4] To minimize redundancies, we use the Porter Stemmer (Porter, 2006) to reduce these nouns to their root form (e.g. "schools" is stemmed to form "school" and all occurrences of "schools" or "school" are considered to be the same). We also remove all URL links and stop words (e.g., "the," "this," "to," "a").

Third, we use latent Dirichlet allocation (LDA) to probabilistically determine the overarching topics from the message content (Blei et al., 2003; Tirunillai & Tellis, 2014). The advantage of LDA is the ability to extract a pre-specified number of topics from a large set of messages where the incidence of each individual word in any particular message is sparse. Thus, this state-of-the-art model is well suited to text analyzing Twitter messages. Our goal using LDA is to describe each message based on these topics, identify the general topic profile of each individual, and assess whether the school message is similar in fit with the topic profiles of the consumers. LDA facilitates our analysis because the method uncovers the latent topics of each message. Conceptually, LDA assumes that a consumer, when creating a Twitter message, wishes to convey one or more topics in a Tweet (e.g., a message about the economy). Conditional on the topic, she will generate words to reflect that topic based on how likely that word is associated with the topic (more likely to use words such as "jobs" or "market" than "coffee" for a message on the economy topic). Given that we observe the words, we can estimate the distribution of topic scores (i.e. probability of belonging to a topic) assigned to a

[1] We specify 10 schools to match the Business Week rankings.

[2] Our average retweet rate is comparable to previous studies. For example, Lambrecht et al. (2015) observed 16% and 53% retweet rates for the messages they disseminated. In addition, Zhao et al. (2011) found that in their representative sample of Twitter messages, the retweet rate ranges from on average 8% to 36% depending on the topic category. Given that our business schools rank in the top decile, it seems reasonable to stimulate slightly greater retweet rates than the average message on Twitter.

[3] We retrieve a maximum of 200 messages due to the limitations of the Twitter API. We note that not all users have 200 broadcast observations. For those with less than 200, we use all their available broadcast observations.

[4] Given the large set of potential words in Twitter messages, a focus on nouns provide more actionable managerial implications through the identification of specific topics as opposed to sentiment. We also run the same analysis using nouns, adjectives, and adverbs. The results are available from the authors.

Table 1
Top 10 business schools summary statistics.

Rank	School	# Tweets	Number of retweeters	# Tweets retweeted	Retweet observations	No-retweet observations	Number of followers
1	University of Chicago	273	131	52%	238	49,175	56,821
2	Harvard University	202	218	71%	319	53,817	255,166
3	University of Pennsylvania	496	414	61%	669	229,475	166,944
4	Northwestern University	225	191	68%	389	53,836	171,462
5	Stanford University	226	278	69%	476	73,652	336,618
6	Duke University	476	111	35%	250	76,386	16,865
7	University of Michigan	255	35	19%	80	21,595	11,566
8	University of California - Berkeley	527	163	30%	283	111,968	39,625
9	Columbia University	284	133	50%	240	51,732	70,414
10	Massachusetts Institute of Technology	55	86	31%	130	7350	132,320
	Total	3019	1760	49% (avg)	3074	728,986	125,780 (avg)

particular word and, based on the words in a message, infer the probability of each message belonging to a particular topic. We can also characterize an individual's topic preferences based on the topic scores of her historical messages. Thus, if a user posts many messages about jobs, the market, or inflation, then she is more likely to be described with an interest in the Topic "economy" if those words have a topic score relating to "economy." For full details of LDA, we refer readers to Blei et al. (2003) and Tirunillai and Tellis (2014).

In the final and fourth step, we identify the optimal number of topics. Given that the number of topics is pre-specified in the LDA model, we compare the performance of models using varying numbers of topics to select the optimal number of LDA dimensions. In our initial effort, we use the harmonic mean estimator (Newton & Raftery, 1994) and perplexity (Blei et al., 2003) to determine the number of topics. The fit statistics provided in Appendix A show that the optimal number of topics is about 50–51. One of the reasons for this large number of topics is that the LDA model itself is prone to overfitting (Blei et al., 2003). While this issue is not as vital for scenarios where text documents already revolve around one particular theme that may have a limited pool of words (e.g., online reviews for one product category), this is especially pertinent with Twitter data given the diverse set of messages and topics that could be discussed. Rather than deciding on the number of topics based solely on classifying the text, our approach is to choose the appropriate number of topics based on the performance of our rebroadcasting model (described in the Model Development section). To do so, we empirically compare the fit statistics of the rebroadcasting model using different numbers of topics from the LDA model, beginning with two topics and increasing the number of topics incrementally until performance suffers. Based on the performance of the rebroadcasting model, we empirically determine that the number of topics is six.[5] For any given message indexed by j, one topic was driven by the predominance of words specific to technology ($F_{j,1}$), the second by words related to schools ($F_{j,2}$), the third related to the economy ($F_{j,3}$), the fourth related to time-sensitive news ($F_{j,4}$), the fifth related to daily activities ($F_{j,5}$), and the sixth related to schools and the political environment ($F_{j,6}$). In Table 2, we show the 30 words with the highest posterior probabilities of being associated with each topic (displayed before stemming for ease of interpretation). Using these topics and the associated posteriors, a topic score was computed for each message in the data. This provided a concise and quantitative description of the message content. Table 3 provides an illustrative sample of messages and the respective posterior probabilities (i.e., topic scores) of belonging in each of the six topics.

As we text analyze the messages from both the schools and users, the topic scores associated with each message also allow us to characterize the interests of each individual user in our data. For each user, indexed by i, we compute the average topic scores corresponding to their 200 most recent original messages denoted $AVGF_{i,1} - AVGF_{i,6}$.[6] We include only new messages created by the user and exclude all rebroadcasted messages in these averages.

Finally, the message-specific topic scores, combined with the variables $AVGF_{i,1} - AVGF_{i,6}$, allow us to characterize the fit between each message and each user in our data. Specifically, we operationalize the fit between the content of a message and a user as being negatively related to the Euclidean distance between the topic scores for the message and the average topic score for the user. The fit between user i and message j is given by:

$$FIT_{ij} = -\sqrt{\sum_{h=1}^{H}\left(F_{j,h} - AVGF_{i,h}\right)^2} \tag{1}$$

where h indexes the topics identified from our text analysis and H is the total number of topics identified (in our case, $H = 6$).

[5] While the optimal number of topics is 6 in our context, managers can still tailor the content at the much granular word level given that the LDA model assigns a probability for each word being associated with a topic. Thus, mangers can develop messages with words that are likely to be associated with popular topics. We also use different number of topics in the LDA component and estimate the model with 2–10 topics as covariates (of course, fit also changes depending on the number of topics). The main results hold, although we do observe different effects of content depending on the granularity of the topics (more topics more granular).

[6] This variable is averaged across all original messages. If a user has less than 200 messages, we compute $AVGF_{i,1} - AVGF_{i,6}$ from all of the user's past messages.

Table 2
Top words for each topic.

Tech	School	Economy	Time-sensitive news	Daily activities	School & politics
google	business	jobs	time	day	mba
apple	school	obama	year	social	video
future	job	market	world	people	free
tech	wharton	india	news	work	blog
book	prof	state	game	twitter	post
mobile	women	million	years	media	online
ceo	data	china	change	life	innovation
company	power	wsj	team	week	education
app	plan	economy	show	morning	class
list	case	america	man	city	startup
reading	kellogg	billion	money	tweet	event
ipad	key	crisis	college	nyc	harvard
web	law	deal	house	problem	stanford
sales	professor	tax	christmas	network	food
music	human	debt	super	link	review
technology	photos	bill	head	mind	duke
party	video	story	chinese	room	romney
cloud	leader	bank	five	heart	penn
phone	corporate	economist	war	campaign	art
holiday	michigan	risk	person	coffee	hbs
point	model	government	fans	fail	paul
apps	berkeley	country	yahoo	park	space
books	london	private	death	friend	january
industry	biz	europe	woman	results	mit
site	club	cost	cnn	age	application
story	role	reuters	coach	film	john
month	dean	africa	baby	weather	gingrich
page	project	banks	members	staff	chat
amazon	haas	fund	car	james	airport
mark	annual	bloomberg	climate	fashion	executive

4.2. Control variables

We also consider a set of variables that characterize the posting activity of each school as well as a set of variables that characterize the social media activity associated with each user. First, we specify $SFREQ_j$ as the total number of posts during the week prior to the posting of message j as a measure of the school's posting frequency. This variable captures school level heterogeneity with respect to posting frequency. However, it is also message specific because it captures the frequency of posting by the school during the previous week (users may react differently when they receive many vs. few messages from the firm during a short time span). On average, schools post 16.3 tweets per week, with substantial variance both across schools and over time (standard deviation is 5.7 across schools and 9.0 across weeks).

Second, we consider three variables that characterize each user's position in their social network: (a) the number of followers a user has ($FOLLOWERS_i$), (b) the number of other individuals the user is following ($FOLLOWING_i$) and (c) a ratio of the number of followers to the number following ($RATIO_i$). The $RATIO_i$ variable is similar to one used by Anger and Kittl (2011) and would capture the effects of being an information seeker versus and information disseminator on social media (Java, Song, Finin, & Tseng, 2007).

Finally, we consider how frequently each user posts to social media. Specifically, we compute $UFREQ_i$ for each user as the average number of posts per week. On average, each user posts 4.5 messages per week, with a standard deviation of 5.6 (less

Table 3
Illustrative messages and posterior probabilities (topic scores).

Illustrative message	F1	F2	F3	F4	F5	F6
IBM's Watson took the prize but it was an extremely close match! #IBMWatson	**0.2905**	0.1190	0.1190	0.1476	0.1619	0.1619
Read how MBAs in our doctoral programs bring a special perspective to their role as future business educators #HBSDBA http://t.co/QN2YkKX9	0.1323	**0.3069**	0.1323	0.1323	0.1323	0.1640
More from Prof. Randy Kroszner on lackluster June jobs report in an interview with http://ow.ly/5CpX7	0.2186	0.1366	**0.2350**	0.1366	0.1366	0.1366
Prof. Austan Goolsbee scheduled to appear on CNBC Squawk Box program tomorrow Thursday during the 7 a.m. hour central time.	0.1366	0.1530	0.1366	**0.2678**	0.1366	0.1694
Prof. Luigi Zingles' latest findings of the Chicago Booth/Kellogg School Financial Trust Index released today http://t.co/NopSgo74	0.1488	0.1488	0.1488	0.1488	**0.2560**	0.1488
Prof. Raghu Rajan contends that private sector profits will not save the US from the politically-divisive federal deficit http://ow.ly/5jimk	0.1590	0.1282	0.1897	0.1282	0.1282	**0.2667**

active than schools). While the average is less than one broadcast per day, we observe up to as many as 40.6 posts per week from highly active individuals.

5. Model development

We assume that rebroadcasting behavior follows a split hazard process (Sinha & Chandrashekaran, 1992). A split hazard specification is well suited to our context as it (1) allows for covariate effects in the probability and (2) accommodates right censoring (Aral & Walker, 2012; Iyengar et al., 2011; Katona, Zubcsek, & Sarvary, 2011; Trusov et al., 2010).

Let $y_{ij} = 1$ be an indicator to denote that individual i is observed to rebroadcast message j, and let t_{ij} be the time at which this occurs (measured as the number of minutes since the original message is posted). The likelihood associated with individual i's behavior can then be written as:

$$L\left(y_{ij}, t_{ij}\right) = \begin{cases} P_{ij} f\left(t_{ij}\right), & y_{ij} = 1 \\ \left(1 - P_{ij}\right) + P_{ij} S\left(T_j\right), & y_{ij} = 0 \end{cases} \tag{2}$$

where P_{ij} represents the probability that user i will rebroadcast message j at some point, $f(t_{ij})$ represents the likelihood of doing so at time t_{ij}, T_j represents the time between when message j was posted and the end of the data observation period, and $S(T_j)$ is the survival function that captures the probability of not retweeting before time T_j.

If $y_{ij} = 1$, user i rebroadcasts with probability P_{ij}, and $f(t_{ij})$ governs when in time the rebroadcast occurs. If $y_{ij} = 0$, we have two possibilities. First, individual i may never rebroadcast, with a probability of $1 - P_{ij}$. Second, individual i may eventually rebroadcast the message, but this event is censored in our data. Thus, $S(T_j)$ accounts for the censoring of those who will eventually rebroadcast the message, which occurs with a probability P_{ij}.

Let us first specify the probability, P_{ij}. We assume that the probability of user i rebroadcasting message j is a function of content effects (C_j), user effects (U_i), and the effects of user i's specific interest in the content of j (FIT_{ij}). Thus, we specify P_{ij} as:

$$P_{ij} = \Phi\left(\beta_0 + C_j + U_i + \delta \cdot FIT_{ij}\right) \tag{3}$$

where $\Phi(x)$ denotes the standard normal c.d.f, β_0 is the intercept, and the δ coefficient represents the effect of the user-content fit on rebroadcasting behavior.

The content effects, C_j, are normally distributed with a mean that is influenced by the original message's content ($F_{j,1} - F_{j,6}$) and the posting frequency of the school during the previous week ($SFREQ_j$). We allow heterogeneity in C_j across different messages, where σ_m^2 is the variance among the content effects for each message. This specification allows us to account for school level heterogeneity through both observed differences (i.e. some schools post more frequently; some schools tend to post about one topic whereas others post about another topic) using the above variables and unobserved differences using σ_m^2.

$$\begin{aligned} C_j &\sim N\left(\mu_j, \sigma_m^2\right) \\ \mu_j &= \beta_1 \times SFREQ_j + \sum_{h=1}^{6} \beta_{1+h} \times F_{j,h} \end{aligned} \tag{4}$$

We similarly specify user effects U_i as being affected by the user's profile as measured by $AVGF_{i,1} - AVGF_{i,6}$. We also control for the user's social network position as measured by the number of $FOLLOWERS_i$, the number of users he or she is $FOLLOWING_i$ and the followers to following $RATIO_i$.[7] We log transform each variable (after adding one to avoid a log of zero) to rescale the highly varied and skewed distributions observed. For the $RATIO$ variable, we add one to the log-transformed elements to avoid dividing by

[7] We use the ratio between followers to following to control for the "type" of users: either an information seeker or information disseminator (Anger & Kittl, 2011; Java et al., 2007), as it is an important factor in one's decision to rebroadcast. Users who have a much larger proportion of followers tend to be an information disseminator. In contrast, those who follow a greater proportion of individuals tend to be an information seeker. Thus, we interpret the ratio as conditional on the absolute number of followers and following, a proxy for information seeking or disseminating status. We also were unable to collect time-varying data with respect to the number of followers and following. These were collected at the end of our observation period.

zero. Finally, we also accommodate the potential effects of a user's poster frequency ($UFREQ_i$) of rebroadcasting behavior. Again, we allow heterogeneity in U_i across different individuals, where σ_u^2 is the variance among the user effects for each individual.

$$\begin{aligned} U_i &\sim N\left(\eta_i, \sigma_u^2\right) \\ \eta_i &= \sum_{h=1}^{6} \beta_{7+h} \times AVGF_{i,h} + \beta_{14} \times \log(FOLLOWERS_i + 1) + \\ \beta_{15} &\times \log(FOLLOWING_i + 1) + \beta_{16} \times \frac{\log(FOLLOWERS_i + 1) + 1}{\log(FOLLOWING_i + 1) + 1} + \beta_{17} \times UFREQ_i \end{aligned} \tag{5}$$

We highlight the content and user effects also capture unobserved heterogeneity across schools. One type of school-level heterogeneity is the different composition of followers, who may have varying preferences or reactions to twitter posts. The individual level effect (U_i) captures these variations among individuals across schools. Other types of school-level heterogeneity may include the type of content different schools disseminate and the frequency of their posts. These are captured by the content level effects (C_j) and the frequency of posting variable ($SFREQ_j$).

While the user interest variables captures a pattern between an individual's profile and his general tendency to rebroadcast, this does not account for an individual's tendency to rebroadcast certain types of content. We account for the latter using the content-user interaction (FIT_{ij}) in Eq. (3).

We turn now to specifying the hazard process. Conditional on the message being rebroadcast, the hazard component of the model governs the timing of rebroadcasts and can allow for time-varying effects (note that our specification of P_{ij} includes only non-time-varying effects). Consistent with Trusov et al. (2010) and Aral and Walker (2012), we include the effects of influence as a time-varying covariate in the hazard component of the model. For identification purposes, we assume that influence only affects the time varying component of the model since it depends on the extent that an individual is exposed to prior rebroadcasting at a specific point in time (e.g., see Iyengar et al., 2011). This assumption is consistent with Libai, Muller, and Peres (2013) who find that seeding programs aimed at influentials drive the rate at which customers accelerate their purchase adoption to an earlier date rather than whether or not to purchase.

We employ a Weibull process for the baseline hazard and capture the effects of any social influence that may result from the actions of previous rebroadcasters ($INFL_{ijt}$) as follows:

$$h_i(t) = \lambda_i c t^{c-1} \exp\left(INFL_{ijt}\right) \tag{6}$$

where λ_i and c are parameters of the Weibull distribution to be estimated, and $INFL_{ijt}$ is a time-varying covariate that characterizes how previous rebroadcasters of message j affect user i at time t (i.e. Total Influence).

We specify the $INFL_{ijt}$ as the influence from all previous rebroadcasters weighted by individual i's susceptibility to that influence[8]:

$$INFL_{ij\tau} = \alpha_i \sum_u \left[\gamma_u \times 1\left(t_{uj} < \tau\right)\right] \tag{7}$$

Specifically, α_i reflects the extent to which the focal individual i is susceptible to influence and $\sum_u [\gamma_u \times 1(t_{uj} < \tau)]$ captures the sum of the influence from all previous rebroadcasters (indexed with u) of message j. γ_u captures the influence of the previous rebroadcaster, u, and the indicator function $1(t_{uj} < \tau)$ ensures we incorporate user u's influence only if she has rebroadcast message j by time τ. This approach is consistent with the method employed by Trusov et al. (2010).

To further explain the intuition behind Eq. (7), we describe how the previous rebroadcasters may influence user i at time τ. First, i refers to the focal individual whose rebroadcast outcome for message j we are modeling and u refers to users who have previously rebroadcasted. Given that i follows the school that sent message j, i is also exposed to the retweets by others such as u who retweet messages sent by that same school. As a result, u can potentially influence i. Second, the total influence depends on i's susceptibility to influence (i.e. α_i). If i is more susceptible, then she will be more likely to be affected by the influence of others (thus $INFL_{ijt}$ is larger). Third, the total influence also depends on the ability of other users (who have rebroadcasted) to influence (i.e. $\sum_u [\gamma_u \times 1(t_{uj} < \tau)]$). Thus, the total influence ($INFL_{ijt}$) increases with the number of influential individuals who have already rebroadcast the message. Our analysis treats the influence effect as time varying, since only users who have rebroadcasted at time $t_{uj} < \tau$ will have the potential to influence individual i at time τ.

Modeling both influence and susceptibility to influence at the individual level presents identification challenges. In our data, each individual is exposed to multiple tweets. Additionally, those tweets may be associated with a varying number of influentials at different points in time. The variation within an individual, across tweets, and over time aids in the identification of the influence variables. Specifically, to identify γ_u, we rely on variation in the subsequent rebroadcast activity of others in the sample. That

[8] We note that all individuals i and users u represents the same set of people in our sample. We use the different subscripts to separate the effects of u's influence (γ_u) from i's susceptibility (α_i) in Eq. (7). i and u are connected through their mutual following of the tweets by a particular business school. Thus, if u retweets from school A, i is exposed to that retweet because i follows the activity of school A (and thus messages that relate to school A).

is, if more individuals subsequently retweet sooner, then user u is considered influential and will tend to have a greater γ_u estimate. To identify α_i, we rely on the variation in the number and influence of users who have previously rebroadcasted. Thus, a consumer has greater susceptibility (α_i) if we observe her having greater retweeting activity as the number of consumers who have previously retweeted (or their respective influence) increases (i.e. greater $\sum_u [\gamma_u \times 1(t_{uj} < \tau)]$). Finally, we are able to identify γ_u from α_i, as there is variation in the influentials who retweet across messages.

In our empirical setting, we are only able to identify the relative magnitude between γ_u and α_i, as our data is sparse (only 3074 retweet observations and 728,986 non-retweet observations). For example, not all individuals have multiple retweet observations. This hinders our ability to identify individual level influence estimates. However, there is sufficient variation with respect to the sum of the influence of users who retweet (Appendix A reports the distribution of $\sum_u [\gamma_u \times 1(t_{uj} < \tau)]$). We achieve identification by specifying our model to allow for heterogeneity in a user's ability to influence by assuming that the user population consists of a discrete mixture of influentials where $\gamma_u = 1$ with probability π_u and non-influentials where $\gamma_u = 0$, with probability $1 - \pi_u$. We similarly specify α_i as a mixture of users who are susceptible to influence from the previous rebroadcasts of others and those who are not. Specifically, we assume that users are susceptible to influence such that $\alpha_i = \alpha$ with probability ϕ_i. With probability $1 - \phi_i$, user i is not susceptible to influence, or $\alpha_i = 0$. This approach is consistent with that employed by Trusov et al. (2010) and allows us to identify γ_u and α_i by restricting the parameter space to 1 and 0 for γ_u and 0 and α for α_i. Details of the sampling procedure are provided in Appendix B.

We specify γ_u to be a mixture of zeros and ones while α_i is a mixture of zeros and α.[9] In other words, we assume that users are either influential or not, and individuals have either some level of susceptibility to the actions of influentials or have no susceptibility. While we expect α to be positive, we assume that α has a diffuse normal prior to allow for scenarios where individuals might have negative susceptibility. That is, if $\alpha_i < 0$, then the rebroadcasting by others may deter subsequent rebroadcasting by individual i as opposed to encourage it.

To estimate the model, we specify diffuse priors for all hyper-parameters: $\lambda_i \sim Gamma(a_{\lambda 0}, b_{\lambda 0})$, $c \sim Gamma(a_{c0}, b_{c0})$, $\beta \sim N(u_0, \sigma_0^2)$, $\delta \sim N(\delta_0, \upsilon_0^2)$, $\sigma_u^2 \sim Gamma(a_{u0}, b_{u0})$, $\sigma_m^2 \sim Gamma(a_{m0}, b_{m0})$. Please refer to Appendix B for the specification of the priors with respect to the influence components (α_i and γ_u). We use a block Metropolis-Hastings algorithm to draw each parameter separately. We run 80,000 iterations and discard the first 60,000 for burn-in. The remaining 20,000 iterations are used to form our posterior results.

6. Model comparisons

6.1. Selecting the number of topics

Before discussing the estimation results, we first explain how we empirically selected the number of topics. As mentioned earlier, instead of using the fit statistics for only the LDA model, our approach is to choose the appropriate number of topics based on their incremental value with respect to the performance of our rebroadcasting model. In Table 4, we compare measures of model fit (deviance information criterion (DIC)[10] and log marginal density calculated on the full model) to assess the value of each incremental topic in the LDA model. Using this selection method, our analysis suggests that the ideal number of topics for analysis is 6, as additional topics in the LDA model do not provide incremental improvement in the performance of the proposed rebroadcasting model. The reduced number of topics (compared with the 50 topics that would be selected based solely on the text analysis) possibly provides a more parsimonious set of topics, offering efficiency in model estimation and aiding interpretability.

6.2. Determining model specification

We also evaluate our proposed model that includes content effects, user effects, and influence components by comparing it to a number of alternative model specifications. Specifically, we compare in-sample and out-of-sample fit across a number of nested models and evaluate each model using fit and prediction metrics (see Table 5). For the in-sample comparisons, we calculate deviance information criterion (DIC) and log marginal density on the entire dataset. For the out-of-sample fit statistics, we estimate our model on a calibration data set and forecast on a holdout dataset. In our data, each individual is exposed to all the messages that are disseminated from the school(s) she follows. That is, if she follows only the University of Chicago and that school broadcasts 100 messages, she is exposed to those 100 messages. This means that there are 100 individual-message specific observations for her. Next, for each individual, we randomly select 25% of the individual-message specific observations (i.e. messages she was exposed) to as the holdout. This means that we use 75% of the individual-message observations as calibration for the model, where we estimate the parameters on this calibration data and use the posteriors to forecast for the holdout observations.[11]

Our out-of-sample fit statistics are as follows. We first consider the likelihood of observed rebroadcasting in the holdout sample using posterior means for each individual and message. To examine forecasting accuracy, we first calculate the hit rate associated

[9] Trusov et al. (2010) estimate α_i as a continuously distributed variable. We use a discrete mixture due to differences in our data sparseness.

[10] DIC may be problematic for mixture models (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). As a result, we also use both log marginal density and a variety of predictive metrics (Table 5) for model comparison.

[11] The covariate measures for the calibration samples are not affected by the holdout. The goal of the holdout comparison is to simply test the forecast ability of the model.

Table 4
Selection of optimal number of topics based on model fit.

Number of topics	3	4	5	6	7	8	9
DIC	86,580	86,116	85,833	**80,452**	85,535	85,737	86,021
Log marginal density	−42,724	−42,468	−42,457	**−40,645**	−42,355	−42,427	−42,516

Bolded values indicate the best fitting model with respect to the number of topics.

Table 5
Model comparison.

Included variables/model	1	2	3	4
Content effects (C_j)	X		X	X
User effects (U_i)		X	X	X
Content-user interaction (FIT_{ij})			X	X
Heterogeneous (λ_i)		X	X	X
Heterogeneous influence ($INFL_{ijt}$)				X
DIC	143,228	83,830	82,561	**80,452**
Log marginal density	−71,826	−41,787	−41,053	**−40,645**
Likelihood of random 25% holdout	−17,254	−10,196	−10,132	**−9741**
Hit rate (25% holdout)	0.734	0.991	0.992	**0.993**
APE (number of retweets, 25% holdout)	0.639	0.210	0.184	**0.160**
MAD (minutes, 25% holdout)	542.317	49.820	47.430	**45.490**

with accurately predicting a retweet or no-retweet observation.[12] Second, we report the accuracy in predicting the total number of retweets by the holdout sample for a message as the average percent error (APE), calculated using the forecasted and the actual number of retweets for each message. Finally, we predict the timing of retweets and assess performance using the mean absolute deviation (MAD) between the actual time and the forecasted retweet time (in minutes).

The first model that we estimate (Model 1) is a baseline model that includes only content effects and ignores any individual-specific, content-user fit, or influence effects. Model 2 includes only individual-specific effects (including user heterogeneity) and ignores content-specific, content-user fit, and influence effects. Model 3 includes both content-specific and individual-specific effects and also incorporates the content-user fit effects but omits any influence effects. Finally, Model 4 adds the influence effects and is the full model proposed in the previous section.

When comparing Models 1 and 2, our results suggest that incorporating user-specific covariates and heterogeneity dramatically improves model fit, more so than simply accounting for content effects. In other words, attributing the virality of a tweet to only the nature of its content may be oversimplifying the factors at play. The inclusion of content-user fit in Model 3 also improves fit. While the improvements are modest, we observe more significant improvements when the role of influence is added in Model 4. Overall, these model comparisons suggest that each component of the proposed model contributes to the ability of the model to capture the behavior observed in the data. Thus, for the remainder of this paper, we will focus on the results associated with the full model (Model 4).

7. Model results

7.1. Content, user, and interaction effects

Table 6 provides estimates for the effects of message content (C_j), user interests (U_i), and the fit between content and user (FIT_{ij}) on the probability, P_{ij}, that individual i rebroadcasts message j at some point in time. First, our results suggest that, broadly speaking, the content of social media messages plays an important role in affecting rebroadcasting behavior. Our analysis provides evidence for significant variation in the baseline rebroadcasting rates across content, as some content (school with $\beta_3 = 0.020$) is more likely to be rebroadcast while other content is less likely to be rebroadcast (tech, time-sensitive news, and daily activities with $\beta_2 = -0.013$, $\beta_5 = -0.013$, and $\beta_6 = -0.011$, respectively). In addition, messages from schools that broadcast more frequently during the past week have less rebroadcasting activity ($\beta_1 = -0.011$), possibly due to information overload diminishing the tendency with which any one message is rebroadcast (Edmunds & Morris, 2000).

Second, our results suggest that rebroadcasting behavior varies significantly across individuals. One user-level characteristic that could potentially drive this variation is the type of content one tends to broadcast. In our context, users interested in school-related topics ($\beta_9 = 0.177$) are generally more likely to rebroadcast than those interested in tech, daily activity, or politics related topics ($\beta_8 = -0.057$, $\beta_{12} = -0.088$, $\beta_{13} = -0.037$, respectively). In addition, users who tend to post more frequently are also more likely to rebroadcast ($\beta_{17} = 0.005$). Finally, with respect to network size, users who have more (fewer) followers

[12] A prediction is accurate if we correctly predict the rebroadcasting outcome (rebroadcast or no-rebroadcast). Specifically, our prediction value is $(1-|\widehat{Y_{ij}}-Y_{ij}|)$, where $\widehat{Y_{ij}}$ is the predicted rebroadcast (1 for yes, 0 for no) and Y_{ij} is the actual value. Thus, the predicted value is 1 if we correctly predict the true rebroadcast outcome. We note that the hit rate for Models 2, 3, and 4 are high due to the fact that rebroadcasts are rare in our empirical context.

Table 6
Model 4 results – decision to rebroadcast.

Component	Variable	Parameter estimate	Standard error
	Intercept (β_0)	**−1.970**[a]	(0.029)
Content	*SFREQ* (β_1)	**−0.011**[a]	(0.001)
	F1 (Tech, β_2)	**−0.013**[a]	(0.007)
	F2 (School, β_3)	**0.020**[a]	(0.005)
	F3 (Economy, β_4)	0.005	(0.010)
	F4 (News, β_5)	**−0.013**[a]	(0.007)
	F5 (Daily Activity, β_6)	**−0.011**[a]	(0.006)
	F6 (Politics/School, β_7)	0.008	(0.011)
User	*AVGF1* (β_8)	**−0.057**[a]	(0.014)
	AVGF2 (β_9)	**0.177**[a]	(0.015)
	AVGF3 (β_{10})	−0.018	(0.013)
	AVGF4 (β_{11})	−0.006	(0.015)
	AVGF5 (β_{12})	**−0.088**[a]	(0.014)
	AVGF6 (β_{13})	**−0.037**[a]	(0.008)
	FOLLOWERS (β_{14})	**0.143**[a]	(0.007)
	FOLLOWING (β_{15})	**−0.163**[a]	(0.006)
	RATIO (β_{16})	**−0.526**[a]	(0.024)
	UFREQ (β_{17})	**0.005**[a]	(0.001)
Content-user fit	*FIT* (δ)	**0.011**[a]	(0.004)

[a] Bold values indicate zero is not contained in the 95% confidence interval.

(following) are more likely to rebroadcast. This is reasonable as consumers who have more followers could potentially rebroadcast to spread information to their followers. In contrast, consumers who have more people they follow are possibly more likely to be information seekers who use the platform to gather rather than disseminate information (Java et al., 2007). Conditional on the amount of followers, we posit that users who are following fewer people may potentially have less content to rebroadcast, thus the negative coefficient on the ratio term ($\beta_{16} = -0.526$).

Finally, and most importantly, we examine whether the content-user fit measure impacts rebroadcasting behavior. After accounting for variation across individuals in their rebroadcasting tendencies, our results provides evidence that users are more likely to rebroadcast content that matches their own interests ($\delta = 0.011$). There are two new and important implications of this result. First, this finding provides evidence for the importance of jointly considering both content effects and user characteristics, as the fit between content and user is an essential driver of rebroadcasting behavior. This suggests that users are more likely to rebroadcast messages with content that are related to their interests than messages with other types of content. Second, this result implies that organizations can tailor content to match the audience's interests in order to increase rebroadcasting activity from them. We further explore the value of tailored content through simulations, which we present in Section 8.

7.2. The role of influentials

While the above results indicate who is more likely to rebroadcast, it does not indicate whether those users are more likely to affect the rebroadcasting behavior of others. In this section, we evaluate the impact of influentials and those susceptible to influence by examining the parameter estimates associated with the hazard component of the model (see Table 7). The role of influence is represented by a combination of parameters. First, π_u reflects the probability that user u is influential. Second, the parameters ϕ_i and α reflect whether individual i is susceptible and the level of susceptibility, respectively. Therefore, for those individuals who are susceptible to influence, *when* they rebroadcast may be affected by the number of influential users who have previously rebroadcast ($\alpha = 0.404$).

Our results provide evidence that there is heterogeneity in both a user's ability to influence as well as a user's susceptibility to influence. Figs. 2 and 3 show the distributions of the posterior means of π_u and ϕ_i across individuals. These figures suggest noticeable heterogeneity across the population on both dimensions.

Table 7
Model 4 results – timing of rebroadcast.

Parameter	Description	Parameter estimate	Standard error
λ	Rebroadcasting rate (scale)	0.010	(0.009)
c	Rebroadcasting rate (shape)	0.796	(0.019)
α	Susceptibility to rebroadcast	**0.404**[a]	(0.019)
		Mean effect	Standard deviation
π_i	Probability of being key rebroadcaster	0.378	(0.131)
φ_i	Probability of being susceptible	0.564	(0.069)

Note. We report the empirical mean and distribution of π_i and φ_i.
[a] Bold values indicate zero is not contained in the 99% confidence interval.

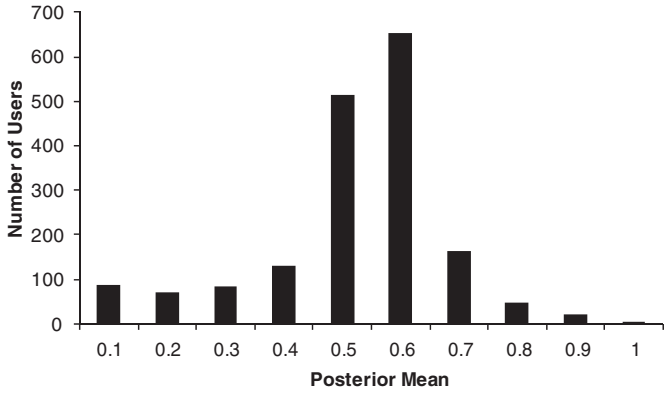

Fig. 2. Distribution of influence (posterior mean π_u).

Overall, our findings provide evidence that not all individuals contribute to the propagation of social media messages in the same way. Some may have a very high probability to be influential while others are more likely to be non-influential. However, even after identifying the influentials, their value on rebroadcasting may not be readily apparent without jointly considering the susceptibility of the population exposed to influence. Influentials seem to be are less likely to be impactful on rebroadcasting if other users exhibit a low level of susceptibility (low ϕ_i) to the prior rebroadcasts of others. This is consistent with Watts and Dodds (2007) who contend that a population of individuals susceptible to influence may play as important a role as a population of influentials in contributing to an information cascade.

8. Evaluating social media messaging strategies

In this section, we discuss how the proposed model can provide managerial insights as to which communications strategies should be employed when seeking to expand the reach of social media messages through rebroadcasting activities. In particular, we follow Lambrecht et al. (2015) and assess how firms can increase the propagation of their messages. However, rather than target early or later propagators as tested by Lambrecht et al. (2015), our framework allows managers to compare the expected efficacy of (1) developing customized content to (2) relying on influentials to spread the message.

To illustrate how our model can be employed, we simulate the rebroadcasting activities resulting from various content design and influential targeting scenarios. Our simulation procedure is as follows:

1. For each of the 2260 individuals in our data, we compute *i's* probability of rebroadcasting (P_{ij}) using individual *i*'s and message *j*'s model-based posteriors for their respective topic scores.
2. We then simulate the decision to rebroadcast the message through Bernoulli draws with probability P_{ij}.
3. For each individual who will eventually rebroadcast, we simulate the timing of *i*'s rebroadcast using posteriors estimates from hazard model component.
4. We designate the individual rebroadcaster with the lowest simulated t as the initial rebroadcaster.
5. We probabilistically determine whether each individual is influential based on his/her posterior π_u.
6. We simulate the timing of the next rebroadcast conditional on the influence that earlier rebroadcasters may have (Eqs. (6) & (7)). Thus, at time t, if users A and B have rebroadcasted, then our simulation of the timing of the next rebroadcast at or after time t is conditional on the influence of A and B (e.g., those who have already rebroadcasted) and the susceptibility of all potential future rebroadcasters.

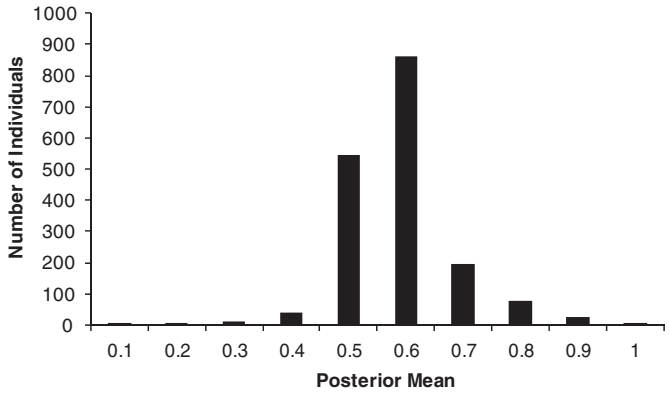

Fig. 3. Distribution of susceptibility to influence (posterior mean ϕ_i).

Table 8
Comparing content in simulated populations.

		Heterogenous audience base		Homogenous audience base					
		Empirical	Balanced	Topic 1 (tech)	Topic 2 (school)	Topic 3 (economy)	Topic 4 (news)	Topic 5 (daily activities)	Topic 6 (politics / schools)
Message content (posterior parameter values in parentheses)	F1 (–0.013*)	**6.434**	**6.413**	**5.834**	**7.817**	**6.520**	**5.773**	**6.179**	**6.356**
	F2 (0.020*)	4%	5%	–1%	7%	5%	6%	5%	6%
	F3 (0.005)	4%	4%	–2%	3%	10%	4%	3%	4%
	F4 (–0.013*)	0%	0%	–6%	0%	1%	6%	1%	0%
	F5 (–0.011*)	1%	1%	–5%	1%	1%	2%	6%	1%
	F6 (0.008)	2%	3%	–3%	4%	3%	3%	3%	8%
	Percent lift	4%	5%	–6%	7%	10%	6%	6%	8%

*Notes. The baseline number of retweets is for a message content of topic F1 (least viral content). The percentages indicate the lift of a different message content relative to the baseline. The Empirical audience base consists of the entire population in the data. The Balanced audience base consists of a population with an equal amount of users who are ranked in the top 10% of each topic. The Topic 1 (or 2/3/4/5/6) user base consists of a population where all individuals are drawn from the top 10% of users ranked by topic 1 (or 2/3/4/5/6). F1 (or 2/3/4/5/6) is a synthetic message that has 100% probability of being a topic 1 message (i.e. designed by the firm). Grey highlight shows the effect of content and fit.

Table 9
Simulation comparing seeding and content-fit strategies.

	No seed	Seed random	Seed influential	FIT influential
2 users (0.1%)	6.563	0.64%	4.58%	7.60%
5 users (0.25%)	6.563	0.76%	7.88%	5.25%
10 users (0.5%)	6.563	0.80%	9.18%	3.00%

*Notes. The numbers under No Seed indicate the total number of retweets for a no seeding strategy across a one week period. The percentages indicate the lift above the no seed condition for the random seed, seed influential, or fit influential strategies. Fit influential refers to developing a message content that matches the preferences of influentials in an attempt to generate organic influential retweet activity.

7. This procedure is repeated so long as the rebroadcasts occur within a specified timeframe (one week).

We use the posterior estimates from the last 1000 iterations from the estimation sampler. For each set of posterior estimates, we simulate the rebroadcasting behavior that would result from a number of strategic scenarios that we will describe in depth below. We average across 500 simulated iterations for each set of posteriors and then across the 1000 estimation iterations to obtain our simulation results which we present next.

8.1. Simulation study 1: focusing on content

The purpose of our first set of simulations is to assess the extent to which firms can rely on customized content to drive rebroadcasting activity. To do so, we generate six synthetic messages, each focusing on only one of the six topics recovered in the LDA. For example, we create a message emphasizing the "tech" topic (Topic F1) by specifying that message to have a 100% probability of belonging to Topic F1 and 0% probability of belonging to one of the other topics (and so on for the other messages).

We disseminate each message in eight synthetic populations with varying content preferences. The first audience consists of the actual population in our empirical setting. The second audience is a balanced population with heterogeneous topic interest, where we rank all users according to their topic scores for Topic 1 and randomly select 1/6 of the user base from those ranked in the top 10% (and repeat for the 6 topics). For the third through eighth audiences, we create a homogeneous population with a focus on one specific topic (e.g., for a Topic 1 audience base, we randomly select users that are ranked in the top 10% of topic 1 scores). For each message topic and audience combination, we simulate a sequence of rebroadcasting behavior using the previously outlined simulation procedure.

The results from this simulation are presented in Table 8, where each element of the table represents the rebroadcasting activity associated with each content (row) and audience (column) combination that is expected within the first week following the original tweet.[13] The results in the first two columns suggest that in a heterogeneous audience, a Topic 2 message generates the greatest rebroadcasting activity. However, when faced with a homogeneous audience (i.e. all users tend to discuss one particular topic), the best strategy for the firm may be to disseminate a message with content that fits the preferences of the audience base. As a result, while previous research has found that certain content is more likely to be shared (e.g., Berger, 2011; Berger & Milkman, 2012; Berger & Schwartz, 2011; Heath et al., 2001), our results suggest that tailoring the message content to match the preferences of the audience (and not necessarily to match that of viral content topics) could be very effective in a homogeneous population.

8.2. Simulation study 2: identifying and seeding influentials

The purpose of the second set of simulations is twofold. First, we examine the impact of seeding influential users (i.e. incentivizing them to rebroadcast the messages immediately). Second, we assess the impact of tailoring content to match the preferences of influentials (i.e. without incentives). We identify influential as those individuals associated with high posterior estimates for $\hat{\pi}_u$.

For the seeding simulations, we assume that influentials targeted by the seeding strategy rebroadcast the message immediately following its original broadcast. The seeded influentials' retweets are not included in the final count. For the content tailoring simulations, we identify the top influentials and disseminate tailored messages that match the preferences of these influentials.[14] We compare these strategies with a random-seed strategy in which we randomly chose individuals to rebroadcast immediately.

Table 9 shows the results of the "no seed" strategy and the resulting percentage increase in rebroadcasts (compared to the "no seed" strategy) using random-seed and influential-seed strategies. We also vary the number of seeded influencers: 2 users (0.1% of population), 5 users (0.25% of population), or 10 users (0.5% of population). Overall, seeding influentials seems to accelerate the rebroadcasting process. In terms of lift, with 10 users, seeding influentials generate 8.4% more rebroadcasting activity than a random seed strategy. In addition, our results suggest that seeding strategies focused on influentials are more efficient than strategies that rely on the volume of individuals seeded (i.e. randomly seeding a large amount of individuals).

[13] The numbers in the first row indicate the 1 week total rebroadcasting activity for baseline message content (topic 1, least viral). The percentages in the subsequent rows indicate the lift for different message content relative to the baseline and the gray highlight shows the effect of content that fits the preferences of the audience. The bottom row indicates the percent difference in rebroadcasting activity between the best and worst performing content. Parameter estimates for content effects are also listed in column 1 (in parentheses) for convenience.

[14] Specifically, when tailoring content to two influentials, we release two different messages (tailored toward the two influentials) to the entire population and record the resulting rebroadcasting activities for both messages. We then average the rebroadcasting activities (excluding the retweets of the influentials) and report the average per-tweet percentage increase above the "no seed" strategy in Table 9.

Table 10
Comparing content vs. influence.[a]

		Heterogenous audience base		Homogenous audience base	
		Empirical	Seed influentials	Topic 2	Seed influentials
Message Content (posterior parameter values in parentheses)	F1 (−**0.013**)	6.434	7.035	7.817	8.308
	F2 (**0.020**)	6.714	7.344	8.397	9.009
	F3 (0.005)	6.703	7.282	8.088	8.529
	F4 (−**0.013**)	6.450	7.032	7.817	8.288
	F5 (−**0.011**)	6.498	7.104	7.885	8.340
	F6 (0.008)	6.580	7.199	8.125	8.584

Bolded values indicate zero is not contained in the 95% confidence interval.

[a] Notes. The numbers indicate the total number of retweets for a certain message content in the specified audience base. The Empirical audience base consists of the entire population in the data. The seed influentials columns use the strategy to seed 10 influentials (0.5%) in an empirical audience base (left panel) or a Topic 2 audience base (right panel). The Topic 2 user base consists of a population where all individuals are drawn from the top 10% of users ranked by topic 2. F1 (or 2/3/4/5/6) is a synthetic message that has 100% probability of being a topic 1 message (i.e. designed by the firm).

Table 9 also reports the results from the content tailoring strategy, which slightly outperforms that of seeding influentials in the 2 seed scenarios. We gain greater rebroadcasting activity by tailoring content to the top 2 influentials than by creating "average" content and incentivizing those 2 influentials to rebroadcast. This highlights the potential benefits of developing content around the interests of an organization's influential followers.

While matching content performs well with 2 individuals, we see an unexpected dip in future rebroadcasting activity (compared to seeding influential users) when we tailor the content to the top 5 or 10 influentials.[15] That is, while influentials may be more likely to rebroadcast content designed for them, subsequent message propagation by the masses, with whom the message may or may not resonate, might be less likely. In other words, the strategy of tailoring content to influential users has its limits.

8.3. Comparing content design with seeding influentials

Next, we assess the relative value of content tailoring versus influential seeding strategies by comparing the rebroadcasting activity under two hypothetical user bases: (1) the actual (and heterogeneous) audience base observed in our data and (1) an audience base with a particular interest in Topic 2. Our results (see Table 10) suggest that when faced with a heterogeneous audience base, the seeding strategy that targets influential will always dominate content manipulation. However, when facing a homogeneous audience base, content manipulation may also be effective. In this setting, content is an important determinant in how much rebroadcasting activity occurs.

9. Conclusion

In this paper, we jointly model the drivers of social media rebroadcasting behavior. We develop a modeling framework that allows us to identify the role that content, users, the fit between content and users, and social influence play in the decisions of whether and when to rebroadcast a social media message. Our results suggest that each of these drivers plays an important role in affecting rebroadcasting behaviors. Our analysis provides evidence that rebroadcasting activity depends on the content of the message. The results also suggest that active rebroadcasters tend to rebroadcast messages with certain content, have many followers (conditional on the ratio of followers to following), and tend to more frequently broadcast messages. Furthermore, our analysis provides evidence that individuals whose profiles closely fit a given message are more likely to rebroadcast it compared to other messages. In addition, we probabilistically determine the potential existence of a limited number of influentials whose rebroadcasting is related to subsequent rebroadcasting by others, underscoring the likely importance of targeting individuals to increase the reach of a social media message.

Our findings make a number of important contributions. First, we jointly consider the role of message content, the role of influentials, and the interplay between content and individuals in rebroadcasting behavior. Specifically, our results suggest that targeting influentials to encourage their rebroadcasting of our message can potentially lead to greater rebroadcasting activity than investing in message content. However, our analysis also provides evidence that, under certain circumstances, tailoring message content to the interests of the influentials can generate even greater rebroadcasting activity. Second, we provide potentially-novel findings on the impact of content-user fit in social media rebroadcasting. While the literature has suggested that certain types of content are on average more viral, it may be critical for managers to ensure that this content fits the preferences of their audience base. Content-user fit can possibly be especially impactful for homogenous followers, for which managers could tailor content to the followers' preferences rather than simply disseminate viral content.

However, our research is not without its limitations. First, similar to other empirical work examining social effects, we do not directly observe influentials affecting others. Instead, we are only able to probabilistically assess those who may facilitate ongoing rebroadcasting. In addition, our definition of influence is limited. We identify influentials in terms of their ability to facilitate the spread of social media. Future research may also look at influence in terms of one's impact on brand health measures (e.g., Schweidel & Moe,

[15] We clarify that we are not seeding or incentivizing these influentials to immediately rebroadcast here. Instead, we simply develop content that matches their interest in the hope that they organically rebroadcast without firm-level incentives. Thus, the dip stems from content that may not appeal to the masses.

2014), the purchasing behavior of others (e.g., Ho, Li, Park, & Shen, 2012), or context specific decisions (e.g., for clothing, for books, or for restaurants). Second, the baseline group of users could be different from the users who retweet due to selection. Future work could randomize treatments that encourage retweet activity. They could also include certain fixed effects to account for heterogeneity across brands. Third, our text analysis only generates overarching topics and does not identify emotional content. Future work could develop better methods to detect emotion, especially in noisy data such as that in Twitter. Fourth, our findings follow from specific modeling assumptions in the context of a specific dataset on business schools. Given that this is one of the first findings on the topic of content and influence in rebroadcasting, future research could evaluate the robustness of these findings across different industries and with different model specifications (e.g., other types of models such as a conditional logit model). We encourage such work, as it will enable us to identify empirical regularities that persist across contexts with respect to the importance of social influence and message content on users' consumption and production of social media.

Appendix A. Additional figures and tables

Table A.1
LDA fit statistics.

No. of topics	Harmonic mean (LMD)	Perplexity	No. of topics	Harmonic mean (LMD)	Perplexity
2	−2,036,648	343	28	−1,186,129	235
3	−1,896,277	335	29	−1,173,089	234
4	−1,796,116	328	30	−1,156,355	231
5	−1,710,188	321	31	−1,149,015	228
6	−1,648,375	316	32	−1,139,397	226
7	−1,601,848	310	33	−1,131,829	223
8	−1,560,814	305	34	−1,128,524	223
9	−1,525,801	300	35	−1,123,445	222
10	−1,492,540	295	36	−1,117,405	219
11	−1,450,975	290	37	−1,102,084	216
12	−1,433,000	285	38	−1,099,104	215
13	−1,402,980	283	39	−1,092,645	214
14	−1,383,978	279	40	−1,089,764	212
15	−1,354,457	274	41	−1,087,208	211
16	−1,340,627	271	42	−1,084,192	210
17	−1,327,496	267	43	−1,073,434	207
18	−1,304,850	263	44	−1,071,335	207
19	−1,287,914	260	45	−1,067,690	207
20	−1,277,280	257	46	−1,061,602	203
21	−1,257,982	254	47	−1,051,965	201
22	−1,246,894	250	48	−1,053,226	201
23	−1,229,140	248	49	−1,047,412	199
24	−1,212,383	244	**50**	**−1,032,194**	197
25	−1,210,077	242	**51**	−1,037,959	**194**
26	−1,187,857	240	52	−1,038,851	196
27	−1,183,961	237	53	−1,043,347	195

Bolded values indicate the best fitting model with respect to the number of topics.

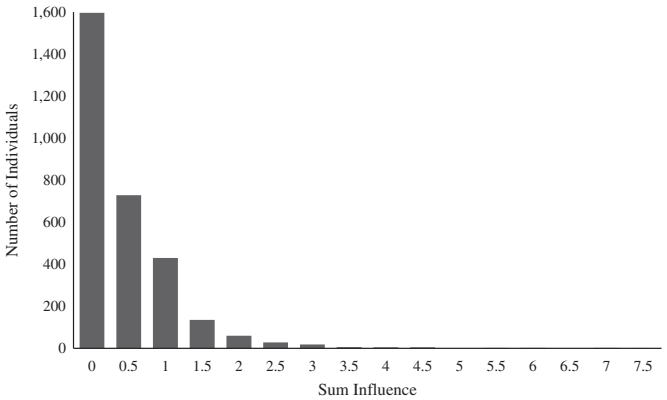

Fig. A.1. Distribution of $\sum_u [\gamma_u \times 1(t_{uj} < \tau)]$.

Appendix B. Sampling details

In this Appendix, we describe our sampling procedures for both γ_u and α_i (representing the influence of user u and the susceptibility of individual i, respectively). Our approach mirrors that used by Trusov et al. (2010).

For each user u, we draw γ_u from a Bernoulli distribution with probability π_u. If $\gamma_u = 1$, user u is considered influential and if $\gamma_u = 0$, user u is considered non-influential. In each iteration of the sampling procedure, the probability of being influential (π_u) is updated as follows:

$$\pi_u = \frac{p_u \cdot \prod_{j\in\{Y_{ij}=1\}} \prod_{i\in\{t_{ij}>t_{uj}\}} L\left(y_{ij}, t_{ij}|\gamma_u = 1\right)}{p_u \cdot \prod_{j\in\{Y_{uj}=1\}} \prod_{i\in\{t_{ij}>t_{uj}\}} L\left(y_{ij}, t_{ij}|\gamma_u = 1\right) + (1-p_u) \cdot \prod_{j\in\{Y_{uj}=1\}} \prod_{i\in\{t_{ij}>t_{uj}\}} L\left(y_{ij}, t_{ij}|\gamma_u = 0\right)}$$

where p_u is a draw from a Beta distribution representing the proportion of the population that consists of influentials with $\gamma_u = 1$. This population distribution effectively provides a prior for each user's π_u estimate.

$$p_u \sim Beta\left(1 + \sum_{U\setminus u} \gamma_k, 1 + n - \sum_{U\setminus u} \gamma_k\right)$$

where $U|u$ denotes the set of all users except user u. Our initial prior is specified as Beta(1,1).

We generate estimates for α_i in a similar manner. Again, for each user i, we draw α_i from a Bernoulli distribution with probability ϕ_i which represents the probability that user i is susceptible to influence. In each iteration, this probability is updated as follows:

$$\phi_i = \frac{q_i \cdot \prod_j L\left(y_{ij}, t_{ij}|\alpha_i = \alpha\right)}{q_i \cdot \prod_j L\left(y_{ij}, t_{ij}|\alpha_i = \alpha\right) + (1-q_i) \cdot \prod_j L\left(y_{ij}, t_{ij}|\alpha_i = 0\right)}$$

where q_i is a draw from a Beta distribution representing the proportion of the population that is susceptible to influence ($\alpha_i = \alpha$). This distribution is updated as follows:

$$q_i \sim Beta(1 + m, 1 + n - m)$$

where m represents the number of users, excluding i, for whom $\alpha_i = \alpha$. Again, this provides a population prior for the user-specific estimate ϕ_i. We specify the initial prior to be $q_i \sim Beta(1,1)$.

Appendix C. Posterior checks

We provide posterior checks to assess how well the model recovers key statistics of the data. Using the last 1000 iterations from the posterior estimates, we simulate predicted rebroadcasting activity across the sample of individuals (similar to the simulation procedure described on p. 34 but over the duration of the observation period). This provides us with a predicted individual-message outcome for each observation. This also allows us to predict the total number of rebroadcasts for each message (denoted "message level in Table C.1) and the total number of rebroadcasts for each individual (denoted "individual level" in

Table C.1). Table C.1 compares the actual rebroadcast at the message, individual, and individual-message level with that forecasted by the posterior estimates through simulation. We also plot the distributions in Figs. C.1, C.2, and C.3. The results suggest that the forecasted rebroadcasting activity is similar to the actual data.

Table C.1
Distribution of retweeting activity for actual data and simulated data.

	Actual data			
	Mean	Std dev	Min	Max
Message level	1.018	1.660	0	18
Individual level	1.383	2.246	0	69
Individual-message level	0.004	0.065	0	1
	Posterior based simulations			
	Mean	Std dev	Min	Max
Message level	1.058	1.177	0.018	15.39
Individual level	1.437	1.382	0.060	46.75
Individual-message level	0.004	0.007	0	0.504

*Note. This table compares the distribution of retweeting activity across messages, across individuals, and across individual-message observations between the actual data and data generated based on simulations from the posterior estimates.

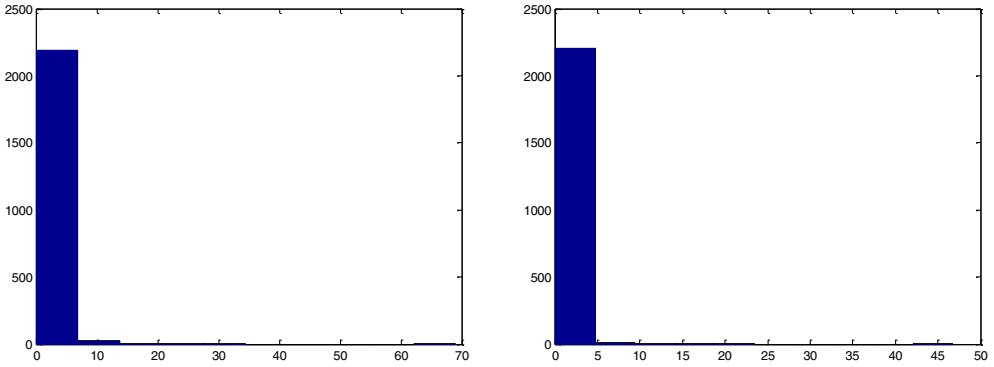

Fig. C.1. Distribution of retweeting activity for actual data (left panel) and simulated data from posterior (right panel) across individuals.

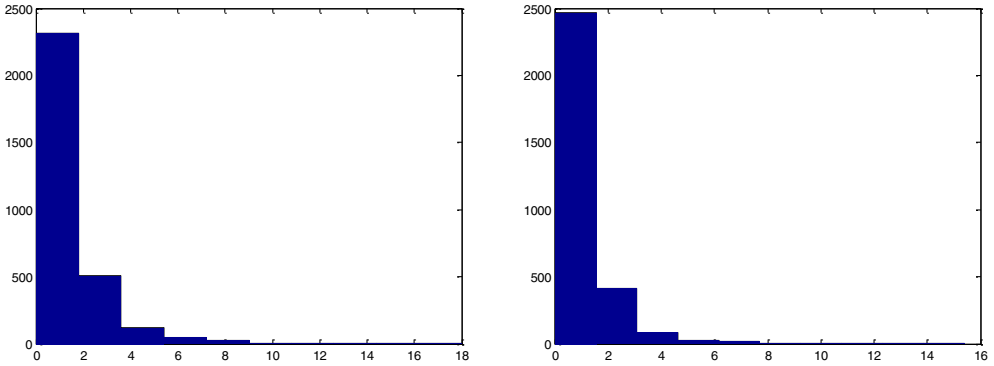

Fig. C.2. Distribution of retweeting activity for actual data (left panel) and simulated data from posterior (right panel) across messages.

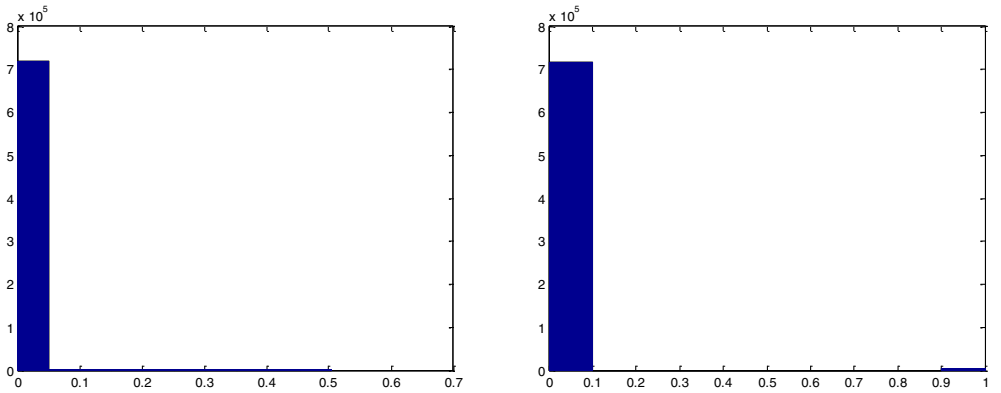

Fig. C.3. Distribution of retweeting activity for actual data (left panel) and simulated data from posterior (right panel) across individual-message observations.

References

Anger, I., & Kittl, C. (2011). Measuring influence on twitter. *Proceedings of the 11th international conference on knowledge management and knowledge technologies*. New York, New York, USA: ACM Press.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, *337*(6092), 337–341.

Bass, F. M. (1969). A new product growth for model consumer durables. *Manag. Sci.*, *15*(5), 215–227.

Bell, D., & Song, S. (2007). Neighborhood effects and trial on the internet: Evidence from online grocery retailing. *Quant. Mark. Econ.*, *5*(4), 361–400.

Berger, J. (2011). Arousal increases social transmission of information. *Psychol. Sci.*, *22*(7), 891–893.

Berger, J., & Milkman, K. (2012). What makes online content viral? *J. Mark. Res.*, *49*(2), 192–205.

Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *J. Mark. Res.*, *48*(5), 869–880.

Bhatia, T., & Wang, L. (2011). Identifying physician peer-to-peer effects using patient movement data. *Int. J. Res. Mark.*, *28*(1), 51–61.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

De Bruyn, A., & Lilien, G. L. (2008). A multi-stage model of word-of-mouth influence through viral marketing. *Int. J. Res. Mark.*, *25*(3), 151–163.

East, R., Hammond, K., & Lomax, W. (2008). Measuring the impact of positive and negative word of mouth on brand purchase probability. *Int. J. Res. Mark.*, *25*(3), 215–224.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organizations: A review of the literature. *Int. J. Inf. Manag.*, *20*(1), 17–28.

Haenlein, M. (2013). Social interactions in customer churn decisions: The impact of relationship directionality. *Int. J. Res. Mark.*, *30*(3), 236–248.

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *J. Pers. Soc. Psychol.*, *81*, 1028–1041.

Ho, T. -H., Li, S., Park, S. -E., & Shen, Z. -J. M. (2012). Customer influence value and purchase acceleration in new product diffusion. *Mark. Sci.*, *31*(2), 236–256.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Mark. Sci.*, *30*(2), 195–212.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, WebKDD/SNA-KDD '07, New York, NY* (pp. 56–65). USA: ACM.

Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *J. Mark. Res.*, *48*(3), 425–443.

Kirmani, A. (2009). The self and the brand. *J. Consum. Psychol.*, *19*(3), 271–275.

Lambrecht, A., Tucker, C., & Wiertz, C. (2015). Advertising to early trend propagators? Evidence from twitter. *Working paper*.

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *J. Mark. Res.*, *48*(5), 881–894.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global twitter heartbeat: The geography of twitter. *First Monday* (April 2013).

Libai, B., Muller, E., & Peres, R. (2013). Decomposing the value of word of mouth seeding programs: Acceleration vs. expansion. *J. Mark. Res.*, *50*(2), 161–176.

Nair, H. S., Manchanda, P., & Bhatia, T. (2010). Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *J. Mark. Res.*, *47*(5), 883–895.

Netzer, O., Ronen, F., Jacob, G., & Moshe, F. (2012). Mine your own business: Market-structure surveillance through text mining. *Mark. Sci.*, *31*(3), 521–543.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc.*, *56*(1), 3–48.

Porter, M. (2006). The English (Porter2) stemming algorithm. Accessed January 10, 2012, Available at: http://snowball.tartarus.org/algorithms/english/stemmer.html

Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *J. Mark. Res.*, *51*(4), 387–402.

Shriver, S. K., Nair, H. S., & Hofstetter, R. (2013). Social ties and user-generated content: Evidence from an online social network. *Manag. Sci.*, *59*(6), 1425–1443.

Sinha, R. K., & Chandrashekaran, M. (1992). A split hazard model for analyzing the diffusion of innovations. *J. Mark. Res.*, *29*(1), 116–127.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc.*, *64*(4), 583–639.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. *Social computing (SocialCom), 2010 IEEE second international conference, Minneapolis, MN* (pp. 177–184).

Symonds, M. (2013). Business schools social media ranking 2013 - Do you like, follow and subscribe? Accessed May 1, 2015, Available at: http://www.forbes.com/sites/mattsymonds/2013/09/30/business-schools-social-media-ranking-2013-which-schools-do-you-like-follow-and-subscribe-to/

Sysmosos (2010). Replies and retweets on twitter. Accessed February 1, 2015, Available at: http://sysomos.com/inside-twitter/twitter-retweet-stats

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Mark. Res.*, *51*(4), 463–479.

Top Business School Rankings: MBA, Undergrad, Executive & Online MBA – Businessweek (2011O). Accessed December 1, 2011, Available at: http://www.businessweek.com/bschools/rankings

Toubia, O., & Stephen, A. T. (2013). Intrinsic versus image-related motivations in social media: Why do people contribute content to twitter? *Mark. Sci.*, *32*(3), 368–392.

Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining influential users in internet social networks. *J. Mark. Res.*, *47*(4), 643–658.

Van den Bulte, C., & Joshi, Y. V. (2007). New product diffusion with influentials and imitators. *Mark. Sci.*, *26*(3), 400–421.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *J. Consum. Res.*, *34*(4), 441–458.

Zaman, T., Fox, E., & Bradlow, E. (2014). A Bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.*, *8*(3), 1583–1611.

Zhao, W. X., Jing, J., Jianshu, W., He, J., Ee-Peng, L., Hongfei, Y., & Xiaoming, L. (2011). Comparing twitter and traditional media using topic models. *Advances in information retrieval: 33rd European conference on IR research, Dublin, Ireland. Proceedings.*