INSEAD

The Business School for the World®

# Faculty & Research Working Paper

## A Comparison of the Effects of Transmitter Activity and Connectivity on the Diffusion of Information over Online Social Networks

Andrew T. STEPHEN
Yaniv DOVER
Jacob GOLDENBERG
2010/35/MKT

# A Comparison of the Effects of Transmitter Activity and Connectivity on the Diffusion of Information over Online Social Networks

Andrew T. Stephen*

Yaniv Dover**

Jacob Goldenberg***

April 29, 2010

Working paper—please do not cite without permission

The authors thank Yakov Bart, Don Lehmann, Moshik Miller, Danny Shapira, Olivier Toubia, and participants in the INSEAD Technology-Operations Management-Marketing brown bag for their helpful comments; Twitter and BackType for allowing access to data and assisting with data collection; and the INSEAD Alumni Fund for financial support. Andrew T. Stephen is the corresponding author.

* Assistant Professor of Marketing at INSEAD, Boulevard de Constance, 77305 Fontainebleau, France Ph: (33) 01 60 72 26 42 Email: andrew.stephen@insead.edu

** Doctoral Candidate in Marketing at the School of Business Administration, Hebrew University of Jerusalem, Israel. Email: yanivd@phys.huji.ac.il

*** Professor of Marketing at the School of Business Administration, Hebrew University of Jerusalem, Israel. Email: msgolden@huji.ac.il

**A Comparison of the Effects of Transmitter Activity and Connectivity on the Diffusion of Information over Online Social Networks**

This paper examines how observable and measurable characteristics of the people who originally transmit information in online social networks affect how far that information spreads. Two characteristics are compared: a transmitter's connectivity (how well connected they are in the network) and activity (how frequently they transmit information over their social ties in the network). Despite extensive past research on connectivity (e.g., the literature on hubs), the role played by activity in driving diffusion is largely unexplored. Across three studies (an experiment, a simulation, and an empirical analysis of link sharing in Twitter) the authors find that (1) a person's transmission activity positively influences diffusion, (2) people who are exceptionally frequent content transmitters—pumps—have a large positive effect on information diffusion, (3) when comparing the activity effect (cf. pumps) to that of connectivity (cf. hubs), activity is at least as strong a driver as connectivity, if not more under a variety of realistic conditions, and (4) the transmitter activity effect on diffusion holds even after controlling for the information's quality and breadth of appeal.

Keywords: diffusion, online networks, social media, social networks, word-of-mouth.

The spread of the Internet has led to a colossal quantity of information posted and shared by people through social media such as forums, blogs, and online social networks. A fast-growing trend among users of online social networks is to use them for sharing information, which often includes referrals or links to content on the web. In the case of Twitter, for example, an increasingly common use is for posting (or "tweeting") links (URLs) to content elsewhere on the Internet (e.g., a video on YouTube, a news article on the New York Times website). This is so common that, according to Twitter, a link to a New York Times article is shared every four seconds over their network. Underscoring the ubiquity of link sharing in networks such as Twitter, a recent Yahoo study of approximately 10 million tweets in July 2009 found that 1.8 million (18%) contained URLs (Singh 2009). In the case of Facebook, as of early 2010, more than 25 billion pieces of content (e.g., photos, videos, links) were shared each month through Facebook (Facebook 2010). Clearly, information sharing is a hallmark of social media. As a result, shared links have become significant traffic sources for many blogs and websites (including mainstream media outlets such as CNN and the New York Times). In some cases traffic to major websites coming from Facebook, Twitter, and other social media sites exceeds traffic coming from Google (Hopkins 2009).

Despite the prevalence of information sharing in social media, very little is understood about what factors might affect the diffusion of this information over these online social networks. While diffusion has been extensively examined in the marketing literature for decades, information sharing through social media and over online social networks is a new and important context that has received scant attention in extant research. A key difference between diffusion of, for example, new products in consumer or industrial markets, and diffusion of digitized information over online social networks is the underlying social transmission process. The micro-level process that drives macro-level information diffusion outcomes in online social networks is more complex than most of those examined

in extant literature, requiring people to *transmit*, *consume*, and *retransmit* (i.e., pass on)[1] information for it to spread. Specifically, (1) a person must bring content into a network from "outside" (e.g., a news website or another network) and *transmit* it over their online social ties (e.g., posting a link on one's Facebook page or tweeting it through Twitter), (2) their contacts are then *exposed* to the content (e.g., seeing a link to a news article on a friend's Facebook page or in their Twitter feed), (3) these "receivers" then decide whether or not to *consume* the content (e.g., by clicking the link to read the news article), and (4) they also decide whether or not to *retransmit* or distribute the content by sharing it with others (e.g., by "retweeting" it to their Twitter followers).

The requirement that information be explicitly retransmitted distinguishes the social transmission process for information sharing in online social networks from other WOM and contagion processes studied in marketing (e.g., Bass 1969; Goldenberg, Libai, and Muller 2001; Watts and Dodds 2007), and in other fields such as physics and sociology (e.g., Coleman, Katz, and Menzel 1957; Dodds and Watts 2004). Typical in the diffusion models in prior work is the assumption that information spreads to potential adopters simply by them being exposed to (or connected to) past adopters. Requiring a person to retransmit information adds a critical extra step to the process. A deeper understanding of this process is therefore needed. A major aim of this paper is to shed light on the micro-level processes that give rise to macro-level diffusion outcomes for information shared through online social networks.

Many potential drivers of diffusion can be considered, and here we concentrate on easily observable and objectively measurable behavior-based attributes of the people who first introduce (transmit) information into these networks. We concentrate on behavior-based

[1] Retransmit means that a person who received information passes it on to another person. E.g., person A transmits information to person B. Then, if person B passes this on to person C we say that person B has "retransmitted" the information. Retransmit does not mean that a single person repeatedly transmits the same information repeatedly.

person characteristics because marketers can use these for targeting purposes where they have less control over other potential diffusion drivers such as information content (which is often user-generated and therefore not within a marketer's control), and also since past diffusion research has often focused on such factors (e.g., the literatures on hubs, opinion leaders, and mavens). Specifically, we compare two major transmitter-related diffusion drivers: a person's network *connectivity* (e.g., how many followers do they have in Twitter or friends in Facebook?), and their transmission *activity* (e.g., how frequently do they post new tweets or status updates?). While connectivity has been the subject of much discussion (and debate) in past literature (e.g., research on hubs), activity, to the best of our knowledge, has not been previously investigated.

To preview our main results, we find that (1) a person's transmission activity positively influences diffusion, (2) people who are exceptionally frequent content transmitters—pumps—have a large positive effect on information diffusion, (3) when comparing the activity effect (cf. pumps) to that of connectivity (cf. hubs), activity is at least as strong a driver as connectivity, if not more under a variety of realistic conditions, and (4) the transmitter activity effect on diffusion holds even after controlling for the information's quality and breadth of appeal.

## *THEORY AND RESEARCH QUESTIONS*

Our objective is to examine how transmitter activity and connectivity compare as drivers of diffusion in the context of information shared over online social networks. Put simply, is a transmitter's connectivity or their activity a better predictor of how far a piece of information they post online in, for instance, Facebook or Twitter, will spread? Past research on diffusion and related work on word-of-mouth (WOM) and consumer-to-consumer social interactions has focused mostly on how diffused information (e.g., online reviews as a type of

online WOM) influences consumer behavior and aggregate marketing outcomes such as sales (e.g., Chevalier and Mayzlin 2006; Dellarocas, Zhang, and Awad 2007; Eliashberg et al. 2000; Reichheld and Teal 1996; see Libai et al. 2010 for a recent review of research on consumer-to-consumer interactions). We are interested not in how diffused information affects such outcomes, but rather on the spread of information itself and what drives this diffusion.

A number of factors conceivably can influence the probability that a piece of information or content will diffuse widely, falling into categories related to the item itself (e.g., how interesting or topical is the information? E.g., Berger and Milkman 2010), the source of the item (e.g., how credible is the source?), and the social network over which it spreads (e.g., how centrally located in the network is the seed person, or how dense is the social network? E.g., Goldenberg et al. 2009; Katona, Zubcsek, and Sarvary 2009; Watts 2002). Although many of these factors likely combine and interact to drive social epidemics and widespread diffusion of information, we focus on network-related factors associated with the transmitter. We concentrate on certain transmitter behaviors because receivers (i.e., people to whom information is transmitted) may take these into account when deciding whether or not to retransmit a given piece of information (e.g., how trustworthy is this person as a source of information that is worth passing on?), irrespective of the nature of the information itself or its perceived quality.

We focus on two kinds of easily observable and objectively measurable individual-level transmitter behaviors that may influence how information spreads over online social networks: *connectivity* and *activity*. Connectivity is a direct function of the network's structure and refers to how centrally positioned a person is in the network. Various measures of connectivity (based on graph theoretic analysis of network structure) are available, with the most common and straightforward being a person's degree (number of social ties a person

has).[2] Degree and other measures based on network structure have attracted a lot of attention in the literature; on the other hand, activity, which is not a direct function of network structure, has attracted sparse attention so far. Activity refers to how frequently a person transmits information or posts messages in an online social network.

**Connectivity and Diffusion**

Transmitter degree has been shown to affect diffusion processes. For example, Goldenberg et al. (2009) studied the effects of hubs (people with exceptionally high degree) on the diffusion of virtual goods in a South Korean online social network, and found that a hub adopting a good positively affected the extent and speed of diffusion. However, in their study the focus was on product adoption where adoptions were publicly observable. A positive relationship between degree and extent of diffusion is plausible in such cases because the process primarily depends on exposure or awareness.[3] If the aim is to maximize reach (i.e., increase exposures and raise awareness) then this is reasonable. However, if information needs to be explicitly passed on then degree may not be the only (or best) criterion for selecting seeds.

As noted above, the information sharing process in online social networks is more complex and involves more steps than simply exposing people to a piece of information and having them adopt it with some nonzero probability. Thus, we cannot automatically assume that degree will be a dominant diffusion driver. Indeed, the importance of degree in driving information cascades and the spread of public opinion has been questioned by Watts and Dodds (2007), who argue that cascades occur not because of so-called "influentials" (their term for hubs) but because of the existence of a critical mass of easily influenced people on

[2] Other measures used in past literature are also based on network structure, and include clustering, betweenness, closeness, proximity, and eigenvector centrality. See Stephen and Toubia (2010) for examples, and Van den Bulte and Wuyts (2007) for a more general discussion of node-level measures that can be computed from network graphs.

[3] This situation is typical in many diffusion models, including in the Bass (1969) model and other commonly used models such as Goldenberg et al. (2001).

the receiving end of transmissions. Despite some controversy over the model used by Watts and Dodds (2007), they nevertheless draw attention to the possibility that a transmitter's connectivity may be neither the only nor the most critical driver of diffusion outcomes, particularly when information (not products) is what is spreading over social ties.

**Activity and Diffusion**

We hypothesize that a person's activity is a valid predictor of their contribution to widespread information diffusion in online social networks. We now explain the rationale behind this hypothesis.

Recall that explicit retransmissions are critical for information being shared over social networks to spread widely. What would make a person who receives some information over a social tie (e.g., they see it on a friend's Facebook page) more or less likely to retransmit it? Research into the characteristics of items that makes them more likely to catch on, diffuse, and be talked about suggests that more provocative, exciting, surprising, novel or even awe-inspiring items tend to spread more (e.g., Berger and Heath 2005; Berger and Milkman 2009; Heath, Bell, and Sternberg 2001; Rogers 2003). People have an inherent desire for novelty (Hirschman 1980; Rogers and Shoemaker 1971), and are motivated to transmit WOM by a need to be listened to by others (Engel, Blackwell and Miniard 1993; Hennig-Thurau et al. 2004; Stephen and Lehmann 2010; Sundaram, Mitra and Webster 1998). Therefore, digital media content and information shared in online social networks should be more likely to be passed on if people perceive it to be more novel or fresh (i.e., apparently current and/or new). Put differently, people should be generally more likely to pass on information when they believe that others will not have already seen it. While it is easy for a person to judge the freshness of some online media content (e.g., a big news story published on a certain date), this is rarely the case given the sheer volume of available content. For example, a video on YouTube might have been viewed a million times, but this

does not mean that a person's friends have seen it or that a person knows how many of their friends have seen it.

We posit that in these situations people look at *who* exposed them to the information (i.e., transmitters) and form judgments about the information based on transmitter characteristics. This is where we believe a transmitter's activity plays a role. Consider the following example. Suppose that a Twitter user follows two people, one who has higher activity and tweets frequently (e.g., a few times a day on average), and one who has lower activity and tweets infrequently (e.g., about once every other week). We hypothesize that, compared to the less active person, the more active person will be perceived as having fresher information. This is because receivers may infer from a person's activity how much information they have to give and people who frequently "pump" out information are presumably doing it because they have something to say and want others to pay attention (Stephen and Lehmann 2010). We predict that people look at how active a transmitter is and use this as a heuristic for judging how fresh information is likely to be. Then, if they judge the information to be sufficiently fresh, it is therefore worth retransmitting and they pass it on to their contacts. (We test this in study 1.)

A transmitter's degree is not expected to have the same effect on retransmission and, when aggregated, diffusion. When the network is undirected (i.e., if A → B then B → A; like Facebook), a transmitter with high degree has a big audience and is exposed to many other people. When the network is directed (i.e., if A → B it is not necessarily the case that B → A; like Twitter), having high degree means either they have a large audience (high out-degree) or are exposed to many others (high in-degree), but not necessarily both (although the correlation between in- and out-degree in directed online networks, such as the WWW, tends to be positive; e.g., Liu, Dang, Wang, and Zhou 2006). In terms of out-degree (audience size), there is no reason why information from a high out-degree transmitter will have a

higher *individual-level* retransmission probability than information from a low out-degree transmitter. While broad awareness or exposures will increase, it is not clear that retransmission probabilities will. In fact, people may be *less* inclined to pass on information from a hub because they know that, by definition, lots of other people will have also received the same information, thus making it less fresh and novel, and therefore less attractive to retransmit. In terms of in-degree (i.e., breadth of exposure to others), although a person with high in-degree has access to many other social information sources, this information will not necessarily be perceived as fresh, maybe because of the multiple appearances in the same instance.

**Research Questions and Overview of Studies**

We address three research questions: (1) What inferences do people draw about shared information based on how active and well connected transmitters are? (2) What are the relative effects of transmitters' activity and connectivity on the probability that a person exposed to information from them will retransmit it to others? And (3) what are the relative effects of transmitter activity and connectivity on the extent of information diffusion in online social networks? All questions (and all three studies in this paper) center on retransmission, which, as discussed above, is essential for information to diffuse over online social networks like Twitter, Facebook or even through "old fashioned" email forwarding. We look at retransmission directly as a micro, individual-level action in study 1, and as an outcome of collective retransmissions that give rise to macro, aggregate-level diffusion in studies 2 and 3.

The first two research questions, which address micro-level psychological aspects of the underlying social contagion process, are explored with a behavioral experiment (study 1) where we test how people perceive information from transmitters who differ according to their activity and connectivity, and how likely participants would be to retransmit this information. The third question, which focuses on aggregate, macro-level information

diffusion, is examined first with an agent-based model (study 2) where we build an individual-level model for sharing and retransmitting information and simulate diffusion over large social networks where people vary in terms of connectivity and activity. Then we further address how activity and connectivity affect diffusion with an empirical analysis of data on link sharing in Twitter (study 3).

## *STUDY 1: AN EXPERIMENTAL TEST OF THE MICRO-LEVEL EFFECTS OF TRANSMITTER ACTIVITY AND CONNECTIVITY ON INTENTIONS TO RETRANSMIT*

### Overview and Experiment Design

One hundred and eight participants from a large panel were recruited for this experiment, for which we used Twitter as an online social network context. All participants were prequalified as current users of Twitter (i.e., they gave Twitter user names that were checked to be valid). The task for participants was straightforward: they were asked to look at information about another Twitter user who they were led to believe was a real person (the "target user"), and were asked questions about retransmission and perceptions of this user and information posted by them.

We manipulated between subjects the target user's connectivity (out-degree/number of followers: low vs. high) and the target user's activity (average number of tweets per day: low vs. high) in a $2 \times 2$ full factorial design. Participants were randomly assigned to one of the four conditions. This information was presented to participants in a table that listed the target user's (1) number of followers, and (2) average tweeting rate (expressed as "1 tweet every $x$ days or about $1/x$ tweets per day").

The levels of both factors were calibrated using real data from a random sample of approximately 2,500 Twitter users (from the same dataset used in study 3). Both degree and

average tweeting rate were heavily skewed, long-tail distributions (approximately power-law, which is typical of many node-level measures in online and social networks; e.g., Barabási and Albert 1999; Stephen and Toubia 2009). For low (high) levels of the connectivity and activity we respectively took the mean number of followers and mean average daily tweeting rate from the first (fourth) quartiles of the respective variables in the data. For connectivity: low = 6 followers, and high = 693 followers. For activity: low = .07 tweets/day, and high = 12 tweets/day (note that high-low ratios are of the same order of magnitude for these two factors).

We asked participants to suppose that they followed the target user and that this user "posted a tweet that contained a link (URL) to some content on the Internet" and that they noticed this in their feed. We deliberately did not provide actual information or content to remove the potential for information characteristics affecting the dependent measures (we leave this for study 3). We measured (1) the likelihood they would click this link to view the content (0 to 100% scale), (2) the likelihood they would share this link with their Twitter followers by retweeting it (0 to 100% scale), and (3) their perceptions of the target user and the information they shared (10 scale items, each 1 = "strongly disagree" to 7 = "strongly agree"). The items are listed in Table 1 and were presented to participants in a randomized order. A factor analysis of the 10 items revealed three factors that explained 82.5% of the variance. The three factors reflect participants' expectations of the information quality of the target user's tweets ("quality"), perceived speed at delivering new information ("speed"), and perceived trustworthiness as a source of information ("trust"). Three reliable composite items were created by taking the means of the items ($\alpha$ = .94 for "quality," $\alpha$ = .88 for "speed," and $\alpha$ = .81 for "trust").

[INSERT TABLE 1 ABOUT HERE]

**Results**

*Manipulation checks.* After the dependent variable and process measures, we asked participants to rate the target user on seven-point bipolar scales with respect to how they thought this user compared to other Twitter users who are regular people (as opposed to celebrities, news organizations, companies, etc) in terms of degree (1 = "has many fewer followers than the average user" to 7 = "has many more followers than the average user"), activity (1 = "is less active in tweeting than the average user" to 7 = "is more active in tweeting than the average user"), and in general (1 = "is very rare, these characteristics are very uncommon" to 7 = "is very typical, these characteristics are very common"). To confirm that our activity and connectivity manipulations worked as intended, we expected that, compared to participants in the low activity (connectivity) condition, participants in the high activity (connectivity) condition would rate the target user higher on the comparison based on activity (degree). We expected no differences between conditions on the general comparison to confirm that the target user was not considered to be particularly unusual or extreme.

We compared means of these three scales for each activity and connectivity condition using a multivariate analysis of variance. The results confirmed our expectations. For the comparison based on degree, there was a significant main effect of connectivity ($M_{low}$ = 2.37, $M_{high}$ = 5.12, $F(1, 104) = 84.86$, $p < .001$) but not activity or the interaction ($p = .18$ and $p = .65$, respectively). For the comparison based on activity, there was a significant main effect of activity ($M_{low}$ = 2.37, $M_{high}$ = 5.40, $F(1,104) = 118.13$, $p < .001$) but not connectivity or the interaction ($p = .08$ and $p = .35$, respectively). Finally, for the general comparison no differences were found ($p > .21$ for all effects).

*Effects on retransmission.* The likelihood of retransmission variable was compared across activity and connectivity conditions using an analysis of covariance, with the probability of clicking on the link to view (i.e., consume) the referred content as a covariate,

Pr(view). Put simply, we examined how the target user's activity and connectivity affect the retransmission probability conditional on intentions to view/consume content.

Least-squares means for retransmission probabilities (adjusted for the Pr(view) covariate) are plotted in Figure 1. Activity had a significant main effect on the conditional probability of retransmission ($F(1, 104) = 4.71$, $p = .03$). Connectivity, however, had no effect ($F(1, 104) = .73$, $p = .40$), and neither did the interaction ($F(1, 104) = .45$, $p = .51$). The effect of the covariate was positive and significant, as would be expected ($F(1, 104) = 18.40$, $p < .001$). Note that the same pattern of results was found if we excluded the Pr(view) covariate from the analysis.

[INSERT FIGURE 1 ABOUT HERE]

The effect of activity but not connectivity on the intended conditional probability of retransmission is consistent with our suggestion that receivers are more likely to pass on information from transmitters who are more active and "pump out" information more frequently than others. We found no evidence to suggest that connectivity affected retransmission.

*Mediation analysis.* In our earlier discussion about activity and how it might affect retransmission behaviors we posited that information shared by more active transmitters might be perceived as fresher and more novel than information shared by less active transmitters, and that this could explain a positive effect of activity on retransmission. We tested this hypothesized process using the composite items based on the 10 perception scale items. We performed a standard mediation analysis following Baron and Kenny's (1986) procedure, and report the results for the separate regression models in Table 2 for "quality," "speed," and "trust" as potential mediators of the effect of activity on retransmission.[4]

[INSERT TABLE 2 ABOUT HERE]

[4] Note that this analysis is not intended to fully describe the underlying process (which is beyond the scope of the current paper), but rather to offer some process-based support for our main arguments.

Our hypothesis suggests that "speed" in particular mediates the effect of activity on retransmission. Consistent with this, we found evidence of complete mediation of the effect of activity on retransmission through speed (Sobel test: $Z = 2.39$, $p = .02$). When the target user has higher activity, participants believed that transmitted information from the target user would be fresher, more novel, and delivered faster, which in turn increased their stated conditional probability of retransmitting information from the target user. We also found support for "trust" in the transmitter as a reliable, trustworthy source of information as a second mediator (Sobel test: $Z = 2.17$, $p = .03$). The perception of the information quality ("quality"), however, did not mediate the activity → retransmission effect (Sobel test: $Z = 1.00$, $p = .32$).[5] Note, however, that we would generally expect quality of content to drive retransmission decisions, but we deliberately did not provide content (and hence no manipulation of content quality) in this study so as to focus entirely on transmitter activity and connectivity. We consider the effect of content quality on diffusion in study 3 when we examine the actual spreading of links in Twitter.

## *STUDY 2: AN AGENT-BASED MODEL OF THE EFFECTS OF TRANSMITTER ACTIVITY AND CONNECTIVITY ON DIFFUSION*

### Overview

Study 1 demonstrated that at the individual level a transmitter's activity (but not connectivity) can positively influence retransmission. We now move to considering macro-level aggregate diffusion consequences of retransmission behaviors. In this study we examine how transmitter activity and connectivity affect aggregate diffusion in a social network. The individual, micro-level social transmission process is based on the process described above.

[5] These findings were based on separate sets of OLS regressions, as reported in Table 2. Regressions with two or more mediators was not possible due to multicollinearity concerns given the moderate to high positive correlations between the composite variables: (speed, trust) = .47, (speed, quality) = .50, and (trust, quality) = .69.

Our focus is on the relative effects of activity and connectivity on macro diffusion outcomes. We use agent-based modeling (ABM) methods that are usually used when the focus is on the collective dynamics of a system that occur as a result of individual behaviors (Lusch and Tay 2004; Rand and Rust 2009). While we report the results of an empirical analysis of link sharing in Twitter in study 3, the full collective dynamics of these complex systems and social interactions cannot be fully understood with actual diffusion data (cf. Garber et al. 2004). Hence, to more rigorously analyze and compare transmitter activity and connectivity in driving the dissemination of shared information in online social networks we take a simulation-based approach.

**Model**

We use a network with *N* nodes and *E* undirected ties between pairs of nodes. Each node (indexed by *i*) represents a person and each tie can be thought of as a "friendship" in an online social network. We endow each node with two characteristics: degree (connectivity) and transmission delay (which is used to represent activity; see below for details).

*Degree.* Denoted by $k_i$, this is the number of ties connecting node *i* to other nodes (and $E = \frac{1}{2}\sum_{i=1}^{N} k_i$). We assume that the social network has the common scale-free property exhibited in many offline and online networks (Barabási and Albert 1999), including e-commerce settings (Stephen and Toubia 2009). This means that $k_i$ is power-law distributed across the *N* nodes; i.e., $P(k) \sim k^{-\gamma}$ with scale parameter $\gamma$ (usually between 2 and 3 in real social networks; Barabási and Albert 1999). This distribution ensures the presence of nodes with exceptionally high degrees.

*Transmission delay*. The notion of activity in this model is operationalized in terms of the speed with which a person transmits or retransmits a piece of information to the people connected to them. In study 1, speed was an important factor driving micro-level retransmission behaviors, and for this reason we focus on speed in operationalizing activity in

this model. When exposed to a piece of information, each node will take some amount of time (here, measured in discrete time periods) before they retransmit that information (if they decide to retransmit the information; see below)—we call this delay, $s_i$. High (low) activity nodes will have low (high) $s_i$. Activity is defined as $a_i = \max\{s_1, s_2, ..., s_N\} - s_i$. Like degree, we assume that activity is power-law distributed (independent of degree); i.e., $P(a) \sim a^{-\lambda}$ with scale parameter $\lambda$ (assumed to be between 2 and 3, similar to degree).[6] Note, as for the degree distribution, the results we report below are robust to different assumptions on the delay distribution, including Poisson and Gaussian distributions as alternatives.

*Social transmission process*. First, a single seed node is selected (representing the person who introduces the piece of information into the network from some outside source in the form of a post (e.g., a tweet, a status update, or similar). This occurs at time $t = 0$. By introducing the information into the network the seed has transmitted the post to its $k_{seed}$ friends, who are now all exposed to the message (this is the transmission mechanism used in Twitter, Facebook, etc).

Second, the $k_{seed}$ nodes exposed to the information each independently decide whether to consume the information with probability $q$. This is equivalent to clicking on the link and viewing the content. This exposure-to-consumption step applies for all successive generations.

Third, the nodes that were exposed to *and* consumed the information then must decide whether or not to pass this information along to others. They retransmit the information to their friends with probability $r$ and, if they do, with a delay (number of periods after consumption) according to the delay distribution. If they retransmit, more nodes are exposed to the information and the process repeats itself, following the same rules in each period.

[6] Assuming a power-law distribution for activity is consistent with our belief that very few people are highly frequent (and short delay) transmitters. This also means that so-called "pumps" in this model are rare. It should be harder to find activity to be a driving force in macro diffusion outcomes if exceptionally active nodes are rare, and thus provides a more conservative test than if we assumed other distributions for activity in this ABM.

Also, consistent with traditional diffusion models used in marketing, we also allow for the possibility that nodes can be exposed to the same information contained in the post from external sources (e.g., advertising, other networks they belong to, searches for information on the web, etc). This exposure occurs with probability $p$ in each period for each node that has not already been exposed to the information.

**Results**

*Simulation setup and procedure.* We allowed this process unfold over several connected networks, varying the numbers of nodes and ties across multiple runs of the simulation for the sake of robustness (ranging from $N$ = 10,000 to 500,000 nodes, and $E$ = 100,000 to 5,000,000 ties). The results reported below qualitatively hold for all different network sizes and densities, as well as for variations on the type of network based on different degree distributions (i.e., scale-free as mentioned above, as well as Poisson and Gaussian).

The main parameters that we varied in the simulation were the seed node's activity, the seed node's connectivity (degree), the scale parameters for the distributions of activity and degree ($\lambda$ and $\gamma$, respectively; both in the empirically common 2-3 range), the probability of a node consuming the information given exposure from another node ($q$), and the probability of a node retransmitting the information given consumption ($r$).[7] We examined combinations of these parameters across a wide range of the joint parameter space (see Table 3).

[INSERT TABLE 3 ABOUT HERE]

For each simulation run with each combination of parameters, a specific node was chosen as the seed that would start the information diffusion. This node had a certain activity (delay) and connectivity (degree). A wide range of combinations of activity and connectivity

[7] We also varied the probability of consuming the information when exposed from an external source ($p$) but kept this very small so that the social transmissions within the network were always dominant drivers of diffusion.

were tested. We then allowed the social transmission process to unfold from this seed node across the entire network. After the diffusion process has ended (i.e. there were no more information consumptions) we observed the extent of diffusion or "cascade size," measured as the proportion of the nodes in the network that had *retransmitted* the information. We focused on retransmission instead of consumption since, as already discussed, in the online social networking context information must be retransmitted for it to spread widely. Therefore, the aggregate macro-level diffusion outcome of interest is reflected by how many nodes retransmit.

*Analysis.* Since multiple simulations (i.e., diffusion processes) were run for each combination of parameters and the diffusion outcome was observed each time, the output of the ABM was a dataset much like that of an experiment: multiple cells, each with multiple observations on a dependent variable of interest. We subjected this data to a regression analysis where we regressed the extent of diffusion (cascade size) for each simulation run on the seed node's activity and connectivity, as well as the other parameters as controls.

[INSERT FIGURE 2 & TABLE 4 ABOUT HERE]

Figure 2 plots the average diffusion/cascade size (vertical axis) as a function of activity and connectivity (averaging over the other parameters and simulation runs). Clearly, as both activity and degree increase, so does diffusion. The question is, what are their relative effects? Also, are there some regions of the parameter space where one dominates the other?

Column 1 in Table 4 reports standardized regression parameter estimates for the full set of simulations (using network size $N$ = 90,000, and $r$ = 1; results hold for different-sized networks and different values of $r$). Both activity and degree positively affected diffusion ($p$s < .001). The estimates for these two effects were very similar (activity .80 vs. degree .86). We investigated whether these effects were nonlinear using quadratic terms. Interestingly, while we found a diminishing marginal effect of increasing degree on diffusion (negative

degree$^2$ effect, $p < .001$), we found an increasing marginal effect of activity (positive activity$^2$ effect, $p < .001$). Compared to seeds with moderate levels of degree, seeds with exceptionally high degree (i.e., hubs) do little to increase how widely information diffuses (because of the diminishing marginal return to increasing degree). However, increasing a seed's activity is beneficial and increases their effect on diffusion (because of the increasing marginal return to increasing activity).

The non-significance of $q$ (probability of information consumption once exposed) is likely because $q$ does not have a critical effect on the final diffusion outcome but rather controls the ease of consuming information after being exposed to it ("easier" with higher $q$). However, the positive interaction of $q$ with activity ($p < .001$) suggests that the effect of activity on diffusion gets stronger as information gets easier to consume. Information is typically very easy to consume in online social networks: e.g., simply clicking a link in a tweet or status update.

These results indicate that, in general, both activity and connectivity play a role and that the positive effect of activity on diffusion is not weaker than that of connectivity. However, we do not see across-the-board evidence of activity dominating connectivity. Since the effects of activity and degree in column 1 of Table 4 are at the averages of the other varied parameters, it may be the case that under certain conditions activity dominates whereas under other conditions connectivity does. The question is whether in instances where activity dominates are the network and distribution parameters' values realistic in that they match conditions of social transmissions in real online social networks. Significant interactions between either activity or degree and the distribution scale parameters $\gamma$ (for degree distribution) and $\lambda$ (for activity distribution) suggest this may be the case. These distribution parameters control the extremes on the activity and degree distributions and the average node's activity and degree, and therefore can affect diffusion. In Figure 3 we illustrate this by

plotting the relative effect of activity versus connectivity (i.e., the ratio of the standardized regression coefficients for activity to connectivity from column 1 of Table 4) against the distribution scale exponents. Activity dominates connectivity when the ratio is above 1. Based on this figure, it appears that activity dominates connectivity in the parameter subspace in which the exponents are low (< 2.5).

[INSERT FIGURE 3 ABOUT HERE]

Is this space where activity dominates a reflection of real online social networks? With respect to the two distribution parameters, for most real networks, both human and non-human, past research finds that the average degree distribution exponent (for a scale-free network) is approximately $\gamma$ = 2.2 (Albert and Barabasi 2002). Since we are interested in online social networks and, in study 3 use data from Twitter, it pays to consider networks in our simulation that are similar to these. Computer science research on Twitter found a power-law degree distribution, thus indicating that Twitter has a scale-free network consistent with the assumption in this ABM study, with Twitter's scale-free exponent approximately $\gamma$ = 2.4 (Java et al. 2007).

Column 2 in Table 4 reports the standardized regression parameter estimates when the degree distribution exponent is $\gamma$ = 2.4; i.e., for a "Twitter-like" network structure. The advantage is now clearly on activity's side with its effect larger than degree's (.84 vs. .70). The other effects discussed above also generally hold, including, importantly, the positive interaction between $q$ and activity indicating that the effect of activity on diffusion is particularly strong and positive for things that have a higher $q$ (e.g., because it is easy to consume them once exposed).

## *STUDY 3: AN EMPIRICAL COMPARISON OF THE EFFECTS OF TRANSMITTER ACTIVITY AND CONNECTIVITY ON LINK DIFFUSION IN TWITTER*

### Overview

The previous two studies have shown that transmission activity can influence how shared information diffuses in online social networks like Facebook and Twitter, and that transmitters' network connectivity (degree) may in fact play less of a role in driving information diffusion and information sharing online than previously thought. Here we again compare transmitter activity and connectivity as diffusion drivers, this time using data on the sharing of links (URLs pointing to websites and other online content) among users of Twitter.

### Data

Our data come from Twitter. In Twitter, users post tweets that are short (up to 140 characters) text messages, which get transmitted to all of their followers (who are other users in the Twitter network). Users decide who to follow, and when and what to post in their tweets. As we mentioned above, a common use for Twitter is to share links to external online content by embedding links in tweets. In this study we focus on the diffusion of links shared through tweets across the Twitter network.

Over a 44-day period in May-June 2009 we observed the activity and network connections for a panel of 2,500 Twitter users. Due to attrition and some accounts being deactivated by Twitter during this period (e.g., due to suspicious behavior), our effective panel size was 2,461 users. The users were randomly selected and screened to ensure that none were media organizations, celebrities, companies, or any other account that was, to the best of our knowledge, not an actual person. Each day for each user we collected data on (1)

number of tweets (posts), (2) number of followers (out-degree), and (3) the text of their tweets. No other information was available on these users.[8]

*Dependent variable: link diffusion.* We focused on the spread of URLs (links) that were embedded in tweets. We do not examine the specific retransmissions of tweets themselves because they cannot be easily tracked, unless a retransmitter included the tag "RT" (for "retweet") in the new tweet (and even then, retweeted tweets are sometimes edited versions of the original; more commonly, users do not always include "RT" even though it is considered good etiquette to do so). According to Yahoo's recent analysis (Singh 2009), "RT" in tweets is rare (only 1% of tweets). This of course does not mean that Twitter users do not retransmit other users' tweets, just that it is difficult to reliably track.[9] On the other hand, tracking the spread of outside-Twitter content linked to by URLs within Twitter (tweets) is easier and more reliable (although still not straightforward). Thus, we focus on the spread throughout the Twitter network of content that exists outside of Twitter on the web (e.g., videos, news, blogs).

Importantly, although we use URLs included in tweets to track the spread of content, we consider the linked-to source itself and not the URL as the unit of analysis, because different URLs can link to the same source. This is particularly common in Twitter due to the 140-character length restriction for tweets and the widespread use of URL shorteners (e.g., http://bit.ly) that take long URLs and convert them into short URLs for posting into tweets.

Full details of how we compiled the diffusion data from URLs contained in tweets are given in Appendix A. Here we provide a brief description of the many steps involved. To compile the diffusion data for this study we first parsed out URLs from the 114,711 tweets that we collected from the 2,461 users over the 44 days. Of these tweets, 21,430 (18.7%)

[8] Note that additional data, such as user demographics, were not available and could not be manually collected. Twitter user profiles, unlike the user profiles in other online networks such as Facebook, are typically very sparse and do not require inputting even basic demographic information.

[9] Twitter now has a formal retweeting system and tracking has improved. However, this did not exist for Twitter during the time of our data collection.

contained URLs. These URLs were mostly short URLs, which we converted back to original long URLs using a tool called LongURL (http://www.longurl.org). For each original long URL we were given data from BackTweets (http://www.backtweets.com) counting URLs posted in Twitter (both in long URL and short URL form). Using the date the URL-containing tweet was posted, we obtained data on the number of times that piece of linked-to content (i.e., whatever was to be found at the original, long URL) was referenced in tweets immediately before it was posted, and then 7 and 14 days after it was posted.

In the analysis below we use the 14-day diffusion data, where $diffusion_{ij}$ is a count of the number of times content *j* was referenced in Twitter after it was first posted by transmitter *i*. The mean 14-day diffusion was 117.64 (SD = 1264.64) and was a heavily skewed distribution ranging from 0 to 30,204. Based on an examination of the data it appears that the diffusion processes quickly ran their courses. Hence, a 14-day horizon is reasonable in this case.

*Transmitter characteristics: independent variables.* Connectivity was measured by the number of followers a user has (out-degree). Activity was measured by the average number of tweets made per day by a user (irrespective of whether they contained URLS). Both of these variables appeared to be approximately power-law distributed, consistent with assumptions made in study 2. Descriptive statistics are reported in Table 5.

[INSERT TABLE 5 ABOUT HERE]

*Information content characteristics: control variables*. Certain characteristics of the content could also drive diffusion. Compared to content that is perceived to be low quality and/or appealing to only a niche audience, content that is perceived to be high quality and/or appealing to a broader range of people should spread further. To control for the effects that these content characteristics might have on diffusion we had judges from a large representative online panel rate each piece of content (i.e., whatever was on the webpage a

URL linked to, such as a video; they did not rate the website itself, such as YouTube) on an overall quality scale (1 = "very bad quality" to 5 = "very good quality"), and a breadth of appeal scale (1 = "content would appeal to almost no one" to 5 = "content would appeal to almost everyone"). The content linked to by each URL was rated by three independent judges. Each judge saw only one piece of content. Given that judges' perceptions of content quality and appeal can be highly subjective, our aim here was not to find agreement among each set of three judges. Rather, for each URL we took mean ratings on quality and appeal for use in the analysis.

*Endogeneity.* A concern when modeling any social interactions data is endogeneity. Our main concern in this context is that diffusion outcomes could influence a user's connectivity and activity. For example, if a person embeds a URL to a YouTube video in one of her tweets and over the subsequent days she observes others also posting this video then this feedback could make her speed up her tweeting (i.e., increase her activity), or conversely, if it does not appear to be spreading she could be discouraged and slow down her tweeting (i.e., decrease her activity). Also, if the users following her see the content and dislike it they might decide to "unfollow" her (i.e., decrease her out-degree), or conversely if other users not following her can trace the content back to her and liked it then they might decide to follow her (i.e. increase her out-degree).

To mitigate the chances of these types of direct and indirect feedback effects on the independent variables, we temporally separated them from the dependent variable. Specifically, we used the first 28 days of data for measuring the two transmitter-related independent variables, and only used content that was posted during days 29 to 44 for diffusion. For transmitter connectivity we used the number of followers they had at the end of day 28, and for activity we took the average number of tweets posted by per day during days

1 to 28.[10] Over the 16 days (days 29 to 44) used for diffusion outcomes, 312 of the sampled users posted 13,810 links.

As a further safeguard against endogeneity-related estimation bias, we only used "fresh" content that had been introduced into Twitter for the first time; i.e., content that, at the time of it being posted, had never previously been referred to by any URLs in Twitter. This helps because (1) it reduces the possibility that outside factors unrelated to transmitter activity and connectivity could be driving results, (2) it means that a piece of content's history in Twitter could not play a role (e.g., as a "social proof" signal), and (3) it allows us to examine only how the activity and connectivity characteristics of the transmitter *who first introduced the content* (i.e., the seed within Twitter) affect that content's diffusion in the network.

The resultant dataset included complete data for 9,656 "fresh" URL-linked pieces of content (70% of all content introduced by the sampled users in days 29 to 44). All 312 users who posted at least one link during days 29 to 44 (and who had nonzero activity and connectivity) introduced this content. The number of pieces of content per user was between 1 and 399 (mean = 30.95, SD = 53.79, median = 15).

**Results**

We used a random effects Poisson regression model to regress diffusion on transmitter characteristics of activity and connectivity, controlling for content characteristics of quality and breadth of appeal.[11] For transmitter *i* who posts unique content item *j* during days 29 to 44, using maximum likelihood we estimated the parameters in the following model:

[10] The choice of 28 days was arbitrary. In robustness checks we split the data at days 14 and 21 and found no qualitative differences in the results.

[11] See Appendix B for an alternative model that, for large networks such as Twitter where users typically have quite large "reach" into the network (i.e., the number of other users they are connected to directly and indirectly through others is large), is equivalent to the Poisson model used here.

$$diffusion_{ij}^{Day14} \sim Poisson(\theta_{ij})$$
$$\ln(\theta_{ij}) = \beta_0 + \beta_1 k_i + \beta_2 a_i + \beta_3 quality_{ij} + \beta_4 appeal_{ij} + \upsilon_i + \varepsilon_{ij}$$
$$\upsilon_i \sim N(0, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where, for transmitter $i$ or URL $j$ posted by transmitter $i$:
$k_i =$ connectivity
$a_i =$ activity
$quality_{ij} =$ mean perceived content quality
$appeal_{ij} =$ mean perceived content breadth of appeal

The parameter estimates and fit statistics are reported in Table 6. The best-fitting model (column 5) shows that, controlling for quality and appeal (both of which have significant positive effects on diffusion), activity, but not connectivity, positively affects diffusion.

[INSERT TABLE 6 ABOUT HERE]

*Interactions.* We also tested whether transmitter and information characteristics interacted in meaningful ways. Specifically, we checked whether there were significant activity × quality and activity × appeal interactions. Both interactions were significant ($p <$ .001). The effect of transmitter activity on diffusion is strengthened by increasing content quality and decreasing breadth of appeal. Put differently, highly active "pumps" are most influential on link diffusion in Twitter when the links they post refer to high quality content of interest to niche or specialized audiences. In other words, these types of transmitters are most effective for spreading quality content that most people may not otherwise find particularly appealing. This suggests that content alone is not enough to get something to spread, but that transmitter characteristics—in particular how active a transmitter they are—also matter. This interaction suggests even content that, on its own, may be unlikely to spread widely (e.g., because of limited appeal), can spread further if transmitted by the right person (e.g., a high activity "pump").

*Robustness checks.* First, we estimated a zero-inflated Poisson model to allow for some of the zeros in $diffusion_{ij}$ to come from the Poisson distribution and others to be a mass at zero. The log-linear regression model for the Poisson rate was the same as above, and the same specification was made for the latent probability that an observed $diffusion_{ij} = 0$ value came from either a mass at zero or the Poisson distribution. We once again found that, controlling for quality and appeal, activity had a positive effect on diffusion but connectivity did not.

Second, we used 7-day diffusion instead of 14-day diffusion as our dependent variable. In the Poisson model with both activity and connectivity regressors the same patterns of effects reported in Table 5 were found, although in both models the activity effect only approached marginal significance ($p = .11$ in both models).

Third, we removed users from the sample who posted large numbers of URLs during days 29 to 44 in case they were driving the results. We used the median average daily tweeting rate in our sample (2.61) times the number of days (16) to determine the cut-off number of URLs posted: just under 42 pieces of content. We excluded users who posted more than 42 URLs. The positive effect of activity but not connectivity was again found.

Fourth, we included links that were both fresh (i.e., with no prior diffusion within Twitter before being posted by transmitter *i*) and not (i.e., with some prior diffusion in Twitter and therefore not introduced by transmitter *i* but by someone else who was not in our sample). The positive effect of activity on diffusion should hold regardless of whether a piece of content is fresh and introduced for the first time to Twitter or whether it has already been around because this effect is driven by the underlying psychological process demonstrated in study 1 whereby freshness and novelty are attributed to information coming from highly active transmitters. Thus, activity should matter irrespective of prior diffusion. Using all links, fresh and pre-existing, we once again found a significant positive activity effect and

non-significant connectivity effect. The activity effect was slightly weaker, which might have been caused by a small number of URLs that had extremely large prior diffusion counts (in excess of 5,000). The activity effect was stronger once these well-established pieces of content were excluded from the dataset. This makes sense since information that already spread a lot would be closer to its maximum penetration, thus resulting in transmitter-related factors being less likely to make a difference.

## *DISCUSSION AND CONCLUSION*

Across three studies—a behavioral experiment, an agent-based model, and an empirical analysis of data we have consistently shown that transmitter activity plays an important role in driving information diffusion over online social networks. Moreover, it appears that the role of activity is at least as strong, if not stronger, than the role of connectivity. In fact, once a transmitter's activity is considered, their connectivity may play less of a role than previously thought in the literature. These findings are consistent with the theory advanced above. Further, they suggest that sharing information and digital content over social ties in online social networks—which requires explicit retransmission decisions after being exposed to and having consumed received information—may not be the same as other kinds of diffusion and social transmission previously studied in marketing. Although past studies have found that connectivity can play an important role in driving diffusion (e.g., Goldenberg et al. 2009), it is evident that this is not always the case and that other transmitter behaviors and characteristics can be at least as important as connectivity, if not more.

An explanation for the dramatic differences between the results supporting the role of connectivity (e.g., Goldenberg et al. 2009) and the current findings in favor of activity is as follows. In a social media environment, information that is seeded in the network at hand diffuses in parallel with other networks. The decision to introduce this information into the

network is therefore critical. Those individuals who are consistently active and have the shortest delay are perceived as more reliable in terms of "information freshness," and by avoiding lagging behind in transmitting information they increase the chances that the information they pump into the system is retransmitted because it is genuinely fresher. In other kinds of online social networks (e.g., the South Korean network "Cyworld" studied by Goldenberg et al. 2009) involving adoptions of virtual goods that do not require retransmissions, activity may not be observable like in other social media, leaving space for hubs to make a difference. Future research should address other kinds of online networks where activity and connectivity are both observable and compare their roles in driving key macro-level outcomes for different types of information or goods and different types of social transmission mechanisms.

In the case of social media and online networks such as Facebook and Twitter, large-scale diffusion of information such as YouTube videos and online news articles relies on people retransmitting this information. This aspect, which is less important in many of the diffusion contexts studied in past literature than it is in social media, suggests that micro, individual-level drivers of diffusion may differ to those previously studied. Theoretically, our findings provide some insight into the characteristics of transmitters that people pay attention to and are affected by when deciding whether or not to retransmit information. We of course do not claim that our findings generalize beyond the current context of information sharing in online networks.

Practically, the robust finding in favor of transmitter activity as a driver of online information diffusion has obvious implications for situations where managers or policymakers want to spread information over online social networks. For example, marketers can relatively easily measure network members' activity rates (e.g., average tweets per day or status updates per week) and use this information to select seeds for kick-starting

campaigns. We do not advocate moving away from connectivity as a criterion for seeding (e.g., hubs are still useful at exposing many people to something in one transmission, which might be good for raising awareness but not for triggering subsequent actions such as retransmission). Instead, we suggest that activity should also be considered in seed selection. Exceptionally active transmitters in online social networks—pumps—may be good seeds for viral campaigns where passing along information is critical to campaign success. Also, it is important to note that while identifying hubs requires network mapping (which can be difficult to do), activity can be measured by looking at nodes alone and observing what they do without information on network structure.

While we have tried our best to provide a robust series of tests of our propositions, using a wide variety of methodologies and analyses, the current research is of course not without limitations. Our aim was to test the hypothesis that activity plays a role in driving diffusion, which we did across three studies. We did not, however, fully explore boundary conditions and potential moderators of this effect. More work is obviously warranted. We encourage work on this and related questions about the roles of activity and connectivity in driving diffusion in social networks, for different types of information and different social transmission processes.

## *REFERENCES*


Barabási, Albert-László and Réka Albert (1999), "Emergence of Scaling in Random Networks," *Science*, 286 (5439), 509-512.

Bass, Frank M., (1969),"A New Product Growth for Model Consumer Durables," *Management Science*, 15(5) 215-227.

Baron, Reuben M. and David A. Kenny (1986), "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations," *Journal of Personality and Social Psychology*, 51(December), 1173-1182.

Berger, Jonah, and Katherine L. Milkman (2010), "Social Transmission and Viral Culture," working Paper, University of Pennsylvania.

Berger, Jonah, and Chip Heath (2005), "Idea Habitats: How the Prevalence of Environmental Cues Influences the Success of Ideas," *Cognitive Science*, 29, 195-221.

Chevalier, Judith, and Dina Mayzlin (2006), "The Effect of Word-Of-Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345-354.

Coleman, James S., Elihu Katz, and Herbert Menzel (1957), "The Diffusion of an Innovation among Physicisians," *Sociometry*, 20 (4), 253-270.

Dellarocas, Chrysanthos, Xiaquan Zhang, and Naveen F. Awad, (2007), "Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21, 23-45.

Dodds, Peter Sheridan and Duncan J. Watts (2004), "A Generalized Model of Social and Biological Contagion," *Journal of Theoretical Biology*, 232 (2005), 587-604.

Eliashberg, Jehoshua, Jedid-Jah Jonker, Mohanbir S. Sawhney, and Berend Wierenga, (2000), "MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures," *Marketing Science*, 19 (3), 226-243.

Engel, James F., Roger D. Blackwell, and Paul W. Miniard, (1993), *Understanding the Consumer*, Seventh Ed., Fort Worth, TX: Dryden.

Facebook, (2010), "Statistics," *Facebook Press Room*, http://www.facebook.com.

Garber, Tal, Jacob Goldenberg, Barak Libai, and Eitan Muller (2004), "From Density to Destiny: Using Spatial Dimension of Sales Data for Early Prediction of New Product Success," *Marketing Science*, 23 (3), 419-428.

Goldenberg, Jacob, Barak Libai, and Eitan Muller (2001), "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters*, 12 (3), 211-223.

Goldenberg, Jacob, Sangman Han, Donald R. Lehmann, and Jae Weon Hong (2009), "The Role of Hubs in the Adoption Process," *Journal of Marketing*, 73 (2), 1-13.

Heath, Chip, Chris Bell, Emily Sternberg (2001), "Emotional Selection in Memes: the Case of Urban Legends," *Journal of Personality and Social Psychology*, 81 (6), 1028-1041.

Hennig-Thurau, Thorston, Kevin P. Gwinner, Gianfranco Walsh, Dwayne D. Gremler (2004), "Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?" *Journal of Interactive Marketing*, 18 (1), 38-52.

Hirschman, Elizabeth C. (1980), "Innovativeness, Novelty Seeking, and Consumer Creativity," *Journal of Consumer Research*, 7 (3), 283-295.

Hopkins, Heather (2009), "Perez Hilton #1 Traffic Source Facebook," *Hitwise Intelligence*, February 25.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng (2007), "Why we Twitter: Understanding Microblogging Usage and Communities," *International Conference on Knowledge Discovery and Data Mining*, ACM: New York.

Katona, Zsolt, Peter Pal Zubcsek and Miklos Sarvary (2009), "Network Effects and Personal Influences: The Diffusion of an Online Social Network," Working Paper, University of California, Berkeley.

Libai, Barak, Ruth Bolton, Marnix Bügel, Ko DeRuyter, Oliver Götz, Hans Risselada, and Andrew T. Stephen (2010), "Customer to Customer Interactions: New Opportunities and Research Directions," *Journal of Service Research*, forthcoming.

Liu, Jianguo, Yanzhong Dang, Zhongtuo Wang, and Tao Zhou (2006), "Relationship Between the In-Degree and Out-Degree of WWW," *Physica A: Statistical and Theoretical Physics*, 371 (2), 861-869.

Lusch, Robert F. and Nicholas Tay (2004), "Agent-Based Modeling: Gaining Insight Into Firm and Industry Performance." In Christine Moorman and Donald R. Lehmann (eds), *Assessing Marketing Strategy Performance*, Cambridge, MA: Marketing Science Institute.

Rand, William and Roland Rust (2009), "Rigorous Agent-Based Modeling in Marketing," *INFORMS Marketing Science Conference*, Ann Arbor, Michigan.

Reichheld, Frederick F., and Thomas Teal. (1996), *The Loyalty Effect*, Boston, MA, Harvard Business School Press,.

Rogers, Everett M. (2003), *Diffusion of Innovations*, New York, NY: Free Press.

Rogers, Everett M., and Floyd Shoemaker (1971), *Communication of Innovations*. Free Press, New York.

Singh, Vik (2009), "Some Stats About Twitter's Content," *Yahoo / Zooie's Blog*, http://zooie.wordpress.com/2009/10/12/some-stats-about-twitters-content.

Stephen, Andrew T. and Donald R. Lehmann (2010), "Drivers of Word-of-Mouth Transmission at the Individual Level and Consequences for Diffusion," working paper, INSEAD.

Stephen, Andrew T. and Olivier Toubia (2009), "Explaining the Power-law Degree Distribution in a Social Commerce Network," *Social Networks*, 31 (4), 262-270.

Stephen, Andrew T. and Olivier Toubia (2010), "Deriving Value from Social Commerce Networks," *Journal of Marketing Research*, 47 (2), 215-228

Sundaram, Dinesh S., Kaushik Mitra, and Cynthia Webster (1998), "Word-of-Mouth Communications: A Motivational Analysis," In E. J. Arnould & L. M. Scott (Eds.), *Advances in Consumer Research*, 527-531.

Van den Bulte, Christophe and Stefan Wuyts (2007), *Social Networks and Marketing*. Cambridge, MA, Marketing Science Institute.

Watts, Duncan J. (2002), "A Simple Model of Global Cascades on Random Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 99 (9), 5766-5771.

Watts, Duncan J. and Peter Sheridan Dodds (2007), "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research*, 34 (4), 441-458.

Wu, Fang and Bernardo A. Huberman (2007), "Novelty and Collective Attention," *Proceedings of the National Academy of Sciences of the United States of America*, 104 (45), 17599-17601.

*FIGURE 1*

*MEAN CONDITIONAL RETRANSMISSION PROBABILITIES IN STUDY 1*

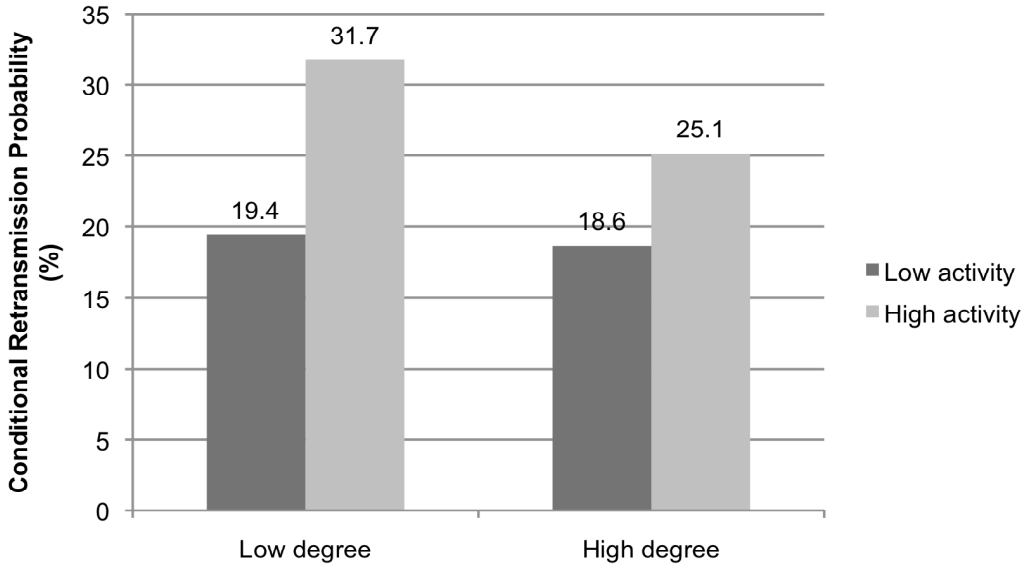

***FIGURE 2***

***DIFFUSION AS A FUNCTION OF ACTIVITY AND CONNECTIVITY IN STUDY 2***

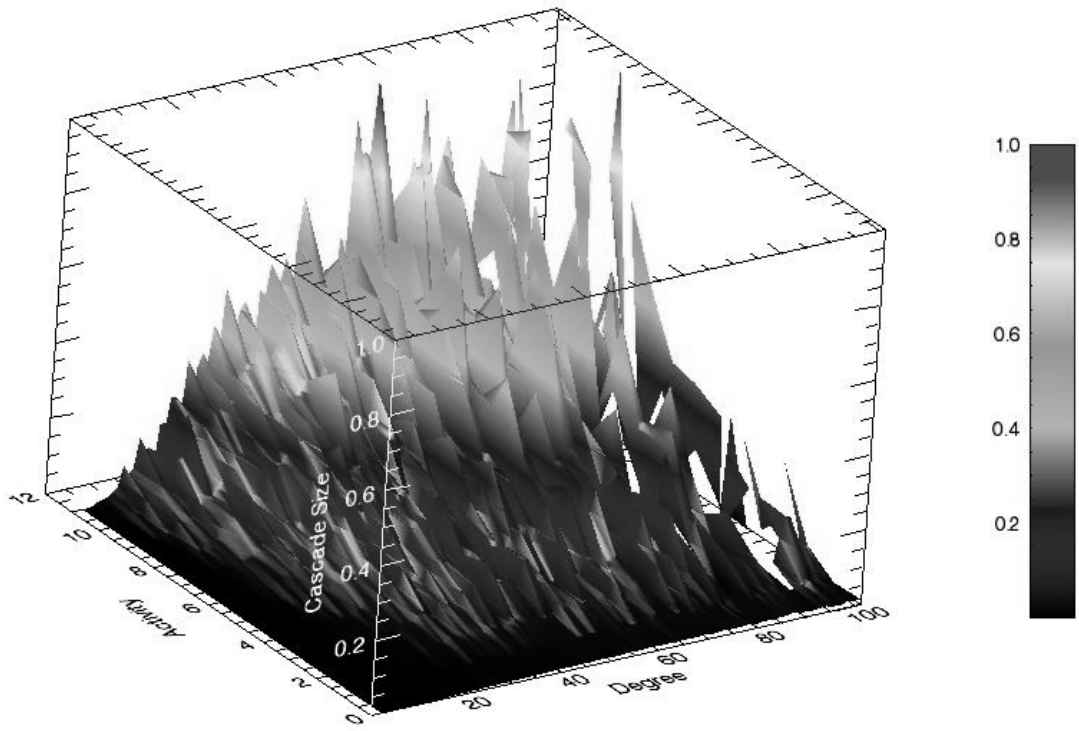

***FIGURE 3***

***COMPARISON OF EFFECT SIZE OF ACTIVITY VERSUS CONNECTIVITY IN STUDY 2***

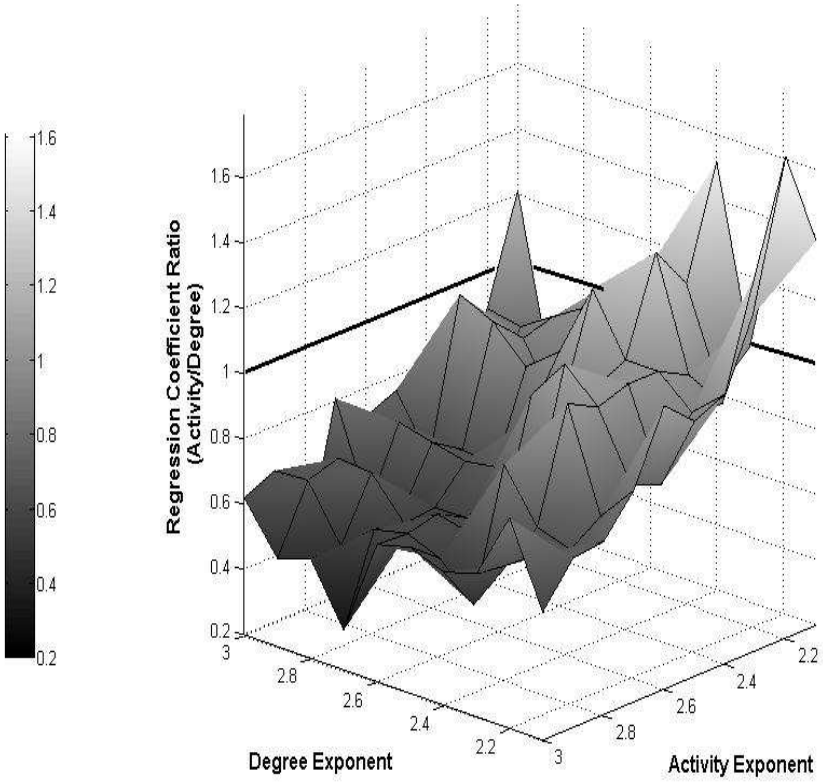

*TABLE 1*

*SCALE ITEMS USED IN STUDY 1*

**QUALITY: Transmitter's information quality** ($\alpha$ = .94)

1. This person's content suggestions are probably good.
2. The information this person posts is likely to be of high quality.
3. The information this person posts is likely to be interesting.
4. The information this person posts is likely to be useful to myself and others.
5. The information this person posts is likely to be informative.

**SPEED: Transmitter's speed** ($\alpha$ = .88)

6. This person is quick to provide information to others.
7. This person gets information sooner than others.
8. This person is faster in finding out new information.

**TRUST: Transmitter's trustworthiness** ($\alpha$ = .81)

9. This person is trustworthy.
10. This person is reliable.

*TABLE 2*

*MEDIATION ANALYSIS IN STUDY 1*

| **Parameter** | **Estimate (std. error)** | | |
|---|---|---|---|
| | **"Speed" mediator** | **"Trust" mediator** | **"Quality" mediator** |
| **Model 1: Activity → *mediator*** | | | |
| Intercept | 2.93 (.26)** | 3.72 (.24)** | 3.79 (.29)** |
| Activity[a] | 1.27 (.33)** | .78 (.31)** | .38 (.37) |
| Degree | -.54 (.39) | .01 (.36) | .19 (.43) |
| Activity x Degree | 1.00 (.49)** | -.69 (.46) | -.31 (.55) |
| $R^2$ | .34 | .09 | .01 |
| **Model 2: *mediator* → Pr(retransmit\|view)** | | | |
| Intercept | -4.62 (5.88) | -16.83 (7.02)** | -11.84 (6.22)* |
| Pr(view) | .24 ( .08)** | .21 ( .07)** | .17 ( .08)** |
| *mediator*[a] | 4.48 (1.47)** | 7.64 (1.77)** | 6.93 (1.65)** |
| $R^2$ | .22 | .28 | .28 |
| **Model 3: Activity, *mediator* → Pr(retransmit\|view)** | | | |
| Intercept | -6.13 (7.01) | -17.73 (7.84)** | -14.29 (7.00)** |
| Pr(view) | .25 ( .08)** | .21 ( .08)** | .16 ( .08)** |
| Activity[b] | 6.78 (6.22) | 6.48 (5.72) | 9.32 (5.50) |
| Degree | .72 (6.75) | -2.09 (6.48) | -3.89 (6.44) |
| Activity x Degree | -8.86 (8.64) | .98 (8.41) | -.92 (8.22) |
| *mediator*[a] | 4.23 (1.81)** | 7.08 (1.87)** | 6.74 (1.64)** |
| $R^2$ | .24 | .24 | .32 |
| **Model 4: Activity → Pr(retransmit\|view)** | | | |
| Intercept | | 2.54 (6.08) | |
| Pr(view) | | .32 ( .08)** | |
| Activity[a] | | 12.34 (5.85)* | |
| Degree | | -.76 (6.87) | |
| Activity x Degree | | -5.85 (8.73) | |
| $R^2$ | | .20 | |

* $p < .05$, ** $p < .01$.

Activity and degree are dummy-coded: low = 0, high = 1.

[a] Effect must be significant for complete mediation.

[b] Effect must be non-significant for complete mediation.

Sobel mediation test for activity → speed → Pr(retransmit|view) process: $Z = 2.39$, $p = .02$.

Sobel mediation test for activity → trust → Pr(retransmit|view) process: $Z = 2.17$, $p = .03$.

Sobel mediation test for activity → quality → Pr(retransmit|view) process: $Z = 1.00$, $p = .32$.

***TABLE 3***

***RANGES OF PARAMETER VALUES TESTED IN STUDY 2***

| **Parameter** | **Range** |
|---|---|
| $q$ (prob. of consuming \| exposure) | .1 to 1.0 |
| $r$ | .1 to 1.0 |
| $\gamma$ (degree distribution exponent parameter) | 2.0 to 3.0 |
| $\lambda$ (activity distribution exponent parameter) | 2.0 to 3.0 |
| $M$ (market potential) | 10,000 to 500,000 |
| $p$ (prob. of consuming due to external factors) | $10^{-5}$ to $10^{-3}$ |

*TABLE 4*

*EFFECTS OF SIMULATION PARAMETERS ON EXTENT OF DIFFUSION IN STUDY 2*

| Parameter | 1<br>Full Parameter Space | 2<br>Twitter-like Network |
|---|---|---|
| Intercept | .270 (.002)* | .127 (.008)* |
| Degree | .859 (.002)* | .700 (.005)* |
| Degree$^2$ | -.035 (.002)* | .005 (.005)* |
| Activity | .796 (.003)* | .843 (.008)* |
| Activity$^2$ | .093 (.002)* | .138 (.006)* |
| Degree × Activity | .284 (.002)* | .342 (.006)* |
| $q$ (prob. of consuming \| exposure) | .002 (.002)ns | -.012 (.005)* |
| Degree × $q$ | -.009 (.002)ns | -.007 (.005)* |
| Activity × $q$ | .049 (.002)* | .043 (.005)* |
| $\gamma$ (degree distribution exponent parameter) | .383 (.002)* | --- |
| Degree × $\gamma$ | .232 (.003)* | --- |
| Activity × $\gamma$ | .015 (.002)ns | --- |
| $\lambda$ (activity distribution exponent parameter) | .471 (.002)* | .533 (.007)* |
| Degree × $\lambda$ | .158 (.002)* | .205 (.006)* |
| Activity × $\lambda$ | .378 (.003)* | .459 (.010)* |
| $\gamma \times \lambda$ | .023 (.002)* | --- |
| Number of total nodes | 90000 | 9000 |
| Number of simulated | 900 | 90 |
| Adjusted $R^2$ | .66 | .84 |

* $p < .0001$, ns not significant. Standardized parameter estimates reported, with standard errors in parentheses.

*TABLE 5*

*DESCRIPTIVE STATISTICS FOR CONNECTIVITY AND ACTIVITY IN STUDY 3*

| Variable | Mean | St.Dev. | Median | Min. | Max. |
|---|---|---|---|---|---|
| Connectivity (number of followers) | 893.76 | 1903.51 | 279 | 2 | 15,957 |
| Activity (average tweets per day) | 8.77 | 14.86 | 2.61 | .04 | 79.28 |

*TABLE 6*

***ESTIMATES FOR RANDOM EFFECTS POISSON REGRESSION IN STUDY 3***

| **Model** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Parameter** | **Estimate (std. error)** | **Estimate (std. error)** | **Estimate (std. error)** | **Estimate (std. error)** | **Estimate (std. error)** |
| Intercept | 1.70 (.14)** | 1.67 (.15)** | 1.52 (.15)** | 1.52 (.16)** | -3.99 (.16)** |
| Connectivity (number of followers) | --- | <.01 (<.01) | --- | <.01 (<.01) | <.01 (<.01) |
| Activity (avg. daily tweeting rate) | --- | --- | .03 (.01)** | .04 (.01)* | .04 (.01)** |
| Content quality | --- | --- | --- | --- | 1.10 (<.01)** |
| Content broad appeal | --- | --- | --- | --- | .39 (<.01)** |
| Random effect variance | 6.13 (.52)** | 6.11 (.51)** | 5.97 (.50)** | 5.97 (.50)** | 5.85 (.49)** |
| -2 log likelihood (x $10^3$) | 7,482.38 | 7,482.36 | 7,482.29 | 7,482.28 | 7,252.85 |
| $N_{users}$ | 312 | 312 | 312 | 312 | 312 |
| $N_{content}$ | 9,656 | 9,656 | 9,656 | 9,656 | 9,656 |

* $p < .05$, ** $p < .01$.

*APPENDIX A*

*DATA COLLECTION AND PROCESSING STEPS FOR STUDY 3*

***Step 1: Prepare list of Twitter users and generate random sample***

- Using Twitterholic.com list of Twitter users and the Twitter API.
- From Twitterholic, get a list of the users with the most friends.
- For each user on this list, get their Twitter user ID.
- For each user on this list, using the Twitter API and the user's ID, get a list of their friends and their friends' Twitter user IDs.
- Continue this web crawling process using the Twitter API. Compile a list of user IDs and remove any duplicates from the list at each iteration.
- This generated a list of approximately 3 million Twitter user IDs in early May 2009.
- Random sample (without replacement) of users drawn from this list of user IDs. The sampled users were checked to ensure that they were "normal" people and not organizations (e.g., media organizations) or celebrities (e.g., Oprah Winfrey).

***Step 2: Collect data from Twitter***

- Using the Twitter API.
- For each user in the sample, gather publicly available data on their activity every 24 hours: total number of posts, current number of friends, current number of followers, time since joined Twitter, list of friends' user IDs (but not followers' user IDs because that is not publicly available), text of posts (and the date of each post).
- Gather the same data on each of a user's friends.

***Step 3: Analyze text of posts***

- Each post is a string a text (up to 140 characters, including spaces)
- For each post, identify if a URL is present (substring starting with “http://”)

***Step 4: Resolve, check, and analyze URLs***

- Using LongURL.com API and BackTweets.com API.
- If a post contained a URL this URL needed to be resolved (if it was a “short” URL that redirected to a true “long” URL), checked (to ensure that the webpage existed), and then analyzed (to see how prevalent it was in Twitter, as a measure of diffusion).
- Each URL was passed through the LongURL API. This resolved short URLs into their true long URL form, or left already true long URLs intact. Short URLs are popular in Twitter (e.g., http://bit.ly/12bG5c). However, a true long URL could have multiple short URLs representing it. Thus, any short URL had to be resolved into its long URL to allow for proper analysis. Each URL was also checked for existence.
- Each resolved URL (all in true long URL form) was passed through the BackTweets API. This API provides diffusion data for any submitted URL. I.e., for a given URL, it will provide a count of the number of posts in Twitter within a given date range that contain that URL. We collected this diffusion count for each resolved URL at a number of time points: immediately before the post was made, and then 7 days and 14 days after the post was made.

***APPENDIX B***

***BINOMIAL PROCESS MODEL OF LINK-SHARING IN TWITTER***

This appendix describes an alternative specification for the random effects Poisson regression model used in study 3. This specification is more closely aligned with the "true" social transmission process described early in the paper (i.e., exposure to information → consumption of information → retransmission of information), however because of data limitations in study 3, this model cannot be estimated without making certain assumptions. We describe this alternative model and estimation details (assumptions and results) for the sake of completeness and to demonstrate that the same results as those reported in study 3 can be obtained under a different model specification.

The social transmission process outlined in the paper, tested in study 1, and used to construct the ABM in study 2 can be summarized as follows: (1) transmitter *i* posts a tweet containing a link to content item *j*, and this tweet is broadcast to all of the transmitter's $k_i$ followers; (2) the probability that any follower consumes the content (i.e., clicks on the link) is *q*; and (3) the conditional probability that follower *j* retransmits the content in a tweet of their own after having consumed it is *r*. Step 2 implies that the number of consumers of transmitter *i*'s content is $X_{ij} \sim \text{Binomial}(k_i, q)$, and the expected number of consumers will be $E(X_{ij}) = qk_i$. Step 3 implies that the number of retransmissions of this content is $Y_{ij} \sim \text{Binomial}(qk_i, r)$, and the expected number of times that content item *j* posted by transmitter *i* will be retransmitted is $E(Y_{ij}) = rqk_i$. The expected extent of diffusion of item *j* throughout the network over a fixed period of time will be the sum of expected retransmissions taken over multiple generations of retransmitters. Further, based on our results from study 1, *r* and *q* may depend on transmitter characteristics. For example,

$q_i = \exp(\alpha_0 + \alpha_1 k_i + \alpha_2 a_i)/[1+\exp(\alpha_0 + \alpha_1 k_i + \alpha_2 a_i)]$ and

$r_i = \exp(\beta_0 + \beta_1 k_i + \beta_2 a_i)/[1+\exp(\beta_0 + \beta_1 k_i + \beta_2 a_i)]$, where $a_i$ is transmitter $i$'s activity rate.

Since we do not have Twitter data on each retransmitter of each piece of content (only who started it and the diffusion in Twitter after 14 days), this constraint of our dataset makes it impossible to estimate the parameters in this model, and further, not knowing the clicks on the links (i.e., consumption) makes it impossible to separately identify *r* and *q*. A simplification can be made whereby the two steps are merged (i.e., *r* and *q* are combined) and assumptions about the total reach or audience size for a transmitter are made. The total reach is the sum of their followers, their followers' followers, their followers' followers' followers, and so on (in graph theory this is the out-domain). For Twitter, it is safe to assume that the total reach for most users, *Z*, is large.[12] An approximation can be made whereby $Y_{ij} \sim$ Binomial($Z$, $(rq)_i$). With these assumptions,

$(rq)_i = \exp(\gamma_0 + \gamma_1 k_i + \gamma_2 a_i)/[1+\exp(\gamma_0 + \gamma_1 k_i + \gamma_2 a_i)]$ and the parameters can be estimated using maximum likelihood with data for $k_i$, $a_i$, and $diffusion_{ij}$ (divided by $Z$). Although the estimates will be approximate, they nevertheless offer general, preliminary empirical insight into whether activity, connectivity, or both affect the probability of consuming *and* retransmitting content in Twitter.

We estimated this preliminary, approximate model as a random effects Binomial regression (with the random effect added to control for repeated content postings by the same transmitters) assuming that the total reach was $Z$ = 1 million users and constant over transmitters.[13] As expected and consistent with the previous studies' findings, transmitter

[12] For example, one of the authors has a mere 64 Twitter followers but, according to a website that measures second-degree followers (i.e., followers-of-followers; see http://www.twinfluence.com), he has 285,535 second-degree followers from just 64 first-degree followers. Hence, we can expect most Twitter users—even those with small out-degrees—to have a large reach, *Z*.

[13] As of July 2009 the average number of followers for a Twitter user was 126 (Guardian 2009). If we conservatively assume reach of only three degrees of separation from an average transmitter their audience is 126 (direct 1st degree followers) + $126^2$ (indirect 2nd degree followers' followers) + $126^3$ (indirect 3rd degree followers' followers' followers) = 2,016,378 (assuming each follower at each degree of separation has a unique,

activity had a significant effect on the joint probability ($\hat{\gamma}_2$ = .035, $p$ < .01). Connectivity, however, did not ($\hat{\gamma}_2$ = -.00001).[14] These results were robust across a wide range of assumed values for $Z$.[15] These results also held when the quality and appeal content-related control variables were added to the linear regression model. Although approximate, these results suggest that, averaging over multiple generations of exposure, consumption, and retransmission, transmitter activity but not connectivity is an important factor affecting the likelihood that information flows through the network.

Note that a property of a binomial random variable is that as the number of trials (i.e., $Z$) gets large, the variable's distribution converges to Poisson. In this case, the total number of successes is $diffusion_{ij}$, the number of trials is assumed to be very large, and we can model the Poisson rate as a log-linked linear function of activity and connectivity. In other words, assuming the underlying behavior process described in the paper and a large potential reach for pieces of content, a Poisson model is appropriate. Of course, a Poisson regression is also a standard model for modeling count data such as our $diffusion_{ij}$ dependent variable. For these reasons we report the Poisson regression model in the paper.

non-overlapping set of followers). If we assume overlapping follower sets (due to clustering; see Watts and Strogatz 1998), then this reach is reduced. E.g., for 25% overlap the total audience is approximately 1.5 million, for 50% overlap it is approximately 1 million, and for a very high overlap of 75% it is around 500,000.

[14] Note that connectivity ($k_i$) and activity ($a_i$) are positively correlated but not strongly ($r$ = .34, $p$ < .001).

[15] We tried total reach = 25,000; 50,000; 75,000; 100,000; 250,000; 500,000 and the results were unchanged.

Europe Campus
Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00
Fax: +33 (0)1 60 74 55 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

www.insead.edu

Printed by INSEAD

INSEAD

The Business School
for the World®