ABSTRACT

Title of Document: ONLINE SOCIAL INFLUENCE

Yuchi Zhang
Ph.D. Candidate in Marketing, 2014

Directed By: Dr. Wendy Moe
Associate Professor
Marketing Department

Dr. David Godes
Associate Professor
Marketing Department

This dissertation studies the behaviors of consumers in an online, social context. In the first essay, we jointly model the drivers of social media rebroadcasting behavior. Our goal in this research is to propose a framework and model of social media rebroadcasting behavior that integrates the various factors shown to influence rebroadcasting behavior. These include the role of message content and influence, factors that have been studied separately with very little integration. The results from our proposed model show that not only does rebroadcasting activity vary with the content of the original message but also that individuals are more likely to rebroadcast content that closely fits with their own interests.

In the second essay, we ask whether online opinions impact consumers' decision quality and assess whether this impact occurs immediately or requires one to undergo learning first. We

focus on a setting where consumers have multiple learning episodes based on their experiences with opinions from both uni- and bi-directional ties (i.e. weak and strong ties). We find that the dynamic effects are dependent on the strength of the tie. Additional strong ties (operationalized as bi-directional links) lead to immediate positive effects on decision quality. In contrast, additional weak ties (uni-directional, follower relationships) as a source of information lead initially to lower decision quality. However, highly-experienced consumers receive, ultimately, higher positive effects on decision quality from weak ties as compared with strong ties.

Finally, in the third essay, we propose a new framework and model for identifying dimension specific influentials. We explicitly model individuals' preferences by estimating their locations on a market map to disentangle purchase behavior due to homophily from that due to influence. Our results show that it is important to estimate dimension specific influence, based on a comparison of model fit with a baseline model that measures influence along one dimension. We also show that individuals have varying levels of influence across dimensions, and an influencer for one dimension is not always influential on all dimensions.

ONLINE SOCIAL INFLUENCE

By

Yuchi Zhang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Associate Professor Wendy Moe, Co-Chair
Associate Professor David Godes, Co-Chair
Professor P.K. Kannan
Associate Professor Michael Trusov
Professor Ginger Zhe Jin

© Copyright by
Yuchi Zhang
2014

Acknowledgements

I am extremely grateful to the many individuals who helped me complete this dissertation. I would especially like to thank the following people. First, I would like to thank my co-advisors, Dr. Wendy Moe and Dr. David Godes, whose guidance, support, encouragement, and patience have enabled me to complete this dissertation. It has been truly a great pleasure to work with such passionate and knowledgeable scholars. Second, I would like to thank Dr. P.K. Kannan, for his insightful advice and teaching. The valuable feedback he provided helped me shape and improve my dissertation. Third, I would like to thank Dr. Michael Trusov, for his knowledgeable feedback and guidance. He has set a great example as an excellent researcher and shared valuable knowledge for tackling research problems. Fourth, I would like to thank Dr. Ginger Zhe Jin, for her generous time and effort serving on my dissertation committee. The feedback she provided is extremely valuable for my dissertation. I am also in debt to Dr. David Schweidel, who provided valuable advice and insights that contributed greatly to my dissertation. I would also like to thank the faculty, students, and staff in the Marketing Department at the University of Maryland for all their helpful advice, feedback, and support. Finally, and most importantly, I would like to thank my fiancée Cherissa. Her support, encouragement, and patience were the foundation for me to pursue and complete this dissertation. I thank my parents, Yiqiang and Chune, for their never-ending support and faith in me. They provided me with the wisdom and guidance to pursue this degree. Also special thanks to my sister Jennifer for her support.

Table of Contents

List of Tables

List of Figures

Drivers of Social Media Rebroadcasting: Investigating the Role of Message Content and Influencers

Introduction

Marketers face an important challenge of generating rebroadcasting activity for their firm created messages in social media (Zaman et al 2013). These behaviors by consumers, such as "retweeting" on Twitter or "liking/sharing" on Facebook (i.e. directly copying a message and disseminating it to one's friends or followers), are key components for propagating firm-generated social media messages. For example, in 2013, 139 billion messages were posted to Twitter, a popular social media messaging platform. Of these messages, about 24% were retweeted by members of the community (Leetaru et al 2013). That means 33 billion messages were able to gain an audience larger than those who already follow the messages of the original broadcaster. However, there is uncertainty as to factors influence the users' decision to rebroadcast those messages.

While few studies explicitly focus on social media rebroadcasting, a few related streams exist in both the offline and online domains. First, there has been extensive work aimed at understanding WOM influence, social contagion, and opinion leaders (Bass 1969; Bell and Song 2007; Nair et al 2010; Shriver et al 2013; Van den Bulte and Joshi 2007). More recently, researchers have placed an emphasis on investigating social effects at the individual level and use social media data to measure a user's influence and susceptibility to that influence (Aral and Walker 2012; Iyengar et al 2011; Trusov et al 2010; Watts and Dodds 2007). Second, a few researchers investigates why individuals share content and increase their WOM transmission for specific types of message content (Heath, Bell and Sternberg 2001; Berger 2011; Berger and Milkman 2012; Berger and Schwartz 2011).

While previous research separately suggests that influencers and message content impact rebroadcasting, there has been very little integration in these streams of research. In this paper, we develop a new approach in modeling rebroadcasting behavior that integrates the effects of influencers with that of the message content. We expect that it is critical to model both the message content and impact of influencers in order to assess the relative impact of each on rebroadcasting behavior. More importantly, by integrating the two streams of research, we also introduce a new measure to capture the effects of the fit between a user's interest and the focal content of the message on rebroadcasting. This allows us to account for the possibility that content may differentially affect an individual's rebroadcasting behavior.

Specifically, we construct a split hazard model to account for the impact of content- and user-specific variables, of content-user fit, and of one's ability to influence on rebroadcasting behavior. We incorporate the impact of influence on rebroadcasting as a function of both a user's ability to influence others and an individual's susceptibility to the influence of other users.

One important challenge with modeling social media message content is the unstructured text of each message. To compare the effects of the message content with that of influencers, we text analyze (Lee and Bradlow 2011; Netzer et al 2012) social media messages to identify the topic(s) featured in each message and convert the text into quantifiable metrics. This allows us to both characterize message content and develop profiles for individuals based on their previously posted messages. The development of quantitative metrics for these user profiles is key to not only accounting for observed heterogeneity across individuals but also in capturing the fit between a user and the content of a given message.

Our empirical results highlight the need to account for variation in content, differences across individuals and effects of previous rebroadcasters when studying social media rebroadcasting behavior. Not surprisingly, we find that content matters, as some topics are rebroadcast more than others. We also find significant variation across individuals in their tendencies to rebroadcast. Individuals who historically broadcast certain content, who have many followers, or who generally tend to post more messages have a higher propensity to rebroadcast in general. In addition to the main effects of the message content and individual's interests on rebroadcasting, we also find that the fit between the user interests and the message content significantly impacts rebroadcasting behavior.

Our analysis also identifies a limited number of influentials whose own rebroadcasting behavior encourages the subsequent rebroadcasting behavior of others, underscoring the potential value of targeting influentials to increase the reach of a social media message. Taken together with the user-content interaction effects, this suggests that matching message content to the interests of influentials can encourage subsequent rebroadcasting of an organization's social media messages. However, appealing to these influentials is not without risk if the message content which they are more likely to rebroadcast differs starkly from that which the broader audience will rebroadcast, as we will illustrate.

To demonstrate the value of our model, we perform a series of simulations in which we vary the type of content of the original message and identify different influentials for seeding purposes (i.e., incentivizing them to immediately rebroadcast). In our context, our simulation results show that, under certain circumstances, tailoring message content to fit the preferences of influential users can be more effective than traditional seeding strategies that simply promote to influential users, but there are limits to the benefits of doing so.

The remainder of this paper proceeds as follows. In the next section, we discuss related literature and show our conceptual framework of social media rebroadcasting. Next, we describe our unique data that includes an important textual component and the method that we use to convert the text into quantifiable metrics. Then, in the Model Section, we detail our modeling framework and describe how we incorporate influencers and content into a unified model. Next, we describe the results of our empirical analysis and present a series of simulated scenarios in which we evaluate the effect of alternative message seeding strategies on rebroadcasting activity. We conclude with a discussion of the implications of our research.

What Drives Rebroadcasting Behavior?

A number of recent studies examine factors that affect rebroadcasting (e.g., online WOM) behaviors. Since rebroadcasting is a specific form of WOM, we briefly discuss the two streams of research that are relevant to our study: (1) the role of influencers (2) and the effects of the message content.

The first stream of research studies the role of influence on consumer behaviors. Prior to the emergence of social media, researchers documented the importance of WOM influence, social contagion, and opinion leaders on product and information diffusion processes (Bass 1969; Bell and Song 2007; Nair et al 2010; Shriver et al 2013; Van den Bulte and Joshi 2007). For example Bell and Song (2007) and Shriver et al (2013) find that consumers are influenced by their geographic neighbors' purchase of online grocery and adoption of solar panels. Likewise, Nair et al (2010) find evidence that opinion leaders affect doctor's prescription choices. More recently, researchers have placed an emphasis on investigating social effects at the individual level (Aral and Walker 2012; Iyengar et al 2011; Trusov et al 2010). Iyengar et al (2011), for example, model the effects of opinion leaders; however they also show that self-reported measures of information leadership often are not strongly correlated with how others in the community perceive you. 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 other users in his or her social network. Likewise, Aral and Walker (2012) conducted a randomized field experiment to identify characteristics of individuals (i.e. age, gender, and relationship status) whose purchase behaviors are observed to alter the purchase behaviors of others. The data from their experiment allowed them to build a demographic profile of influential users.

An important issue raised in modeling influence is the need to separate the effects influence from that of one's susceptibility. Watts and Dodd (2007) suggest that under a number of circumstances, influentials have little impact on overall diffusion. Instead, they contend that it is the easily susceptible masses that drive large cascades. This finding further complicates the measurement of influence as any method employed would need to separate the effects of influence from the effects associated with the susceptibility to influence to assess the levels of one's influence accurately (e.g. Trusov et al 2010, Aral and Walker 2012). Otherwise, we may mistakenly deem one to be influential when, in reality, the subsequent rebroadcasting behavior observed is primarily driven by the susceptibility of others.

Our study differs from the previous research in a number of aspects. While Trusov et al (2010) identify dyadic influence and Aral and Walker (2012) identify demographic characteristics of influentials, we uncover one's ability to influence the rebroadcast timing of all other users who follow a firm's social media messages. Our research also differs from previous studies by explicitly modeling the message content and assessing how the fit between individual interests and the message content impact rebroadcasting behaviors. By controlling for the message content and the fit between individuals and content, we are able to disentangle rebroadcasting due to influence from that attributed to a baseline preference to share certain types of message content.

The second stream of research investigates why individuals share content and transmit more WOM for specific types of content. For example, Heath, Bell and Sternberg (2001) find that emotional content is more likely to be shared in social media environments. Researchers also show how content, that incites emotions and arousal or is cued by the environment, is also more likely to be shared (Berger 2011; Berger and Schwartz 2011; Berger and Milkman 2012). Furthermore, Toubia and Stephen (2013) find evidence that individuals share content to gain intrinsic- and image-related utility.

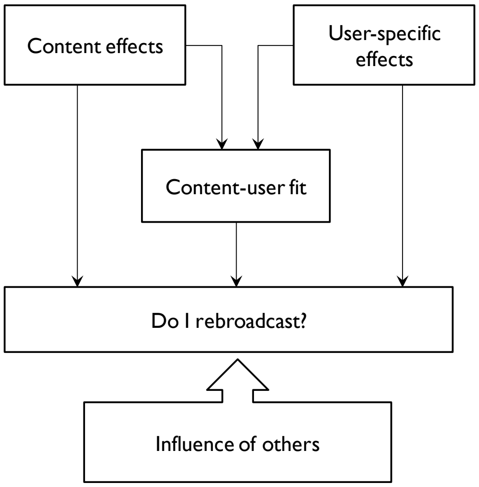
While the previous studies have looked at how different types of content are more or less likely to be shared, limited research has examined how the message content interacts with the user's interests (reflected by the content of her historical messages). If social media users rebroadcast messages to create their personal brands (Toubia and Stephen 2013), it seems reasonable that they may be more likely to rebroadcast messages with content that is consistent with their image (e.g., Kirmani 2009). Thus, in this research, we examine both the effects of content on users' rebroadcasting activities as well as the effect of how the content fits with a user's interests.

Modeling Framework

Our approach to modeling rebroadcasting behavior integrates the two separate research streams discussed above. Figure 1 shows our conceptual framework, outlining the factors that drive rebroadcasting behavior. The main focus of our research is to jointly model the effects of influencers and the effects of content on social media rebroadcasting and assess how these two factors are related. Unique to our research, we also examine how the content effects interact with the user's profile. We hypothesize that it is critical to model the fit between the message content and individuals. Consider a scenario where we observe an individual rebroadcasting a social media message and, subsequent to this, the message is rebroadcast by many individuals. One potential explanation for this is that the original rebroadcaster influenced others' behaviors by accelerating the pace at which the social media message was rebroadcast. However, an alternative explanation would posit that the original social media message features content that is likely to be shared or that subsequent individuals who rebroadcast have a high tendency to do so for messages that feature this particular content.

The previous example shows that neglecting the variation in rebroadcasting activity associated with message content and individual differences – such as the fit between content and individual – may lead one to erroneously conclude that some individuals are key influencers or that certain types of message content are more prone to spread by word of mouth for all individuals. To the best of our knowledge, no previous research has jointly modeled how the message content and influencers affect WOM transmission or examined how these two factors may be related in terms of their fit.

Figure 1. Factors that drive rebroadcasting behavior



Data

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 which 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.

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" 2011). 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. Table 1 provides some descriptive statistics of the retweet behavior and shows significant variation across schools.

To understand the drivers of rebroadcasting behavior, we must compare it to instances in which a user chooses not to rebroadcast. However, only individuals who have rebroadcast at least one or more of the schools' messages are included in the feed of Twitter data. As a result, we do not have any individuals who have never rebroadcast any message in our data. We therefore supplement the data from PeopleBrowsr with observations from a random sample of followers (50 from each school) who did not rebroadcast any of the schools' messages during our observation period. The behavior of these individuals is collected and added to our dataset, thus allowing us to incorporate factors associated with the decision of whether or not to rebroadcast. As a result, our final data set includes the 1,760 unique rebroadcasters found in the

PeopleBrowsr dataset and the 500 additional users randomly sampled from the follower population.

Table 1. Top 10 Business Schools Summary Statistics

Rank	School	Number of Tweets	Number of Retweeters	# Tweets Retweeted	# Retweet Observations	# No-Retweet Observations
1	University of Chicago	273	131	142	238	49,175
2	Harvard University	202	218	144	319	53,817
3	University of Pennsylvania	496	414	303	669	229,475
4	Northwestern University	225	191	153	389	53,836
5	Stanford University	226	278	157	476	73,652
6	Duke University	476	111	166	250	76,386
7	University of Michigan	255	35	49	80	21,595
8	UC Berkeley	527	163	158	283	111,968
9	Columbia University	284	133	142	240	51,732
10	MIT	55	86	40	130	7,350
	Total	3,019	1,760	1,454	3,074	728,986

Overall, our data include 732,060 observations where each observation is either a rebroadcast or a non-rebroadcast of an original business school message (see Table 1). For each of the 3,074 rebroadcast observations, the time elapsed, measured in minutes, since the original message broadcast is recorded. The median time to rebroadcast across all rebroadcast observations is 322 minutes, ranging from 1 minute 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. Since followers may rebroadcast a message at a time after our observation period has ended, our data are right censored. We will describe how we accommodate this characteristic of our data in our empirical analysis in the Model Development section.

Text Analysis and Variable Specification

For each observation, we construct a set of variables that describe the original message and a set of variables that describe the user. The message-specific variables are intended to describe the content of the posted message and are obtained using text mining procedures similar to those used by Lee and Bradlow (2011) and Netzer et al (2012). Our goal is to characterize the content of each posted message and identify a set of themes discussed among the various schools and the users in our data. To do this, we follow a four step process. First, we create a dataset consisting of all original messages for the 10 schools and the most recent 200 messages (excluding retweets) from each user in our sample, resulting in 268,609 messages[1]. Second, we create a word bank consisting of all nouns that appear more than once in these messages and use the Porter Stemmer (Porter 2006) to reduce these nouns to their root form to characterize the focal topic of the message (e.g. "schools" is stemmed to form "school" and all occurrences of "schools" or "school" are considered to be the same). This yields a final word bank of 759 unique words. Third, we create a 268,609 message x 769 word data matrix in which either a zero or one indicates whether each of the 3,019 school messages and 265,590 user messages includes each of the 769 words. Fourth, using this data matrix, we factor analyze the words, with varimax rotation, to identify the underlying themes reflected by the text (Netzer et al 2012).

[1] We note that not all users have 200 broadcast observations. For those with less than 200, we use all their available broadcast observations.

Table 2. Words Loading onto Each Factor

1) School	2) Finance	3) Politics
school	private	romney
mba	equity	mitt
prof	sector	santorum
business	capita	gingrich
kellogg	invest	iowa
dean	fund	votes
wharton	hedge	republican
booth	firm	paul
harvard	pension	tax
columbia	prof	obama
haas	partner	resident
stanford	debt	candid
ross	bank	carolina
sloan	crisis	perry
graduate	booth	election
applicant	market	race
mit	europe	campaign
berkeley	industry	poll
executive	wealth	politics
class	economy	attack

Our factor analysis resulted in three factors with eigenvalues greater than 1. In Table 2, we show the 20 words with the highest loadings for each factor (displayed before stemming for ease of interpretation). Factor one (F_1) is made up of terms specific to the schools. Thus, we refer to content scoring high on F_1 as school-related content. Factor two (F_2) includes words that indicate a focus on finance, and factor three (F_3) indicates of a focus on politics. Factor scores (with a mean of zero and standard deviation of one across all messages) are computed for each message, providing a concise and quantitative description of the message content.

We use the resulting factor scores, denoted F_{j1}, F_{j2} and F_{j3}, to characterize the content of message *j*. The content of most messages focused on school-specific topics, with 513 original messages scoring at least one standard deviation above the mean F_{j1}. Political topics were also popular with 216 messages scoring one standard deviation above the mean F_{j3} and finance-related topics were the least frequently posted with only 81 original school messages scoring at least one standard deviation above the mean F_{j2}. Tables 3, 4 and 5 presents a sample of representative Twitter messages[2] from each school for each text factor, along with the mean and standard deviation of the factor scores across all messages for each school.

We also consider the posting-frequency of the school posting the original message ($SFREQ_j$) as a message-specific variable. This variable is calculated as the total number of posts during the week prior to the posting of message *j* and accounts for differences in rebroadcasting activity that may exist across schools and across time due to the recent volume of social media messages sent. The average of this variable is 16.3 broadcasts per week, with substantial variance both across schools and over time (standard deviation is 5.7 across schools and 9.0 across weeks).

Table 3. School Related Messages

School	Mean F_1	Std Dev F_1	Illustrative Message
University of Chicago	2.51	1.58	Check out the transcript from the @ftbuseducation web chat with #ChicagoBooth prof Linda Ginzel http://t.co/bDZrS2ev #mba #business
Harvard University	1.09	1.16	Q&A with Dean Nitin Nohria on enhancements to the MBA curriculum and his role as dean [@nytimes] http://t.co/4rsglvg
University of Pennsylvania	1.81	1.34	From the Dean: 'An Extraordinary Opportunity' -- Dean Robertson's message to the incoming MBA Class of 2013: http://t.co/q4jOgi4
Northwestern University	2.72	1.68	B-school application tweets? http://ow.ly/6ooyx #Kellogg Prof Rakesh Vohra questions whether schools gain authenticity with brevity
Stanford University	1.55	1.29	FT: Bschools now trying to recruit a different type of MBA student. Quotes StanfordGSB dean @Saloner. http://t.co/nulUeoA
Duke University	1.12	1.13	London alumni: Join us for a reception &discussion with Fuqua's new dean Bill Boulding. Oct. 20. Learn more & register: http://ow.ly/6Qwrq
University of Michigan	1.43	1.70	The Ross School's Executive and Part-time MBA programs place among the @BWbschools Top 10 again. http://t.co/9FntzMWU
UC Berkeley	1.06	1.27	Five Things We Like About the Haas School of Business | MBA Admissions Blog by MBA Game Plan http://ht.ly/5QHDW
Columbia University	1.07	1.21	Check out Columbia Business School Gear the School's virtual store for B-school branded merchandise! http://t.co/L8jaSrwD
MIT	1.68	1.52	Interested in an MIT Sloan MBA? We have an MBA LinkedIn group that we'd love for you to join: http://t.co/Va92tZNP

[2] These messages are selected from the top 10 messages within each school that have the highest factor score for the respective factor.

Table 4. Finance Related Messages

School	Mean F_2	Std Dev F_2	Illustrative Message
University of Chicago	0.59	0.93	Prof. Raghuram Rajan said if sovereign-debt crisis leads to a banking crisis it would spread to the US very quickly http://ow.ly/6wIaD @WSJ
Harvard University	0.07	0.23	Bob Pozen discusses money market funds under regulatory attack [@FT] http://t.co/CkoXaLl
University of Pennsylvania	0.13	0.64	Do investors benefit from private meetings with management? New research from Wharton prof. Brian Bushee and others: http://t.co/W44r0aL
Northwestern University	0.26	0.34	Europe the euro and an unstable economic climate http://t.co/Cv1BliBE #Kellogg Prof Sergio Rebelo goes behind the scenes on the debt crisis
Stanford University	0.43	1.48	Prof Admati: Bankers have confused us equating equity/capital (how they're funded) w liquidity (how funds invested) http://t.co/8xaarpq
Duke University	0.09	0.98	Prof @camharvey assesses the chances of a 2nd banking crisis. Where would you put the odds? http://ow.ly/6dxCN
University of Michigan	0.03	1.13	Private Equity Post-Op http://t.co/vUOdtKWI
UC Berkeley	-0.05	0.16	Haas Socially Responsible Investment Fund Now Fully Invested http://ht.ly/5gWLn
Columbia University	0.09	0.65	Investors can mine news stories to capitalize on info that financial analysis overlooks. Paul Tetlock in Ideas at Work: http://ow.ly/52V9W
MIT	-0.03	0.37	MIT Sloan Professor Simon Johnson on how the middle class pays for financial market mistakes: http://t.co/mCSJFhz1

Table 5. Politics Related Messages

School	Mean F_3	Std Dev F_3	Illustrative Message
University of Chicago	0.40	0.89	Prof. Raghuram Rajan tells @Forbes_india both Democrats and Republicans are unrealistic in their fiscal prescriptions http://ow.ly/6fM9R
Harvard University	0.06	0.27	Bob Pozen discusses a list of problems with Herman Cain's 9-9-9 tax plan to reform the tax code [@HuffingtonPost] http://t.co/zIllLdG4
University of Pennsylvania	0.04	0.18	Wharton MBA student Keya Dannenbaum and @UofPenn PhD Paul Jungwirth found a start-up to "make politics more simple."
Northwestern University	0.21	0.24	Financial Trust Index shows a majority of Americans oppose raising the debt ceiling http://ow.ly/5RKXu #Kellogg Prof Paola Sapienza
Stanford University	0.12	0.27	Republicans warn tax hikes will kill jobs but this not as lethal as practices like downsizing says Prof Pfeffer http://t.co/OSmsAjf
Duke University	-0.02	0.18	President Obama is in Durham today for his visit with LED lighting manufacturer Cree Inc. http://ow.ly/5gvNm #ObamaCree
University of Michigan	0.03	0.18	@HoffmanAndy discusses Rick Perry politics culture & consensus on @MPRnews. http://ow.ly/6bsLd (via @erbinstitute) #sustainability
UC Berkeley	-0.01	0.14	What it will take for President Obama & big biz to bring back US jobs? Prof. Laura Tyson discusses in The Washington Post http://ht.ly/67I6M
Columbia University	0.04	0.26	Prof. Mayer and Dean Hubbard speak to @NPR about the potential reach of President Obama's #refinancing plan. http://t.co/cV9lrwH6
MIT	-0.03	0.17	Our own Simon Johnson weighs in on the recent debt ceiling vote outcome: http://ti.me/qq7O6Q ^MP

In addition to the message-specific variables above, we also consider a number of user-specific variables, including variables that describe a user's interests and network characteristics. To create variables that represent a user's interests, we again make use of the factors identified in the text analysis described previously. For each user, we compute the average factor scores corresponding to their 200 most recent original messages denoted $AVGF_{i1}$, $AVGF_{i2}$, $AVGF_{i3}$[3]. This only includes new messages created by the user and excludes all rebroadcasted messages. These variables provide a user profile that describes the user's tendency to post messages related to school, finance or politics. Many of the users included in our data can be described as having school-related interests (178 profiles have $AVGF_{i1}$ scores that are at least one standard deviation above the mean), with only 68 profiles that can be described as finance-related and 109 as politics-related.

We also include three variables that characterize an individual's network position: (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 ($RATIO_i$). An individual following many others, for example, may be more likely to seek information rather than disseminate it (Java et al. 2007) and hence may be less likely to rebroadcast. In contrast, one with many followers may have greater rebroadcasting tendencies. The $RATIO_i$ variable is similar to one used by Anger and Kittl (2011) and would capture the relative size of followers and following, thus controlling for absolute differences in the users' level of Twitter activity.

We also consider the frequency of the user's most recent 200 (or less) broadcasting observations ($UFREQ_i$) as a measure of their posting behavior within their social network. To

[3] This variable is averaged across all original messages. If a user has less than 200 messages, then we compute $AVGF_{i1}$, $AVGF_{i2}$, and $AVGF_{i3}$ from all their past messages.

calculate this variable, we record the time span that the 200 most recent broadcasting observations fall under. Within that time frame, we calculate $UFREQ_i$ as the average number of posts per week (i.e. 200 / number of weeks in time span). If the user posts less than 200 messages, then we perform the same calculation with their total number of broadcasted messages instead of 200. On average, each user posts 4.5 messages per week, with a standard deviation of 5.6 (less 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.

Finally, we create a content-user interaction variable that captures the fit between a message content with a user's general interests using the user profile factors ($AVGF_{i1}$, $AVGF_{i2}$ and $AVGF_{i3}$) and message content factors (F_{j1}, F_{j2} and F_{j3}). Consider a user who historically broadcasts social media messages on a certain topic and encounters a firm's broadcast on the same topic. As the message is consistent with the types of messages that the individual typically broadcasts, we would expect that this user would be more likely to rebroadcast this message compared to individuals who haven't revealed a similar interest in the topic. Conversely, we expect that a user would be less likely to rebroadcast a message that is very different than what that user typically broadcasts. Consistent with prior research that has operationalized similarity as being inversely related to the distance between objects (e.g., Tversky 1977; Hutchinson and Mungale 1997; Schweidel et al. 2006), we operationalize the fit (i.e., similarity) between individual *i* and message *j* as negatively related to the Euclidean distance between *i*'s content profile and the content of message *j* across the three factors denoted by *h*:

$$FIT_{ij} = -\sqrt{\sum_{h=1}^{3}\left(F_{j,h} - AVGF_{i,h}\right)^2} \tag{1}$$

Model Development

In this section, we discuss our model of social media rebroadcasting. Overall, rebroadcasting behavior is fairly uncommon. While there is variation in *when* users choose to rebroadcast a given message, there is also a large mass of users in our sample who do not rebroadcast. Therefore, we use a split hazard specification (Sinha and Chandrashekaran 1992) where, conditional on individual *i* rebroadcasting message *j*, we assume that the time at which the message is rebroadcast, t_{ij}, follows a hazard process represented by f(t_{ij}).

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 be written as:

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

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 represent 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 during period 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 but this event is censored in our data. Thus, S(T_{ij}) accounts for the censoring of those who have a probability, P_{ij}, of eventually rebroadcasting.

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_{j1}, F_{j2} andF_{j3}) and the posting frequency of the school ($SFREQ_j$). We allow heterogeneity in C_j across different messages, where σ_m^2 is the variance among the content effects for each message.

$$C_j \sim N\left(\mu_j, \sigma_m^2\right)$$
$$\mu_j = \beta_1 \times SFREQ_j + \sum_{h=1}^{3} \beta_{1+h} \times F_{j,h} \tag{4}$$

We similarly specify user effects U_i as being affected by the user's profile as measured by $AVGF_{i1}$, $AVGF_{i2}$ and $AVGF_{i3}$. 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 $RATIO_i$. 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 zero. Finally, we also accommodate the potential effects of a user's poster frequency (*UFREQ*) of rebroadcasting behavior. Again, we allow heterogeneity in U_i across different individuals, where σ_w^2 is the variance among the user effects for each individual.

$$U_i \sim N\left(\eta_i, \sigma_w^2\right)$$
$$\eta_i = \sum_{h=1}^{3} \beta_{4+h} \times AVGF_{i,h} + \beta_8 \times \log\left(FOLLOWERS_i + 1\right) +$$
$$\beta_9 \times \log\left(FOLLOWING_i + 1\right) + \beta_{10} \times \frac{\log\left(FOLLOWERS_i + 1\right) + 1}{\log\left(FOLLOWING_i + 1\right) + 1} + \beta_{11} \times UFREQ_i \tag{5}$$

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 equation (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). Thus, we include the role of influence in this component of the model much like Trusov et al. (2010) and Aral and Walker (2012). 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 et al (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:

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

where λ and c are parameters of the Weibull distribution to be estimated, and $g(INFL_{ijt})$ is a function that characterizes how previous rebroadcasters of message j affect users i at time t.

An important issue raised in modeling influence is the need to separate the effects influence from that of one's susceptibility. Watts and Dodd (2007) suggest that under a number of circumstances, influentials have little impact on overall diffusion. Instead, they contend that it is the easily susceptible masses that drive large cascades. We adopt an approach similar to that used by Trusov et al (2010) in measuring the role of influence. We assume that some users are more influential than others. At the same time, some users may be more susceptible to the influence than others. Thus, at time t=τ, we specify *g(INFL$_{ij\tau}$)* to include both differential influence among previous rebroadcasters as well as heterogeneity among individuals in their susceptibility to previous rebroadcasts as follows:

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

where γ_u is the influence of user *u*, the indicator function $1(t_{uj}<\tau)$ captures if *u* has rebroadcast message *j* by time τ, and α_i reflects the extent to which individual *i* is susceptible to this influence[4].

To explain the intuition behind Equation 7, we focus on calculating an influence effect on individual *i* at time τ. First, the influence effect 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. Second, the influence effect depends on the ability of other users to impact *i*. This is reflected in γ_u, where larger γ_u reflect a greater ability for user *u* to influence individual *i*. Finally, the influence effect is time varying, since at time τ, only users who have a rebroadcast time $t_{uj}<\tau$ will have the potential to influence *i*. This specification helps us identify both α_i and γ_u because each individual may be affected by the influence of many different users and each user has the ability to influence many different individuals.

When specifying γ_u, we expect that only a small number of users who are considered influential may exist (Van den Bulte and Joshi 2007; Trusov et al 2010). Therefore, we model heterogeneity in a user's ability to influence by assuming that the user population consists of a

[4] 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 Equation 7.

discrete mixture of influentials where $\gamma_u = 1$ with probability π_u and non-influentials where $\gamma_u = 0$, with probability $1-\pi_u$. For more details about the estimation procedure, we refer the reader to the Appendix.

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 $\alpha_i=0$. Again, details of the sampling procedure are provided in the Appendix.

To disentangle and identify the effects of influence and susceptibility, we specify γ_u to be a mixture of zeroes and ones while α_i is a mixture of zeroes and α[5]. 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: $\lambda \sim Gamma(a_{\lambda 0}, b_{\lambda 0}), c \sim Gamma(a_{c0}, b_{c0}), \beta \sim N(\beta_0, \sigma_0^2), \delta \sim N(\delta_0, \tau_0^2),\ \sigma_u^2 \sim Gamma(a_{u0}, b_{u0})$, $\sigma_m^2 \sim Gamma(a_{m0}, b_{m0})$. We use a block Metropolis-Hastings algorithm to separately draw each parameter. 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.

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

Model Findings

Before discussing the estimation results, we first evaluate our proposed model by comparing it to a number of alternative model specifications.

Model Fit Comparisons

We estimate our proposed model on the Twitter data from the 10 business schools in our sample. The primary goal is to assess the value of each model component (e.g., message content, influencers, and content-user fit) by estimating a number of nested models (see Table 6) and compare fit. We use a number of different evaluations to compare each model. These include the deviance information criterion (DIC), log marginal density, likelihood of a holdout sample, and hit rate of a holdout sample. For the holdout sample, we randomly select 25% of our observations (individual-message specific) and use each model to estimate the parameters. Then, we calculate the likelihood of the observed rebroadcast in the holdout sample using the mean of the posterior parameters. We also forecast each individual's rebroadcast decision for each message in the holdout sample. We assign a value of 1 (or 0) if we accurately (or inaccurately) predict the rebroadcast decision. The hit rate reported in Table 6 is the average across individuals and messages in the holdout sample.

We begin with Model 1, a baseline model which only includes content effects and ignores any individual-specific, fit, and influence effects. Model 2 only includes individual-specific effects and ignores content, fit, and influence effects. We expect that content, user, and fit effects are important drivers of rebroadcasting. Therefore we compare the value of each component to assess the extent to which they improve model fit. 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.

We find that with each added component, model fit improves in terms of deviance information criterion (DIC), log-marginal density, and the likelihood of the observed data on a holdout sample conditional on the posterior estimates. These results clearly show the value for each component in our proposed model specification. We next discuss the results associated with the full model specification (Model 4).

Table 6. 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 $g(INFL_{ijt})$				X
DIC	2,550,541	84,486	83,281	**81,759**
Log Marginal Density	-1,275,515	-42,048	-40,675	**-40,227**
Likelihood of random 25% holdout	-138,335	-11,012	-10,575	**-10,036**
Hit Rate (25% holdout)	0.734	0.991	0.992	**0.993**

Content, User, and Interaction Effects

Table 7 provides estimates for the effects of message content (C_j), user interests (U_i) and the fit between the two (FIT_{ij}) on the probability, P_{ij}, that individual *i* rebroadcasts message j at some point in time. First, our results (see Table 7) show that broadly speaking, the content of social media messages plays an important role in affecting rebroadcasting behavior. Specific to our empirical example, we find significant variation in the baseline rebroadcasting rates across content, as some content (school and politics with $\beta_2 = 0.087$ and $\beta_4 = 0.009$, respectively) is more likely to be rebroadcast while other content is not (finance with an insignificant β_3). In addition, messages from schools that broadcast more frequently have less rebroadcasting activity ($\beta_1 = -0.007$), possibly due to information overload diminishing the tendency with which any one message is rebroadcast (Edmunds and Morris 2000).

Second, our results show that rebroadcasting behavior varies significantly across individuals. One user-level characteristic that drives this variation is the type of content one tends to broadcast. Specific to our empirical example, users interested in "school," "finance," or "politics" differ in their baseline probabilities of rebroadcasting ($\beta_5 = 0.071$, β_6 is not significant, and $\beta_7 = 0.059$, respectively). Finally, users who tend to post more frequently are also more likely to rebroadcast ($\beta_{11} = 0.052$).

Finally, we also examine whether content that fits with an individual's interests impacts rebroadcasting behavior. After accounting for variation across individuals in their rebroadcasting tendencies, we find that users are more likely to rebroadcast content that matches their own interests ($\delta = 0.041$). There are two important implications of this result. First, this finding shows 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 and explains that some users may be more likely to rebroadcast messages with certain content (potentially due to their interests) than others. Second, this result suggests that organizations can tailor content to match their followers' interests in order to increase rebroadcasting activity from them. We further explore the value of tailored content in the simulation section.

Table 7. Model 6 Results -- Decision to Rebroadcast

Component	Variable	Parameter Estimate	Standard Error
	Intercept (β_0)	**-1.093****	(0.012)
Content	*SFREQ* (β_1)	**-0.007****	(0.001)
	F_1 (β_2)	**0.087***	(0.004)
	F_2 (β_3)	-0.001	(0.006)
	F_3 (β_4)	**0.009****	(0.004)
User	$AVGF_1$ (β_5)	**0.071****	(0.008)
	$AVGF_2$ (β_6)	0.005	(0.008)
	$AVGF_3$ (β_7)	**0.059****	(0.007)
	FOLLOWERS (β_8)	**0.032****	(0.007)
	FOLLOWING (β_9)	**-0.033****	(0.007)
	RATIO (β_{10})	0.007	(0.017)
	UFREQ (β_{11})	**0.052****	(0.006)
Content-User Fit	*FIT* (δ)	**0.041****	(0.013)

**zero is not contained in the 99% confidence interval
*zero is not contained in the 95% confidence interval

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 8). 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 will be affected by the number of influential users who have previously rebroadcast (α = 0.418).

Table 8. Model 6 Results -- Timing of Rebroadcast

Parameter	Description	Parameter Estimate	Standard Error
$\bar{\lambda}$	Rebroadcasting Rate (scale)	0.006	(0.010)
c	Rebroadcasting Rate (shape)	0.945	(0.009)
α	Susceptibility to Rebroadcast	**0.418****	(0.016)
		Mean Effect	**Standard Deviation**
π_u	Probability of Being Influential	0.458	(0.208)
ϕ_i	Probability of Being Susceptible	0.312	(0.102)

Note. We report the empirical mean and distribution of π_u and ϕ_i
**zero is not contained in the 99% confidence interval

Our results indicate that there is considerable heterogeneity in both a user's ability to influence as well as a user's susceptibility to influence. Figures 2 and 3 show the distributions of the posterior means of π_u and ϕ_i across individuals. These results suggest that there are relatively few individuals who are highly likely to be influential or are highly susceptible to influence.

Overall, our findings suggest 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 is not readily apparent without jointly considering the susceptibility of the population exposed to influence. Influentials 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, 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.

Figure 2. Distribution of Influence (Posterior Mean π_u)

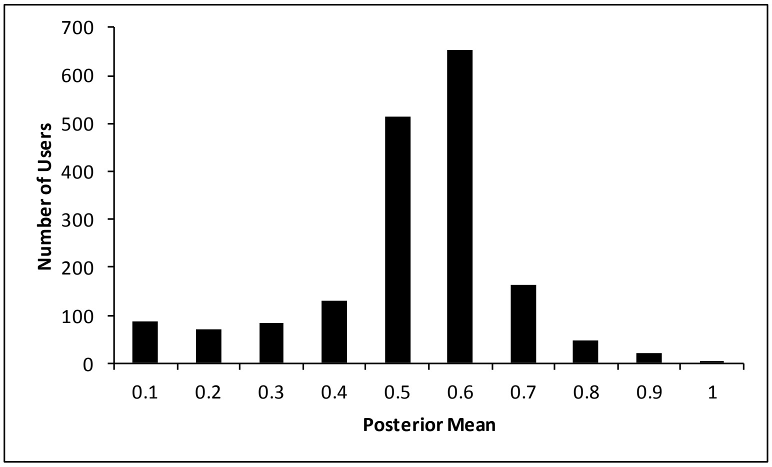
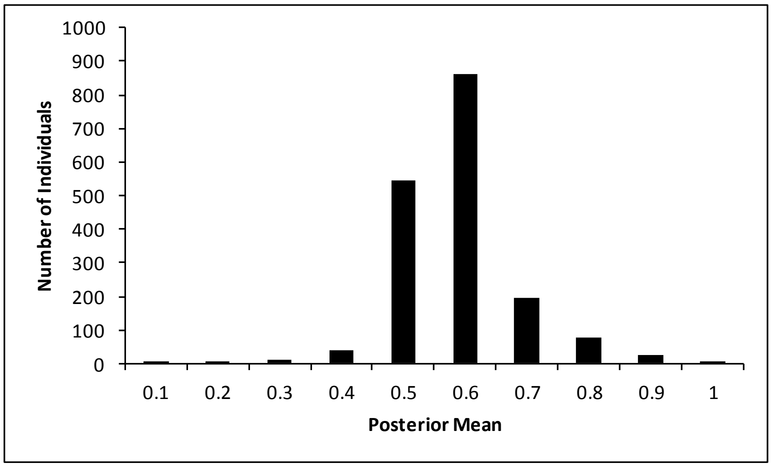


Figure 3. Distribution of Susceptibility to Influence (Posterior Mean ϕ_i)



Simulating Social Media Messaging Strategies

The proposed method described above allows managers to jointly model the effects of content, fit, influence, and susceptibility on rebroadcasting activities. In this section, we provide insight into the relative value of each component by generating simulated scenarios using the model estimates. This will help managers to make decisions in investing their resources in only content design, only seeding influencers, or a combination of both. Traditionally, researchers have recommended various seeding strategies consisting of targeting certain users to spread their messages or purchase decisions. For example, they recommend that marketers should seed influential users (Aral, Muchnik, and Sundararajan 2013; Libai et al 2013) or users with a certain local network characteristic (Goldenberg et al 2009; Hinz, Skiera, Barrot, and Becker 2011; Katona et al 2011; Libai et al 2013). However, our results suggest that in addition to identifying and targeting influential users with incentives, marketers interested in increasing the reach of their social media messaging through users' rebroadcasting should also consider the following. First, firms can design their message content to focus on topics more likely to be rebroadcast. Second, given the importance of content-user fit in encouraging rebroadcasting behavior, marketers can also create content designed to appeal to the influential population.

We consider each of the aforementioned scenarios as a potential strategy to increase the reach of social media messages and use simulations based on our posterior estimates to assess their relative effectiveness. We simulate the rebroadcasting behavior related to a specific message by the 2,260 individuals in our sample using the following method. First, for each individual, we compute *i's* probability of rebroadcasting (P_{ij}) using individual *i*'s and message *j*'s model-based posteriors and simulate the decision to rebroadcast the message through Bernoulli draws with probability P_{ij}. Second, for those who will eventually rebroadcast, we simulate the timing of every *i*'s rebroadcast using the posteriors from their hazard component and designate the individual with the lowest simulated *t* as the initial rebroadcaster.

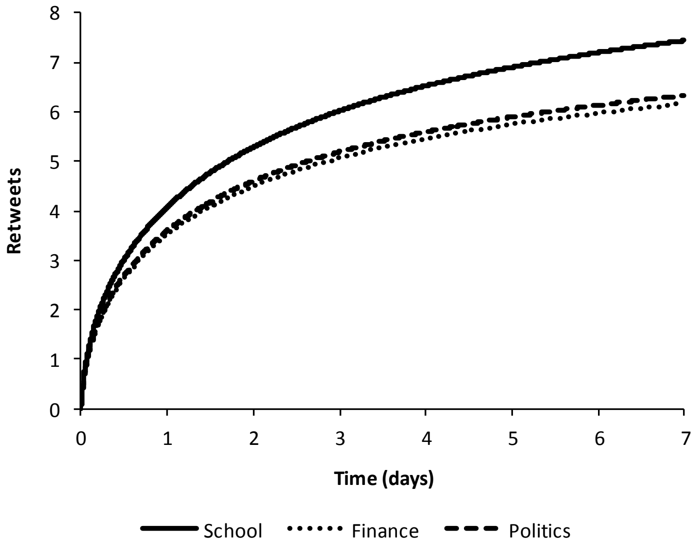
An important feature of our simulations is that we account for the impact that influentials who have already rebroadcast the message may exert on others. To do so, in each simulation iteration, we probabilistically determine whether each individual is influential based on his/her posterior π_u. Then, following the initial rebroadcast of a message, we simulate the timing of the next rebroadcast conditional on the influence that earlier rebroadcasters may have. 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. This procedure is repeated so long as the rebroadcasts occur within a specified timeframe (one week). We use the posterior estimates from the last 1,000 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 which we will describe below. We average across 500 simulated iterations for each set of posteriors and then across the 1,000 estimation iterations to obtain our simulation results which we present next.

Simulation Study 1: Focusing on content

In our first simulation, we consider the rebroadcasting behavior associated with different messages, each varying in terms of content. This simulation highlights the extent with which firms can rely on designing content to drive rebroadcasting activity (without the impact of influencers). We generate messages by using the factor scores to replicate three different messages focused on school, finance, or politics. To create a message emphasizing the "school" topic, we specify that message to be one standard deviation above the mean across our entire sample of messages on the school factor. We also specify that message to have a finance and politics score that is the average value across our entire sample. Thus, the "school" message will have a school factor score of 1 and politics and finance scores of 0. Likewise, we specify a finance (or politics message) with factor scores of 1 for finance (or politics) factor and 0 for the other factor scores. With these message factor scores, we calculate *Pij* and simulate a sequence of rebroadcasting behavior using the previously outlined simulation procedure.

The results from this simulation are presented in Figure 4 and suggest that a school-related message is expected to be rebroadcast more than other content, increasing rebroadcasting activity by approximately 20% over finance- and politics-related messages over a one week period of time in our empirical example. While these rebroadcasting results are specific to our context, they highlight the importance of identifying and designing proper content that inherently generates more rebroadcasting activity.

Figure 4. Simulation Comparing Content



Simulation Study 2: Identifying and seeding influentials

In the second simulation, we examine the impact of targeting (or "seeding") influentials on rebroadcasting (without designing content). For the purposes of our simulation, individuals identified as influentials and targeted by the seeding strategy will rebroadcast the message immediately following its original broadcast (in practice, firms can pay or incentivize influential to rebroadcast their messages immediately). We identify influentials using our posterior estimates for $\hat{\pi}_u$, with higher values corresponding to those who have the highest likelihood of being an influential.

In this simulation, instead of varying the content as we did in the first simulation, we assume that a firm releases a message that is average on all factor scores (i.e. has 0 for all factor scores). We then target the top 2 individuals most likely to be influentials (0.1% of our total population) as ranked by π_u and simulate a rebroadcast from them at *t* = 0. We compare this targeting strategy with a random-seed strategy in which we randomly chose individuals to immediately rebroadcast.

We record all those who rebroadcast within the first week, excluding the initial seeded rebroadcasts. Table 9 shows the results of the no-seed strategy and the incremental percentage increase in rebroadcasts using random-seed and influential-seed strategies. We vary the number of seeded individuals (2, 5, or 10 users).

Consistent with expectation and our parameter estimates, seeding influentials accelerates the rebroadcasting process during the initial week. Interestingly, randomly selecting individuals for seeding, in the 2 seed (0.1%) case, performs only slightly better than a no seeding strategy, highlighting the value of identifying influentials. In addition, we find that seeding 5 (or 10) random users produces similar 1 week cumulative rebroadcasting activity as seeding 2 (or 5) influentials. This suggests that seeding strategies focusing on influentials are more efficient than strategies that rely on the volume of individuals seeded (i.e. randomly seeding a large amount of individuals). To compare these results with those of Simulation 1, we find that creating effective content (i.e. “school”) outperforms randomized seeding (at the levels we specify). However, seeding ten influentials results in greater rebroadcasting activity during the first week when compared to a “school” related message.

Table 9. Simulation Comparing Seeding and Content-Fit Strategies

	No Seed	Seed Random	Seed Influential	FIT Influential
2 users (0.1%)	6.93	+1.2%	+4.8%	+6.9%
5 users (0.25%)	6.93	+4.5%	+7.2%	+8.1%
10 users (0.5%)	6.93	+6.9%	+10.0%	+5.1%

Matching Content to Influentials' Profiles

Of key interest in our results is the fit between content and user. Given the importance of the user-content interaction in rebroadcasting, we perform another set of simulations to assess the extent to which marketers can benefit from both designing content and targeting influencers. Specifically, we test whether developing content that matches the interests of the influential a viable alternative strategy.

In this simulation procedure, instead of seeding influentials – as done in the second set of simulations – we instead develop content that closely fits with the profile of influentials. Thus, instead of relying on incentivizing influential, firms could alternatively focus their resources on creating content tailored to these influential to speed potential rebroadcasting activity organically. Operationally, we identify the top influentials and generate different messages, with each matching one of the influentials. For example, in the "seed 2 influential" condition, we release two different messages (targeted at the two influentials) to the entire population and record the resulting rebroadcasting activities for both messages. We then average the rebroadcasting activities of the non-targeted users across the two messages and report the percentage increase above the no seed strategy in Table 9. We choose to release two different messages in order to target two influentials with potentially different preferences. This allows us to compare this content targeting strategy with the other seeding strategies that also targets two influential users. In the 5 (or 10) influential condition, we release 5 (or 10) different messages targeted at the top 5 (or 10) most influential individuals.

Table 9 also shows the rebroadcasting activity resulting from these simulations. We also compare the results to those from the influential-seeding strategy. The results from the content-fit strategy slightly outperform that of seeding influentials in the 2 and 5 seed scenarios. That is, in our simulation, we gain greater rebroadcasting activity by tailoring content to the top 2 or 5 influentials than by creating “average” content and incentivizing those 2 or 5 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 or 5 individuals, we see an unexpected dip in future rebroadcasting activity when we target 10 users. A key factor contributing to this may be that content suited to influentials does not appeal to the masses. We explore this explanation by examining the fit between content that appeals to influentials with the entire population. First, we create a message with factor scores that equal the average factor scores across all messages in our data set (i.e. an average message). Second, we record the factor scores of the 10 messages that were used in the content-targeting simulation (10 influentials condition). Finally, we calculate the fit between each individual in our population with (a) the “average message” and (b) each of the 10 targeted messages. Overall, the average content-user fit with the “average message” in (a) is -1.74. In contrast, the average content-user fit (averaged across all individuals and across the 10 messages) with the targeted messages in (b) is -2.11. In other words, while influentials may be more likely to rebroadcast content that has been designed for them, subsequent message propagation by the masses, with whom the message may or may not resonate, is less likely. As such, there appear to be limits to how effective an organization can be when they adopt strategies that try to leverage influential users in the spread of social media messages by tailoring message content for them.

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 show that each of these drivers plays an important role in affecting rebroadcasting behaviors. We find that rebroadcasting activity depends on the content of the message. We also find that active rebroadcasters tend to rebroadcast messages with certain content, have many followers, and tend to more frequently broadcast messages. Furthermore, we find 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 existence of a limited number of influentials whose rebroadcasting is related to subsequent rebroadcasting by others, underscoring the 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 and the role of influentials in rebroadcasting behavior. Specifically, we show how targeting influentials to encourage their rebroadcasting of our message can lead to greater rebroadcasting activity than investing in message content. However, we also show that, under certain circumstances, tailoring message content to the interests of the influentials can generate even greater rebroadcasting activity. This has implications for marketers, suggesting that strategies targeting influentials can yield greater rewards compared to investing in message content.

However, our research is not without its limitations. First, similar to other empirical work examining social effects, we do not directly observe influentials impacting others. Instead, we are only able to probabilistically assess those who may facilitate ongoing rebroadcasting. Second, 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 the purchasing behavior of others.

Learning From Online Social Ties

Introduction

Online opinions have been shown to influence a wide variety of consumer choices. Consumers commonly read about others' purchase experiences and make their decisions based on this information. This is especially prevalent on online platforms such as Amazon.com, Rottentomatoes.com and Tripadvisor.com, where many different opinions are readily accessible by a wide audience. Due, in part, to the importance of online opinions to consumers' decision-making processes, marketing scholars have studied a great deal the impact of online opinions on observed actions. Researchers have demonstrated that online opinions and ratings play a significant role in purchase decisions across a range of markets and settings (e.g. Godes and Mayzlin 2004; Chevalier and Mayzlin 2006; Clemons et al. 2006; Liu 2006; Dellarocas et al. 2007; Duan et al. 2008; Li and Hitt 2008; Du and Kamakura 2011; Chen et al. 2011; Moe and Trusov 2011).

While it is clear that online opinions *influence* consumers' purchase decisions, we have limited knowledge as to the outcomes of those decisions. Specifically, there has been no inquiry into the extent to which access to opinions helps *improve* future decisions and, if so, whether individuals undergo learning experiences to aid this improvement. On one hand, from a rational perspective, we would expect that opinions would improve decision quality since consumers continue to use reviews in a wide range of settings. If consumers experience systematically-worse decision quality, we would expect they would refrain from using this information. So, we expect that, on average, decisions should improve when facilitated by online opinions. This may be especially true of information from weak ties, as the previous literature (i.e. Granovetter

1973) suggests that these opinions may be more novel and beneficial. On the other hand, it is not obvious that *all* opinions will be equally helpful. In particular, we consider the extent to which one needs to learn, in the form of decision-making experiences, before others' opinions become useful.

We expect that learning may be critical because using another poster's opinions to infer one's own expected satisfaction with a product may be difficult, particularly if the consumer has limited previous interactions with the other poster, knowledge of her preference structure, or experience in the purchase domain. Early on in one's participation in a community, the ties one makes and the opinions derived from them may not immediately lead to strictly better decisions due to the difficulty in interpreting these opinions. We define interpretability of opinions, in this context, as the reader's ability to infer a mapping from the expressed opinions (e.g., online ratings) to her own preference structure. All else equal, greater interpretability should help improve decisions.

The complexity associated with interpretability may derive from heterogeneity across individual posters. For a given product (book, movie, hotel, e.g.), one commonly finds a great deal of variance across the opinions of posters. Therefore, it may be difficult to interpret the review from one poster and map that information to one's own preferences without undergoing learning experiences. Complexity in interpreting opinions may also arise due to significant heterogeneity in ratings across different online communities which would, again, require learning and expertise-building on the part of new members of the community before they would be able to fully appreciate, and benefit from, the information. Without understanding these norms, we expect that consumers will have a more-difficult time mapping others' opinions to their preferences: it may be difficult to figure out what ratings are associated with a "good" book.

However, consumers should gradually improve their ability to interpret the opinions and make better decisions as they learn the norms of both the community and of specific posters. Given this, we expect that consumers who have more learning experiences with decision making within a community will more-effectively use and interpret the opinions provided in the community in order to make better decisions.

In addition to learning, we also investigate the impact of the *source* of the opinion -- whether arriving from strong or weak ties -- on decision quality. We consider the extent to which one builds ties with -- and gathers information from -- others with preferences that are, or are not, relatively well known (and possibly more-similar) to the consumer. That is, we investigate the relative impact on decision quality of building more strong vs. weak ties. While the extant literature suggests that one may learn more from the latter, we expect the initial interpretability of information to be higher from the former: it should be easier *ceteris paribus* to assess the match between one's preferences and those of strong ties as compared with weak ties. That is, in the absence of learning experiences, one should be able to better interpret opinions from friends than from acquaintances. Therefore, we expect opinions from strong ties to be, initially, more valuable.

However, opinions from weak ties may be particularly valuable due to the novel information they make available (Granovetter 1973). Critical to their usefulness, however, is the extent to which consumers are able to learn how to interpret and utilize information from these ties. We see this as a skill that may need to be learned and developed over time. Therefore, we suggest that opinions from weak ties may initially result in lower levels of improvement, if any improvement at all, in decision quality compared to those from strong ties -- to the extent that the information from weak ties is less-easily interpretable by the reader. However, opinions from weak ties should provide greater improvement, conditional on undergoing learning experiences that improve their interpretability.

In order to study the impact of social information on decision quality within a social network, we investigate the actions taken by a set of consumers from the time they join an online opinions community. We explore this dynamic learning process by exploiting the unique characteristics of the online review platform, Goodreads.com. Unlike the majority of online review websites such as Amazon.com, Goodreads provides not only ratings but also a social network which allows for interaction among individuals. We follow the actions of new Goodreads users as they establish a social network and learn how to make use of the information their network makes available. Consumers on Goodreads can form either bi-directional "friend" relationships or uni-directional "source" relationships. We assume that the bi-directional relationship reflects more-frequent interaction and a closer relationship since both individuals observe each other's opinions. In contrast, the uni-directional relationships suggest more-limited interactions since one sees the opinions of one's sources but not vice versa. To capture the quantity of information received from friends or sources, we measure the evolving size of the networks built by those in our sample over time. Thus, we are able to assess the marginal improvement in decision quality of adding new information sources.

After controlling for the ratings environment, individual-level factors, book-level dynamics, self-presentation effects, and, importantly, potential endogenous network variables, we find that consumers report higher satisfaction -- i.e., they post higher ratings -- the more bi-directional ties they have previously formed. We interpret this as suggesting that opinions drawn from a larger set of strong ties lead to better decisions. In comparison, this is not always true of weak ties. As consumers develop more uni-directional ties, their decision quality initially declines; decisions get worse the more weak ties they have formed. However, as one develops more experience in the community, information from weak ties, in fact, improves decision quality. Indeed, the improvement effected by information from weak ties, for sufficiently-experienced users, eventually outstrips that for strong ties. These results suggest that, when one lacks experience using socially-acquired information, friends may be the better source for acquiring it. However, conditional on developing knowledge of the norms and customs associated with the network -- once the user is capable of clearly interpreting the meaning of the messages being transmitted -- information from weak ties may be more useful than that from strong ties.

These results make several important contributions to the literature. To our knowledge, we are the first to explicitly address the impact on decision quality of the development of online relationships and the use of online word-of-mouth. Moreover, the role of consumer learning and its connection to decision quality has not been addressed in the literature. We also demonstrate that there is a significant and important difference between the value of information one accesses through deeper, stronger two-way relationships as compared with weaker one-way relationships. The extant literature on online reviews has implicitly treated all ties as equal while it is clear from our results that this may not always be correct. Finally, and perhaps most important, our results suggest a dynamic process whereby consumers not only make better decisions as they forge more relationships but they also *learn to make increasingly-better decisions as their expertise is enhanced*. In this sense, our results suggest that relationships, while beneficial, are not equally-valuable to all. Those who invest in developing the expertise to use the information these relationships make available will benefit most from them.

The rest of the paper proceeds as follows. In the next section, we review how our work fits into the relevant literature. Then, we develop in more detail the theory behind our empirical inquiry. Next, we describe the dataset. In the Model Sections, we develop our model and discuss the estimation approach. Finally, we present our results and conclude with a discussion of managerial implications, possible extensions, and limitations.

Related Literature

Our paper relates to three streams in the existing literature: online opinions, herding and cascades, and learning. A wide variety of studies have consistently shown that online social interactions and opinions drive purchase decisions. Most existing research on online reviews may be categorized into one of two groups: those that have analyzed the impact of online opinions on sales, or choices in general, on one hand, and the product-level dynamics of posted ratings, on the other. With respect to the former, there exists strong empirical evidence linking aggregate measures of online opinions (i.e. the valence, variance, dispersion or volume of posted ratings) with choices for television shows (Godes and Mayzlin 2004), books (Chevalier and Mayzlin 2006; Li and Hitt 2008), movies (Liu 2006; Dellarocas et al. 2007; Duan et al. 2008; Chintagunta et al. 2010), beer (Clemons et al. 2006), cameras (Chen et al. 2011), and bath and beauty products (Moe and Trusov 2011).

Specifically, Godes and Mayzlin (2004) report that broader dispersion of opinions across different communities is associated with higher television ratings. Others have shown that positive valence, in the form of higher average ratings, increases sales (Chevalier and Mayzlin 2006; Dellarocas et al. 2007; Chintagunta et al. 2010; Chen et al. 2011; Moe and Trusov 2011). Liu (2006) and Duan et al. (2008) also consider the impact of the volume of ratings on movie box office revenue. While these studies suggest that online opinions have a significant impact on consumer choices, they provide no evidence with respect to their impact on the quality of these choices. Theoretically, consumers will only continue to place weight on others' opinions if the information helps them make better choices. Our study offers the important contribution of identifying the existence of this improvement as well as moderators of the effect.

Recent research has also analyzed the impact of product-level dynamics in the time series of ratings (Li and Hitt 2008; Wu and Huberman 2008; Godes and Silva 2012; Moe and Schweidel 2012). Most notably, there appears to be a robust downward dynamic pattern in ratings. While Li and Hitt (2008) argue that the root cause of the pattern is self-selection, Wu and Huberman (2008) suggest that it is an outcome of strategic decisions by posters who only post when the review will have sufficient impact. Godes and Silva (2012) decompose the effect into a temporal and sequential effect and suggest that the latter is due to purchase errors increasing as more reviews arrive. Notably, their theory is tested based on the assumption that ratings are a proxy for decision quality. Though our model is a dynamic one, the nature of the dynamics we investigate is quite distinct from the focus of the existing dynamics literature. Specifically, these papers have all studied *product-level* (in most cases, book-level) dynamics. That is, they attempt to explain why, for a given book, ratings appear to exhibit a downward trend. On the contrary, we focus on *individual-level* dynamics: how does an individual's decision quality (as captured in her reported ratings) change as she gathers more experience and creates more ties in a community? Of course, we control in all analyses for the product-level dynamic factors found to be important in the existing literature.

It is worth noting that we are not the first to study the dynamics of decision quality as a function of information gathered from others in a social setting. The cascades and herding literature shows that more information may not always help consumers to make the right decision. Bikhchandani et al. (1992) show that under certain situations, individuals will make choices based on the observed behavior of previous consumers, while ignoring their own private information. Banerjee (1992), similarly, calls this "herding behavior," where individuals use other's observed choices as information on which they rely instead of their own private information. In both cases, decisions made based on observational information can potentially cause consumers to make incorrect choices.

Our study differs in a number of important respects. First, we demonstrate that decisions overall improve as one builds more relationships online. Of course, we also show that decision quality may decline in some cases, particularly early during one's tenure in a community. Importantly, however, the mechanism that decreases decision quality in our results is quite different from that driving the results in Bikhchandani et al. (1992) and Banerjee (1992). They both focus on *observational learning* and their results are driven by inferences drawn based on ambiguous outcomes[6]. In contrast, we study a context in which *information* is transmitted among members of an online community. Thus, while the herding/cascades literature focuses on uncertainty that future buyers have with respect to the set of information that earlier buyers had, the mechanism we investigate is based on consumers' inability to interpret the information they acquire from others in the online community. There is, in some sense, no inter-consumer information asymmetry here but only an initial inability to map the observed information to one's preferences. We stress again, as well, that here we focus on intra-consumer (i.e., individual-level) dynamics as opposed to the inter-customer process modeled in the herding/cascades literature.

[6] Specifically, both the results in Bikhchandani et al. (1992) and Banerjee (1992) are based on one's uncertainty as to whether those preceding them have acted upon private information or not. In Bikhchandani et al. (1992), for example, one flips a fair coin when previous observations leave one indifferent. However, future buyers do not know that she flipped a coin and, thus, place positive probability on the fact that her decision was based on the preponderance of information.

Our work here also relates closely to the broader literature on learning. According to our theory, consumers engage in an individual-level learning process when using online opinions over time. In order to improve decision quality, we expect that consumers learn how to use these opinions[7]. Learning, in general, may follow two different paths: learning by instruction and learning by doing (Nokes and Ohlsson 2005). Given that there are no clear instructions on how to use online opinions, we focus on the latter and expect that consumers learn how to make better decisions using online opinions over time, based on their experiences. Learning by doing places emphasis on repeated trials where individuals develop more-efficient or more-effective solutions (Anzai and Simon 1979).

While researchers have not explored learning by doing in a social media and online opinions context, they have demonstrated the existence of such a process in numerous other settings. For example, Foster and Rosenzweig (1995) show that farmers learn how to better use new seed varieties through past experiences. Borgatti and Cross (2003) show that scientists learn how to seek information from others.

The learning process we propose may be driven by the acquisition of both declarative and procedural knowledge (Lakshmanan et al. 2010). In the context of online reviews, declarative knowledge may be the norms of a certain ratings community or specific individual. By having this knowledge, consumers can better understand what the posted ratings, made within an online ratings community or made by certain opinion posters in their network, represent. In contrast, procedural knowledge may allow consumers to better make use of specific opinions. Through

[7] In another related paper, Zhao et al. (2013) find that consumers place more emphasis in a choice context on information from online reviews than past experiences with comparable products. Our papers should be seen as complements in that (i) Zhao et al. (2013) study the effects of reviews on choice while the outcome with which we're interested is the quality of purchases decisions; and (ii) the learning effects in their paper are with respect to preferences for products and product categories while our conceptual model reflects learning about how to make use of the information made available by new network ties.

experience, a consumer may learn how to use opinions from specific individuals in their network to more effectively make purchase decisions. For example, she may find that some individuals are effective sources of information on one genre of books but not on another. More generally, she may learn how to find “experts” in a given genre in which she has not previously purchased.

Since opinions may differ widely across both individual posters and ratings communities, we expect that online consumers may initially need to learn about the norms of a community or specific posters in order to better match the opinions to their preferences. It is especially critical to learn how to learn from weak ties, as we expect that it is initially difficult to interpret the reviews they provide. Alba and Hutchinson (1987), in a study of expertise, argue that familiarity leads to experience; experienced individuals are better at evaluating information because they process it at a deeper level and distinguish between relevant and irrelevant information. Experience has also been shown to decrease cognitive costs involved in learning (Johnson et al. 2003). As a result, we expect that with experience, consumers will *learn how to learn* from online opinions in order to become better at using this information in purchase decisions. To our knowledge, we are the first to investigate the relationship between experience and the value of socially-acquired information.

Theoretical Development

When an individual joins, and begins to develop relationships in, a new community, we expect that she will be influenced by the information provided by these relationships. Indeed, this information is likely to be one of the motivations for joining the network to begin with. Our primary focus here is on the learning process associated with experience in using online opinions to make decisions. We expect that, on average, an individual acquiring information from others should be able to better interpret, and make use of, that information the more experience she has in the community and with that information. With additional experience interacting with others in a community, we hypothesize that consumers engage in “learning by doing” (Anzai and Simon 1979; Nokes and Ohlsson 2005; Lakshmanan et al. 2010) and, thus, improve their ability to use information from these social ties. In turn, this should improve the outcomes of their decisions which are based on this information.

While we expect all relationships to be characterized by improved interpretability as the parties get to know each other -- as our relationship deepens, we are better able to assess what someone means by a “good movie,” for example -- we expect the nature of this dynamic improvement to vary across different types of relationships. A particularly-important source of variation across relationships is the strength of the tie characterizing a dyad (Granovetter 1973). Researchers have typically defined stronger ties -- often referred to as “friends” -- as those characterized by higher frequency of interaction, more trust and the ability to work together more productively (Granovetter 1973; Lin et al. 1981; Brown and Reingen 1987; Krackhardt 1992). Research has also demonstrated that, in established relationships, weaker ties are often associated with the transmission of more-useful and valuable information (Granovetter 1973; Brown and Reingen 1987; Levin and Cross 2004). This follows, in part, from the impact of homophily: since we typically form strong ties with those with whom we are most similar, they are less likely to tell us something new (McPherson et al. 2001). In contrast, weak ties give us more novel information, which may be beneficial in improving decision quality. Based on this, we hypothesize that -- in the long run -- forming weak ties should, all else equal, lead to a greater improvement in decision quality as compared with the formation of strong ties.

While this may be the case in the long run, we expect that information gathered from strong-tie relationships will be relatively-easily interpretable at early stages of the relationship. This may be due to several reasons. First, it may simply be due to homophily: we are more similar to our strong ties, thus, increasing the interpretability of opinions from these ties. The analogy we use here would be that of a college freshman. Upon arrival on campus, she is likely to form relationships with others who are, say, from her hometown or hometowns similar to hers as well as with others from entirely different places. Initially, information from the former will be easily understood: they reference similar benchmarks or are calibrated against known and shared ideals. On the contrary, information gained from the latter may be more difficult to understand[8]. Second, differences in the interpretability across ties may also be due to differences in the frequency of interactions: because we interact more often with strong ties, even those with different backgrounds, we are able to learn quickly how to process, interpret and apply the information gathered from these relationships. In contrast, we may have fewer interactions with weak ties, hindering our ability to interpret the other's preferences and to apply their insights to our own decisions.

So, while we expect the initial level of interpretability to favor strong ties, we expect that the dynamic improvement will be greater for weak ties. Specifically, we hypothesize that, when one initially joins a community, information acquired from strong ties is more valuable than information gained from weak ties due to its easier interpretability. However, as the individual gains experience in the community and is better able to interpret the information gained from the latter, the role of novelty will dominate, implying that weak-tie relationships deliver more-valuable information. The key implication of this point is that *strong-tie relationships deliver*

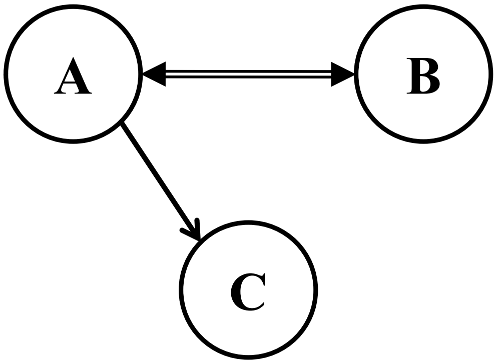
[8] As an admittedly-exaggerated example, a student from New York City would have no problem interpreting recommendations from another student from New York City regarding an Italian restaurant near campus while a recommendation from a student from a rural area of the Midwest United States may be harder to assess.

information that is more-immediately useful while weak-tie relationships require a period of learning -- within the community -- before bearing fruit. In the online community context, then, we expect that the formation of strong-tie relationships will lead to immediate improvements in decision quality while, due to the learning process, the formation of weak-tie relationships will take longer to yield an impact. However, to reiterate, after this learning process runs its course, all else equal, we expect weak-tie relationships to be more impactful than strong-tie relationships.

The foregoing theoretical discussion notwithstanding, it is important to acknowledge that, in reality, tie strength is difficult to assess in secondary data. While traditional survey-based approaches such as those employed by Granovetter (1973) and Godes and Mayzlin (2009) allow for precise and direct assessment of the nature of the tie, those working with secondary data, as we do here, must rely on proxies. A common approach used by several researchers has been to approximate tie strength by the frequency of interaction in a dyad (Granovetter 1973; Lin et al. 1981; Brown and Reingen 1987; Krackhardt 1992). In our analysis of an online review community, however, (observable) interactions do not occur, on a one-to-one basis, precluding this approach. Thus, we make use of a different approximation for the strength of the tie. As in many network settings, we are able to observe two different types of relationships between individuals: uni-directional and bi-directional, where the latter is designated as a "friend." In an online setting, a uni-directional relationship may form when one follows the opinions of, i.e., "listens to," another without the relationship being reciprocated. For example, it is common in Twitter for one to follow others without the reverse being true. Such a relationship is shown in Figure 5, where A follows C but C does not follow A.

We propose that such a uni-directional link represents a weaker tie than a bi-directional link[9]. We expect such a relationship to be characterized by a lower degree of homophily, and relatively-limited interactions between the two individuals. Indeed, it seems reasonable to expect that a strong-tie relationship, such as that characterized by a "friend", would be reciprocated (Granovetter 1973; Lin et al. 1981; Brown and Reingen 1987). Thus, we expect A to have a more-limited understanding of the views, biases or preferences underlying the opinions offered by C and will, therefore, find it more difficult to interpret the information transmitted via such a relationship. In contrast, bi-directional relationships are mutually accepted, where both parties must confirm the tie. These relationships have generally been classified as "friends" (e.g. "Facebook friends," "Goodreads friends"). While the uni-directional tie is more common in Twitter, all relationships in Facebook, for example, are bi-directional in that they require mutual consent. In Figure 5, A and B have such a bi-directional relationship which, we assume, represents a stronger tie than the uni-directional link.

Figure 5. Network Ties



In summary, we propose an additional dimension in the comparison of information from strong- vs. weak-ties: interpretability, or mapping. We argue that it is easier to learn how to use,

[9] We also consider alternative measures of tie strength, based on similarities across demographics in age and gender. However, these models (available from the authors) did not fit as well as that based on uni- and bi-directional ties. This is expected, given that these, and potentially other, demographic measures tend to capture observed heterogeneity rather than tie strength.

to interpret, the opinions offered by those with whom we interact more frequently. This may be due to homophily (Lin et al. 1981; McPherson et al. 2001), a higher probability of exchanging information (Krackhardt 1992) and/or better information transfer and sharing (Uzzi 1997; Kraatz 1998; Hansen 1999; Reagans and McEvily 2003; De Bruyn and Lilien 2008). As a result, while the extant literature predicts that more-useful information comes from weak ties, this may not be the case at earlier stages of a relationship. In an online review context, we expect one to learn quickly the norms of strong ties -- for example, the ratings cut off points that an individual uses when determining the rating to assign to a given product -- but more slowly those of weak ties. As a result, the marginal impact on decision quality of forming new strong-tie relationships -- as captured by bi-directional ties -- will be observed early on in one's experience in a community. Conversely, the impact on decision quality of forming new uni-directional ties -- those one may simply follow -- may be greater than those of the bi-directional tie but will exist only following a more-lengthy learning process.

Data

Our data are collected from Goodreads.com, an online book review website that integrates online opinions, in the form of ratings and reviews, with a comprehensive social networking platform. Since we expect individual dynamic learning effects to occur early in one's participation in the community, we perform the data collection from the individual's perspective rather than the book's. This allows us to capture one's earliest experiences in the community. We chose to collect new users because we expected higher within-subject variation in their social network over our collection period. Moreover, due to the power law associated with learning, we expect that learning will improve rapidly at first, but future incremental improvements will require more time and effort (Newell and Rosenbloom 1981; Johnson et al. 2003). Our sample consists of the first 25,000 user id's who joined Goodreads in April 2010. We use such a large sample since, as is typical in online platforms (see, e.g., the “90-9-1 rule for participation,” Nielson 2006), the majority of new users had little to no activity on the site or removed their accounts from the community. Our large sample size means we are only able to obtain social network data every 2 days in order to comply with Goodreads' Application Programming Interface (API) limitations. Using these data, we track the dynamic social structure that evolved over 50 days in April and May of 2010 for each individual in the sample in order to obtain the dynamic evolution of their complete social network along with their rating observations. We removed all users who were inactive or deleted their account during our observation window and only focus on individuals who have observable site activity. Of the 25,000 individuals in the initial sample, 5,389 individuals posted reviews or had any activity on their account; as our focus is on decision quality, those that did not post provide no insight since they show no evidence of decisions[10].

The social network on Goodreads contains several types of relationships[11]. First, Goodreads labels those connected through bi-directional relationships as “Friends”, similar to the connection between A and B in Figure 5. Friends are able to easily view each other's posted reviews. Second, Goodreads labels those who follow one's reviews -- connected through uni-directional relationships -- as “followers.” Finally, Goodreads also provides a list of people one is following, again through uni-directional ties, which we call “sources.” In the example in Figure 5, A is the follower while C is the source; therefore, C provides information to A while

[10] To be clear, we are unable to estimate a formal selection model since their non-participation implies a lack of observable data.

[11] While individuals are able to log in using their Facebook and Twitter accounts, this is only to display one's Goodreads posts on other social networking websites. Goodreads' social network is separate from Facebook or Twitter, and one needs to determine whom to add as a “friend” or “source” (although you can use facebook/twitter/email to identify potential individuals to link to).

the reverse is not true. We assume that the friend connection is stronger than the source or follower connection due to the bi-directional nature of the tie. We expect, for example, these ties to engage in more-frequent interaction and to be better at interpreting each other's opinions.

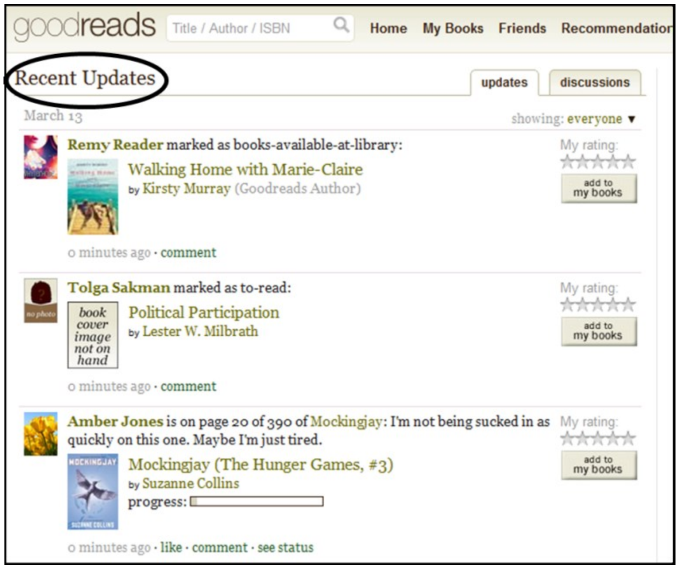
Table 10. Friends versus Sources

	Friends	Sources
Number of Ratings	**132.73***	**205.6***
Average Rating	4.01	4.05
Frequency of Ratings (ratings/day)	0.53	0.49
Time on site (days)	**419.21***	**549.18***
Number of Ties	240.10	250.72

* Indicates a significant difference between the means of Friends and Sources using a two sample t-test with a 95% confidence

To explore whether there may be other differences between the friends and sources who provide opinions to the reader (e.g., expertise), we analyze a set of observable characteristics of friends and sources who are connected to any of the 5,389 users in our sample. Some individuals may be both a friend (to one user) and a source (to another). Therefore we remove all individuals who are both a friend and a source from the comparison and evaluate only the those who fall under the friends or sources category in order to highlight greater potential differences between the two types of ties. For each group of individuals, we calculate the average of the number of ratings, average ratings, frequency of rating (ratings per day), time since they have been a member on Goodreads, and the number of ties. Table 10 shows the means across friends and sources for these measures. Importantly, there is no significant difference across the two types of ties in terms of average ratings, frequency of ratings, or the number of ties. Since some studies propose that experts post more negatively (Schlosser 2005; Moe and Schweidel 2012), this finding may suggest that sources have no more expertise than do friends. However, the sources in our sample have posted significantly more ratings than those who are classified as friends. This seems to be due to their having been members of the site longer, on average, than friends since their rating frequency is not significantly different. In addition, about 14% of all sources eventually turn into friends within the 50-day window. This frequent evolution from source to friend provides additional support for our assumption that friends are stronger ties (who may be more similar in nature or form due to increased tie strength) while sources are weaker ties (who may provide new and novel opinions).

Figure 6. Recent Updates



March 13

Remy Reader marked as books-available-at-library:

Walking Home with Marie-Claire
by Kirsty Murray (Goodreads Author)

0 minutes ago · comment

Tolga Sakman marked as to-read:

book cover image not on hand

Political Participation
by Lester W. Milbrath

0 minutes ago · comment

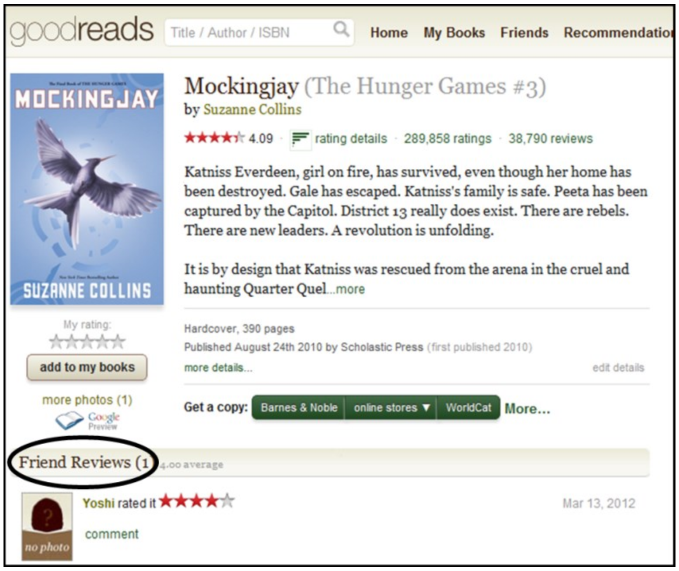
Amber Jones is on page 20 of 390 of Mockingjay: I'm not being sucked in as quickly on this one. Maybe I'm just tired.

Mockingjay (The Hunger Games, #3)
by Suzanne Collins
progress:

0 minutes ago · like · comment · see status

The use of opinions from one's social network is facilitated on the Goodreads.com platform as the site prominently displays reviews provided by one’s friends and sources on the consumer’s home page in a “Recent Updates” list (Figure 6). In addition, reviews from one's social network are shown first when a user navigates to the *reviews* section for a specific book (Figure 7). Given the countless reviews available for certain books, we assume that this display format will prioritize social network reviews. Previous research has explored how consumers consider and react to a large supply of information in an online search context. For example, Morahan-Martin (2004) demonstrates that individuals tend to stay on the first page of search results when seeking information and rarely go on to subsequent pages. Granka et al. (2004) confirm this behavior in an eye-tracking study and find that people spend exponentially less time on lower-ranked search terms. We expect similar behavior on the Goodreads.com platform, where consumers are likely to focus on the ratings from their social network since they are shown before other ratings from "strangers."

Figure 7. Social Network Reviews



Mockingjay (The Hunger Games #3)
by Suzanne Collins
★★★★☆ 4.09 · rating details · 289,858 ratings · 38,790 reviews

Katniss Everdeen, girl on fire, has survived, even though her home has been destroyed. Gale has escaped. Katniss's family is safe. Peeta has been captured by the Capitol. District 13 really does exist. There are rebels. There are new leaders. A revolution is unfolding.

It is by design that Katniss was rescued from the arena in the cruel and haunting Quarter Quel...more

Hardcover, 390 pages
Published August 24th 2010 by Scholastic Press (first published 2010)
more details... edit details

Friend Reviews (1) 4.00 average

Yoshi rated it ★★★★☆ Mar 13, 2012

The final sample of 5,389 consumers had established a total of 13,294 friends, 354 followers, and 960 sources at the end of the 50-day window[12]. The sparse social network is consistent with other online activity, where a large majority of users do not frequently participate (Nielson 2006). The users in our sample were responsible for the production of 120,843 ratings across 16,595 books during the 50 days. We also collect every review written for each book in this set of 16,595 on the Goodreads.com platform for a total of 47.7 million reviews. This

[12] We note that this is different than from Table 1 because these individuals are a random sample whereas all individuals in Table 1 are already a friend or a source and more prone to making social ties.

additional data allows us to control for the prevailing $VALENCE_{ibt}$ and $VARIANCE_{ibt}$ when individual *i* posts her review for book *b* at time *t*. We define $VALENCE_{ibt}$ as the average and $VARIANCE_{ibt}$ as the variance of the rating for book *b* at time *t* before individual *i* posts her rating. Since we are able to capture the order of posted ratings across books, $VALENCE_{ibt}$ and $VARIANCE_{ibt}$ are calculated from all posted ratings prior to individual *i 's* rating for book *b*.

Recall that our theory is one of learning dynamics: individuals learn how to make use of the information gathered via their network ties for purchase decisions. We therefore need a measure to capture a user's evolving level of experience interacting with others and making decisions in the social network. As one would expect, some members of the sample over the 50 days of data collection used the platform a great deal while others did not. To represent this important construct, we use the number of reviewing sessions that a consumer has experienced prior to time *t*[13]. Specifically, our dynamic experience variable is given by $EXPERIENCE_{it}$, which, consistent with Moe and Schweidel (2012), we define as the number of "reviewing sessions" prior to time *t*. Each reviewing session consists of a day where individual *i* posts one or more reviews. Therefore, the first reviewing session occurs on the day that *i* posts her first review. This differs from the number of reviews written prior to time *t* in that a reviewing session may include multiple reviews. We prefer the former to the latter as our goal is to capture learning effects. It is unlikely that the second of two reviews written in the same reviewing session would reflect learning that occurred since the first review[14]. We use this proxy because we expect that learning (i.e. how to improve decision quality) occurs only if the consumer uses online opinions to make purchase decisions and evaluates those decisions. Simply reading opinions or forming

[13] We also test alternative proxies for experience in Section 6.1. However, we expect that the number of reviewing sessions to be a more suitable measure because it captures actual usage experience.

[14] Moreover, we only observe daily timestamps for a review, precluding our ability to accurately order reviews that arrived on the same day.

network ties may not result in equivalent learning since these activities lack the feedback associated with a decision. Of course, we do not directly observe the use of opinions or actual purchase decisions. Therefore we rely on ratings observations to infer that an individual made a purchase and assessed the quality of that decision.

Our primary dependent variable -- decision quality -- is operationalized as the rating given by individual *i* to book *b* at time *t*. Thus, we assume that better purchase decisions will lead, all else equal, to higher ratings. In order to ensure that all else is, indeed, equal, we control for individual- and book-level factors that may also drive ratings behavior independent of decision quality. To control for observable book level dynamics, we create two variables: $TIME_{ibt}$ and $ORDER_{ibt}$. Godes and Silva (2012) find that while ratings decrease over the review order, they increase over time when controlling for calendar year. We note that $ORDER_{ibt}$ is equivalent to the number of previously posted reviews for book *b* at time *t*. In addition, we need not control for calendar year because our set of posted ratings all fall under one calendar year. Again, these measures are all collected at the time of posting for all posted ratings in our dataset.

For social network measures, we define $LAGFRIENDS_{it}$, $LAGFOLLOWERS_{it}$ and $LAGSOURCES_{it}$ as the lagged number of friends, followers, and sources, respectively, for individual *i* posting at time *t*. Since learning effects from friends and sources may take some time to develop -- one needs to take time to buy the book, make an evaluation, and write a review --, we lag these values by two weeks. In contrast, we only lag the number of followers by one day since we expect that a poster evaluates who is watching immediately before posting a rating. Thus, the rating at time *t* is a function of decisions made two weeks earlier, along with more immediate reputational effect. Higher levels of *LAGFRIENDS* and *LAGSOURCES* imply a broader set of bi- and uni-directional ties and thus more strong and weak ties, respectively.

Theoretically, these ties provide information to the consumer. We see these measures as representing proxies for the amount of information gathered from these two different information sources.

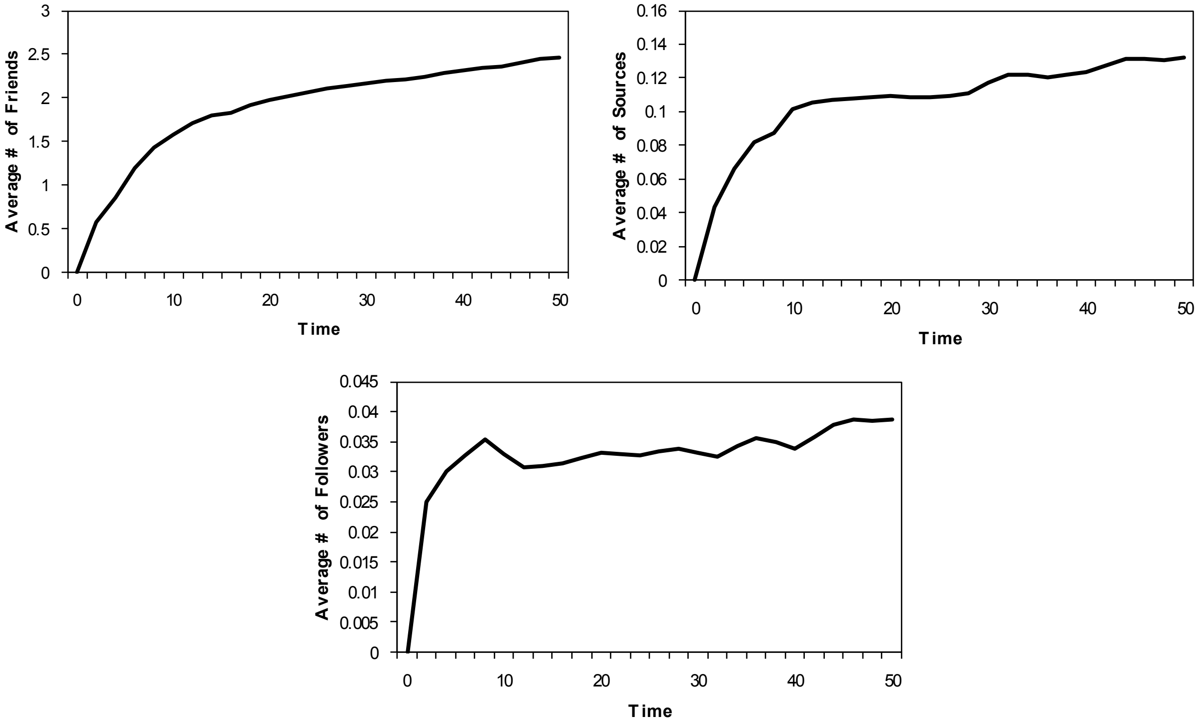
We note that, by virtue of the unique characteristics of our dataset, we are able to investigate directly the impact of reputational concerns on rating behavior. Indeed, while not a focus of our study, we would suggest that our ability to control for these factors is novel in the literature on online ratings. Concerns for reputation may result, for example, in a negative adjustment in ratings for those who are concerned about self-image (Schlosser 2005; Moe and Schweidel 2012). Researchers have captured adjustment effects in a lab setting (Schlosser 2005) or in the heterogeneity in the evaluation adjustment due to the ratings environment (Moe and Schweidel 2012). Other researchers have also suggested incidence effects in the posting of reviews (Ying et al. 2006; Moe and Schweidel 2012; Wojnicki and Godes 2011), potentially selecting for more-positive ratings. Our dataset allows us to control for reputation precisely by capturing in *LAGFOLLOWER*, a proxy for the number of others who are reading the reviews of consumer *i*. In contrast, most existing research has used aspects of the review environment (such as valence, variance, and volume) to proxy for reputation concerns. Without a measure of the number of others "watching," it may be difficult to disentangle more-substantive impacts of the environment (for example, an anchoring effect of the prevailing mean or uncertainty induced by the prevailing variance) from true reputational concerns. Since we have separate measures for the environment and the number of followers, we are able to disentangle and control for these effects.

Table 11 provides summary statistics for our dataset. The mean and standard deviation are calculated across all observations, including those in the beginning of our observation period where most individuals have not made any social network ties. Figure 8 shows, for each type of relationship, the evolution of the social network over time. We note that the latter highlights an important challenge associated with our dataset (and, likely, with all large social networks): the sparseness of network relationships. Many individuals choose to have very few, or no, social ties. This may occur for those who have yet to start making social ties. Alternatively, many individuals do not actively participate in contributing online information (Nielson 2006), thus potentially reducing their need to make social connections. This presents some estimation challenges in terms of our ability to control for unobserved heterogeneity. We discuss the implications of this in more detail below.

Table 11. Descriptive Statistics and Correlations

Variable	Mean	Std. Dev.	Min	Max	1	2	3	4	5	6	7	8	9
1. Rating	3.87	1.09	1.00	5.00	1.00								
2. VALENCE	3.89	0.29	1.00	5.00	0.10	1.00							
3. VARIANCE	1.00	0.23	0.00	4.00	-0.08	-0.55	1.00						
4. EXPERIENCE	1.68	1.98	1.00	33.00	0.01	-0.04	-0.12	1.00					
5. TIME (days/100)	34.11	30.98	0.00	348.07	0.00	0.05	-0.04	0.03	1.00				
6. ORDER (/1000)	8.55	10.47	0.00	38.32	0.01	0.23	0.18	-0.21	-0.07	1.00			
7. LAGFRIENDS	0.37	2.95	0.00	369.00	0.02	0.00	-0.04	0.26	0.00	-0.07	1.00		
8. LAGFOLLOWERS	0.01	0.23	0.00	20.00	0.01	0.01	-0.01	0.10	0.00	-0.01	0.27	1.00	
9. LAGSOURCES	0.08	0.64	0.00	39.00	-0.01	-0.01	-0.04	0.24	0.00	-0.06	0.21	0.03	1.00

Figure 8. Social Network Evolution



Model

Our main goal is to explore the impact of information gathered from an evolving social network on decision quality. In doing so, we use a consumer's posted rating as a proxy for decision quality under the assumption that, all else equal, higher reported ratings follow from better purchase decisions, on average. The use of rating as a proxy for decision quality is consistent with previous work (Ying et al. 2006; Li and Hitt 2008; Moe and Schweidel 2012; Godes and Silva 2012). Of course, past studies have shown that online reviews may be impacted by other factors for which we must control (Li and Hitt 2008; Moe and Schweidel 2012; Godes and Silva 2012). Therefore, we model ratings as a function of the ratings environment, individual level effects and book level effects described in the Data Section.

Our dependent variable, the posted rating *EVAL$_{ibt}$*, is a discrete measure from one to five stars, where five represent the highest possible rating. This discrete ordered variable is appropriately modeled using the ordered probit model, where we specify individual $i's$ continuous latent evaluation for book *b* at day *t*, as:

$$\begin{aligned} EVAL_{ibt}^{*} = \beta_{0i} + \beta_{1i}VALENCE_{ibt} + \beta_{2i}VARIANCE_{ibt} + \beta_{3i}EXPERIENCE_{it} + \gamma_{0b} \\ + \gamma_{1b}TIME_{ibt} + \gamma_{2b}ORDER_{ibt} + \delta_{1}LAGFRIENDS_{it} + \delta_{2}LAGFOLLOWERS_{it} \\ + \delta_{3}LAGSOURCES_{it} + \delta_{4}EXPERIENCE_{it}LAGFRIENDS_{it} \\ + \delta_{5}EXPERIENCE_{it}LAGFOLLOWERS_{it} + \delta_{6}EXPERIECE_{it}LAGSOURCES_{it} \\ + \epsilon_{ibt} \end{aligned}$$

where $\epsilon_{ibt} \sim N(0,1)$. $EVAL_{ibt}^{*}$ is the latent evaluation that will serve as a proxy for decision quality after controlling for the individual-level intercepts (β_{0i}), ratings environment effects (*VALENCE*,*VARIANCE*), individual-level experience effects (*EXPERIENCE*), book-level intercepts (γ_{0b}), book-level dynamic effects (*TIME*, *ORDER*), and reputation (*LAGFOLLOWERS*). The individual-level intercepts, β_{0i}, capture unobserved heterogeneity across individuals including those unobservable characteristics driving an individual's tendency to generate more-positive or more-negative reviews. This parameter thus allows us to control for a consumer's "ability" to select a book that matches her preferences, conditional on the vertical (average) "quality" of the book. Some people tend to make better purchasing decisions while others may be inherently bad at selecting books that will deliver high utility, potentially causing variation in the ratings evaluation. This also captures, of course, any inherent positive or negative bias a consumer may have in her rating tendency. The book-level intercepts, γ_{0b}, capture the baseline quality of the book and other unobservable book-level time invariant factors.

To test whether consumers may improve their decisions due to access to more information mainly available through the creation of new relationships, we include the two social network variables, *LAGFRIENDS* and *LAGSOURCES*. We assume that friends are a relationship between stronger ties due to the bi-directional link and more frequent interaction that tend to occur between friends. In contrast, sources are weaker relationships than friends. We expect that both the interpretability and the novelty -- and thus the usefulness -- of information varies across these different types of ties. We also include *LAGFOLLOWERS* to control for reputational effects. We note that *LAGFOLLOWERS* may also help control for self-selection in posting behavior. According to extant research, individuals may only post if they expect their review to have an impact (Wu and Huberman 2008; Godes and Silva 2012). The number of followers provides a measure -- a proxy for the number of people who are exposed to one's reviews -- to control for one's concerns about the impact of a potential review.

Our focus here is on one's ability to learn to use information gathered from the evolving network, which is captured in the dynamic interactions between the social network variables and *EXPERIENCE*. That is, we test for the impact of the evolving social network on ratings as the consumer gains more experience. Our analysis focuses on the interactions of the variables between *LAGFRIENDS* and *LAGSOURCES*, on one hand, and *EXPERIENCE*, on the other[15]. These coefficients capture the change in the marginal impact on decision quality of adding new ties as one gain more experience. It is also possible that experience itself would improve decision quality. It is known, for example, that consumers may become more skilled in processing relevant information (Alba and Hutchinson 1987), which may help them make better choices. This effect, for non-social information, will be captured in the main effect for *EXPERIENCE*.

[15] We mean center all interaction variables for both the interaction and main effect terms.

Endogeneity and Identification

Consider a simplified version of our model, assessing the impact of the number of friends on ones' decision quality[16]:

$$EVAL_{it}^{*} = \delta_0 + \delta_1 LAGFRIENDS_{it} + \epsilon_{it}$$

There are several potential threats to identification in this model. First, endogeneity problems may arise to the extent that the actions of agents within a network are driven by unobservable individual effects that influence both ratings and network decision. For example, consider one's ability to make a good choice. It may be the case that those who have high ability to select good books also tend to create more friendship ties. Similarly, people who have similar abilities may tend to form ties with each other. If unaccounted for, this would cause $LAGFRIENDS_{it}$ to be correlated with ϵ_{it}, calling into question the estimates of δ. A second source of potential endogeneity would come from unobservable correlated time-varying shocks to both the posted ratings and friendship formation. As a slightly-exaggerated example, if individual *i* is having a "good day" and is very satisfied in general, she may post higher make more new friends. Again, this causes the number of friends to be correlated with the error term. Third, identification issues also occur if, in addition to the impact of networks on ratings (as captured in Equation (1)), it is also the case that posted ratings cause the formation of friendship ties. One's decision to form a bi-directional link with individual *i* may be a function of her posted ratings evaluation, for example. This raises the possibility of reverse causality, again resulting in the number of friends being correlated with ϵ_{it}. Indeed Shriver et al. (2013) find that social network activities (i.e. blogging) influence the number of ties one makes[17]. Even the use of a lagged social network

[16] Please refer to Hartmann et al. (2008) and Nair et al. (2010) for a general treatment of identification issues in a setting characterized by social interactions.

[17] Toubia and Stephen (2013) also suggest that the number of followers may be endogenous with content creation.

variable, as we do here, does not adequately deal with reverse causality if there is serial correlation in the latent evaluation errors.

We address these issues as follows. With respect to time-invariant, correlated unobservables such as ability or taste, we specify our evaluation model to contain individual- and book-level random effects to control for individual-level and book-level differences (Hartmann et al. 2008; Nair et al. 2010). Thus, the estimates of δ should be seen as conditional on an individual's skill at selecting books as captured in the intercept. Neither time-varying correlated shocks to the ratings nor the potential for reverse causality are controlled for using random effects. To resolve these concerns, we employ two approaches: instrumental variables (IV) and latent instrumental variables (LIV).

In the IV approach, we employ instruments that are correlated with *LAGFRIENDS*$_{it}$ but uncorrelated with ϵ_{it}, allowing us to estimate δ consistently. As instruments for *LAGFRIENDS*$_{it}$ we use the size of individual *i's* friends' networks. This is analogous Shriver et al. (2013) who use the number of friend request of *i's* friends as an instrument for the number of *i's* friend requests). Precisely, let $\Theta_{\mathrm{ut}} = \{$set of friends of individual *u* at time $t\}$. Then the instrument for *LAGFRIENDS*$_{it}$ is: $Z_{it} = \sum_{u\in\Theta_{\mathrm{it}}}(N[\Theta_{\mathrm{uT}}] - 1)$, where we define *N*[*S*] as the cardinality of set *S* (i.e. the number of individuals who are friends with *u*). Thus, for each individual *i*, we identify all individuals *u* who are friend of *i* at time *t*, and collect the total number of friends that each *u* will have by the end of the observation window[18]. Again, our use of this instrument is in keeping with the goal of staying as close as possible to the existing literature (Shriver et al., 2013; Oestreicher-Singer and Sundararajan, 2012). We have also estimated a model using as an instrument the average number of friends *i's* friends have: $Z_{it} = \frac{1}{N[\Theta_{\mathrm{iT}}]}\sum_{u\in\Theta_{\mathrm{it}}}(N[\Theta_{\mathrm{uT}}] - 1)$. The

[18] Note that this data collection was performed after the end of the observation period and we are therefore unable to determine the timing of network formation for *i's* friends.

results of this analysis are qualitatively equivalent to those presented here and are available from the authors.

The validity of our instrument is dependent on our assumptions that (a) the size of one's network is correlated with the size of one's friends' networks but that (b) one's choice of ratings has no effect on the size of her friends' networks. In our research setting, the latter assumption seems reasonable given that only ratings from first degree ties (i.e. immediate friends and followers) are easily accessible on the book review page whereas opinions from second degree ties are “hidden” with the other mass reviews (each book in our sample has, on average, 3,000 reviews). The tie formation between *i's* friends (e.g. j) with others who are not connected to *i* (e.g. *k*) involve a joint decision by both *j* and *k*. We expect that individual *k* (and others not connected to *i*) will not be directly affected by *i's* ratings because those ratings are not readily observable. Even if they are observable, we would expect that *i's* ratings should affect a potential tie between *i* and *k* and not between other people (e.g. *j* and *k*). In a similar sense, we expect that the network size of *i's* friends do not directly impact *i's* ratings because *i* cannot easily observe those tie formations, alleviating concerns for potential reputation effects due to a larger group of second-degree ties. We obtain analogous measures for one's followers and sources to create instruments for *LAGFOLLOWERS* and *LAGSOURCES* (the number of ties for ones followers is calculated at time t-1 since *LAGFOLLOWERS* only has a one day lag). These instruments are positively correlated with the total number of network ties (ρ= 0.79, 0.57, 0.63 for friends, followers, and sources respectively), as a greater number of ties is related to the network structure of those ties.

Of course, the consistency of our IV estimates depends on the validity of our instruments. If, however, the network structure of *i's* friends, followers, or sources is directly related to *i's* ratings evaluation, then the IV assumptions may be compromised and our estimates potentially inconsistent. Therefore, as a second approach, we also estimate our model using the LIV method (Ebbes 2004; Ebbes et al. 2005; Zhang et al. 2009; Rutz et al. 2012). This modeling approach deals with endogeneity without relying on the validity of explicit, observable instruments. Following Ebbes (2004), Ebbes et al. (2005), Zhang et al. (2009), and Rutz et al. (2012), we utilize data augmentation (Tanner and Wong 1987) to estimate a binary, latent instrumental variable to decompose the potentially-endogenous network variables into two components, one that is uncorrelated and another that is correlated with ϵ_{it}. Again, to simplify notation, we discuss the simplified model with one endogenous variable as follows:

$$EVAL_{it}^{*} = \delta_0 + \delta_1 LAGFRIENDS_{it} + \epsilon_{it}$$
$$LAGFRIENDS_{it} = \mathrm{\theta Z_{it}} + \mathrm{\nu_{it}}$$

where Z_{it} is a latent categorical variable with category means of θ. We assume that Z_{it} has two categories[19], is orthogonal to $\mathrm{\nu_{it}}$ and ϵ_{it}, and follows a binomial distribution with probabilities (π_1, π_2). Here, π_c is the probability that the c$^{\text{th}}$ latent instrument is one (i.e. belongs to category c) and $\sum_c \pi_c = 1$. Furthermore, we assume that the error terms follow a multivariate normal distribution with mean 0 and variance-covariance matrix:

$$\Xi = \begin{bmatrix} \sigma_{\epsilon\epsilon} & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma_{\nu\nu} \end{bmatrix}$$

Thus, the likelihood function is specified as:

$$p(k_{it}|\delta, \theta, \Xi) = (2\pi)^{-1}|\Xi|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(k_{it} - u_{it})'\Xi^{-1}(k_{it} - u_{it})\right]$$

[19] The number of categories needs to be equal or greater than two for identification purposes; however, the model is robust against misspecification of the number of categories (Ebbes 2004). We follow Ebbes (2004), Ebbes et al. (2005), and Rutz et al. (2012) and set the number of categories to be two. For more details on identifiability of the LIV, please refer to Ebbes (2004) and Ebbes et al. (2005).

where $k_{it} = (EVAL_{it}^{*}, LAGFRIENDS_{it})'$ and $u_{it} = (\delta_0 + \delta_1 LAGFRIENDS_{it}, \theta Z_{it})'$. Please refer to the Appendix for more details on the LIV approach.

Estimation

Since the posted ratings, $EVAL_{ibt}^{*}$, is ordinal, we use data augmentation to map the ordered ratings onto a continuous scale (Tanner and Wong 1987). To transform the evaluation from a five-point scale, we model the probability of a given one to five-star rating as:

$$
\begin{gathered}
p(EVAL_{ibt} = 1) = \Phi(\tau_{1i} - EVAL_{ibt}^{*}), \\
p(EVAL_{ibt} = 2) = \Phi(\tau_{2i} - EVAL_{ibt}^{*}) - \Phi(\tau_{1i} - EVAL_{ibt}^{*}), \\
p(EVAL_{ibt} = 3) = \Phi(\tau_{3i} - EVAL_{ibt}^{*}) - \Phi(\tau_{2i} - EVAL_{ibt}^{*}), \\
p(EVAL_{ibt} = 4) = \Phi(\tau_{4i} - EVAL_{ibt}^{*}) - \Phi(\tau_{3i} - EVAL_{ibt}^{*}), \\
p(EVAL_{ibt} = 5) = 1 - \Phi(\tau_{4i} - EVAL_{ibt}^{*}).
\end{gathered}
$$

where $EVAL_{ibt}^{*}$ is the latent evaluation of consumer *i* for book *b* at time *t*. We specify τ_{1i}, τ_{2i}, τ_{3i}, and τ_{4i} as unobserved individual-level cutoff points, which separate the latent utilities into discrete segments over the normal distribution. These individual-level cutoffs allow us to account for individual differences in the rating process such as scale usage heterogeneity. Some consumers may only rate a “5” for extremely-high utility values whereas others may use a “5” more liberally. To allow for heterogeneity across individuals in the cutoffs, we follow Ying et al. (2006) and assume that the log of the differences between adjacent cut offs to have normal distribution: $\log(\Delta\tau_i) \sim MVN(\bar{\tau}, \Gamma)$. We also assume that $\bar{\tau}$ follows a diffuse normal prior and Γ follows an inverse Wishart prior. As is standard (Albert and Chib 1993; Ying et al. 2006; Moe and Schweidel 2012), we identify the three cutoff points and one intercept per individual by setting $\tau_{1i} = 0$.

To allow for heterogeneity across individuals and correlation among the individual-level parameters, we assume that $\beta_i \sim N(\bar{\beta}, \Sigma)$. The β_i parameters include the individual-level intercept from Equation (1). We also allow for heterogeneity across books and correlation among the book-level parameters and assume that $\gamma_b \sim N(\bar{\gamma}, \Omega)$. The γ_b include the book-level intercepts from Equation (1). The identification of the two separate intercepts -- individual and book -- also presents a challenge. To appreciate this, note that one could add a constant to each individual-level intercept and subtract the same constant from each book-level intercept without changing the model. To solve this identification problem, we set the first book level intercept to 0. This allows us to identify the remaining cut off points and intercepts.

For the rest of the hierarchical model, we assume diffuse normal priors for the mean effects of the individual-level and book-level parameters in the model (i.e. $\bar{\beta}$ and $\bar{\gamma}$). We also assume that Σ and Ω follow Inverse Wishart priors. Furthermore, for the aggregate level estimates on the social network variables (δ), we assume a diffuse normal prior: $\delta \sim N(\delta_0, \Psi_0)$.

It is important to note that a limitation of our social network data is that there exists little variation within individual over time for the social network variables. While there is sufficient individual-level variation within the ratings environment and user characteristics, the social network variables do not have adequate variation across time and within individual to identify individual level estimates. As a result, we estimate δ at the aggregate level[20]. The proposed model requires us to estimate individual-level, book-level, and aggregate-level estimates. In order to do so, we cycle through the Gibbs sampler (Gelfand and Smith 1990) and use a block Gibbs sampler to iteratively estimate each set (individual-, book-, and aggregate-level) of parameters. The full conditional posteriors are available from the authors.

We run 30,000 iterations, where the first 20,000 serves as the burn-in period and remaining 10,000 iterations are used to obtain inferences about the posterior. The iterations

[20] We also estimate a model with latent segment level δ_s and find similar results (available from authors).

quickly converge to a stable posterior, as assessed by examining the posterior iterations and the Gelman and Rubin diagnostic (Brooks and Gelman 1998).

Results

Table 12 presents our main results. Model 1 contains our base model while Models 2 and 3 reflect our corrected estimates obtained via IV and LIV, respectively. Importantly, the three models demonstrate a high degree of consistency in qualitative results. Notably, the IV and LIV results are very similar, providing confidence that the two methods -- while based on different assumptions -- are properly controlling for potential endogeneity. Moreover, note that these models yield stronger results on the social variables of interest than does Model (1), suggesting again that the instruments have "bite." We first focus on the results for the social network variables, which show significant static and dynamic findings. First, as the number of friends increases, consumers post higher ratings about their past purchases, which we interpret as higher experienced utility. This suggests that one's decision quality increases as she develops more bi-directional relationships. Second, as the consumer follows more sources, her experienced utility declines. This provides evidence that not all information from one's social network is equally useful in improving one's decisions. The negative coefficient on LAGSOURCES, in particular, suggests that information from these uni-directional relationships may actually lower a consumer's decision quality.

These findings demonstrate the importance of accounting explicitly for differences across tie types in terms of not only the novelty of the information received from ties (Granovetter 1973; Brown and Reingen 1987) but also the interpretability of that information. While the literature is clear in suggesting that information from weaker ties -- i.e. uni-directional links -- would be more valuable in driving decision quality, this view may ignore the fact that

Table 12. Main Results

Model	(1)	(2)	(3)
Method	No Instrument	IV	LIV
VALENCE	**0.699****	**1.075****	**1.088****
	(0.201)	(0.353)	(0.349)
VARIANCE	**-0.397****	**-0.393***	**-0.378***
	(0.109)	(0.158)	(0.160)
EXPERIENCE	0.337	-0.073	-0.072
	(0.186)	(0.111)	(0.109)
TIME	**0.002***	0.000	0.000
	(0.001)	(0.000)	(0.000)
ORDER	-0.012	0.000	0.000
	(0.011)	(0.000)	(0.000)
LAGFRIENDS	**0.027***	**0.004***	**0.007***
	(0.013)	(0.002)	(0.003)
LAGFOLLOWERS	0.125	-0.001	0.033
	(0.106)	(0.017)	(0.025)
LAGSOURCES	**-0.098^**	**-0.017***	**-0.040***
	(0.067)	(0.008)	(0.019)
EXPERIENCE*LAGFRIENDS	-0.009	**-0.0011***	**-0.004***
	(0.009)	(0.0005)	(0.002)
EXPERIENCE*LAGFOLLOWERS	**-0.110***	**-0.005***	**-0.029***
	(0.042)	(0.003)	(0.013)
EXPERIENCE*LAGSOURCES	**0.051***	**0.002***	**0.015***
	(0.024)	(0.001)	(0.006)
Number of Individuals	5,389	5,389	5,389
Number of Books	16,595	16,595	16,595
MAD	0.779	0.771	0.771
RMSE	1.263	1.267	1.250

Notes. The posterior standard errors are reported in the parentheses

^ Indicates that 0 is not contained 90% Bayesian Credible Interval

* Indicates that 0 is not contained 95% Bayesian Credible Interval

** Indicates that 0 is not contained 99% Bayesian Credible Interval

information acquired from strong ties may be more-immediately useful because friends may have a higher likelihood of interaction, a greater sense of trust and understanding as well as mutual confiding (Granovetter 1973; Lin et al. 1981; Brown and Reingen 1987). As such, a consumer may be initially better at interpreting and applying information from friends to a specific decision as compared with using information from more-anonymous or less-familiar acquaintances, via a one-way connection to their sources. This latter relationship is likely to be characterized by less interaction and, thus, a greater difficulty in mapping the information they provide to one's preferences.

These results conflict with both the traditional view of weak ties and our observed empirical pattern of (online) tie formation: people build, and acquire information from, many weak ties. It would be surprising were it the case that these ties were not in some way beneficial. Thus, we next consider the effects of the consumer's dynamic learning process via a set of interaction terms between experience and the network tie measures. We interpret these estimates as capturing the marginal change in the benefit of adding a new network tie as one gain more experience. We expect, in particular, that, as consumers develop greater experience with the platform and in their network, they become better at interpreting information from weak ties, and, as a result, information gathered from weak ties should begin to yield benefits for decision making.

Indeed, we find that the coefficient on *EXPERIENCE X LAGSOURCES* is positive and significant, suggesting that as one gains more experience in the community, developing more weak-tie relationships improves decision quality. Combined, our results thus suggest that, for relatively-inexperienced consumers, developing more weak-tie relationships leads to worse decisions (reflected through the negative estimate for *LAGSOURCES*). Yet, over time, as one gains experience in the community, these relationships ultimately lead to better decisions. This provides evidence of a rich dynamic consumer learning process.

The dynamics of information acquired from friends appears to be more complicated, as our model unexpectedly yields a negative coefficient on *EXPERIENCE X LAGFRIENDS*. We interpret this as potentially arising from a reputational concern that experienced users may exhibit. Unlike sources, friends are able to view one's ratings so there is the potential for one to adjust ratings to manage reputation. Consistent with this view, we find that the coefficient on *EXPERIENCE X LAGFOLLOWERS* is significantly negative, suggesting that as individuals become more experienced and gain more followers, they tend to become “more negative”[21]. Thus, the *EXPERIENCE X LAGFRIENDS* term may represent the net effect of an improvement due to learning and an adjustment due to reputational concerns. Unfortunately, we are unable to directly disentangle the two effects in Models 2 and 3. In Section 6.1 below, we address this issue by studying the impact of sources that eventually evolve into friends. This approach allows us to compare when there are no reputational effects (sources cannot see one's ratings) versus when reputation effects may exist (friends can see one's ratings) for the same individuals. Note that our estimates of the effect of sources are not similarly impacted by these potential reputational effects: while I see my sources' ratings, they don't see mine. Thus, our results for sources seem to provide evidence of decision-quality improvement due to learning effects[22].

One possible alternative explanation for these results may be that members of the community are not only experiencing different levels of utility from their purchases but that they instead (or also) provide more ratings and/or purchase more books. Another explanation may be that the increasing embeddedness in a community leads one to experience (or, perhaps, to simply communicate) higher satisfaction with past purchases. While both of these mechanisms seem plausible and may in fact be at work in our data, it is unlikely that they would give rise to the specific pattern of results we observe. With respect to the rating-selection process, while the

[21] We acknowledge that this finding should be interpreted with caution as only 4% of all individuals have any number of followers.

[22] We also test the robustness of our findings in order to rule out the possibility that our dynamic interactions are picking up effects that are simply non-linearities. That is, if the social network variables impact one's latent decision quality in a non-linear fashion, the interaction of *EXPERIENCE* with the variable may be a proxy for the non-linear effect. To check this, we re-estimate Model 3 and include quadratic terms ($FRIENDS^2$, $FOLLOWERS^2$, and $SOURCES^2$; available from the authors). However none of the estimates for the quadratic terms are significant. This suggests that these variables are not proxies for hidden non-linearities.

development of network ties may, indeed, impact one's willingness to provide a review for an underlying experience, it is not at all clear why such an effect would be non-monotonic as we observe for both weak and strong ties[23]. Similarly, while one may be increasingly positive as one becomes embedded in a community, we would not expect this effect to be non-monotonic. More important, we would not expect weak ties to have a larger effect on one's happiness than would strong ties, as we find here.

As for our controls, we find that *VALENCE* is significantly positive and *VARIANCE* is associated with lower posted ratings. These estimates are consistent with previous research (Chevalier and Mayzlin 2006; Dellarocas et al. 2007; Chintagunta et al. 2010; Moe and Trusov 2011). Our book level estimates are also consistent with Godes and Silva (2012), showing that posted ratings increase with *TIME* after controlling for *ORDER*[24]. While, unlike Godes and Silva (2012), we do not explicitly control for calendar year, our data were gathered over a two month period. Therefore, all posted ratings lie within the same calendar year[25].

Extension 1: Sources who turn into friends

Our main results suggest that, when first joining a community, consumers make better (worse) decisions as they add more friends (sources). As they gain experience, however, adding more sources eventually improves their decision quality. We also find that, as experience evolves, adding more friends (strong ties) leads to lower ratings. We argue in the previous

[23] In order to check whether there exists any evidence that members post more reviews as a function of our focal variables of interest, we estimated a Probit model in which the dependent variable was the (daily) decision to post a rating. There appears to be some evidence that individuals with more followers are more likely to post a rating. However, critically, there is no evidence that such a process would explain our main results in that there seems to be no impact of friends or sources or their experience interactions. These results are available from the authors.

[24] We also find that posted ratings decrease with *ORDER* if we remove *VARIANCE*. This is consistent with Godes and Silva (2012) in that larger values of *ORDER* means greater dissimilarity between individuals, which may be captured through *VARIANCE*.

[25] We also run Model 4 and include a *VALENCE*VARIANCE* interaction. The coefficient on this variable is significantly positive, consistent with Sun (2012), and all other estimates remain the same.

section that this may be due to the offsetting impact of social factors rather than an impact of network ties on decision quality. Since friends are bi-directional relationships, I both benefit from the information they provide and am concerned with how my ratings make me look to them. The former should increase my ratings while the latter, according to the literature (Schlosser 2005; Moe and Schweidel 2012), might cause me to shade down my ratings. Moreover, one would expect that concerns for reputation in a community would be enhanced as one becomes more embedded in that community. Thus, we suggest that the negative coefficient on *EXPERIENCE X LAGFRIENDS* may be the net effect of these two forces.

In order to investigate this in more depth and to assess the extent to which friends, and not sources, would bring about reputation concerns, we attempt to decompose the effects of information and reputation by isolating those ties that migrate within our 50-day window from a source to a friend. In our dataset, 14% of those classified as a source eventually become a friend[26]. We create a distinct category of ties and analyze their impact on ratings both before the source-to-friend transition and after, which we label as *PRE-S2F* and *POST-S2F*, respectively (these variables are also lagged). Note that the key impact of the tie transition is that the relationship goes from one in which the focal user was previously reading the source's reviews to one in which the source becomes a friend who is also reading the focal user's reviews. Thus, critically, the information available to the focal user does not change and the primary impact of this transition should be driven by concerns for reputation. While this represents, using the labels we've employed in the paper, a migration from weak ties (sources) to strong ties (friends), it is important to note that we should not expect a significant change in the value of the information provided from the tie following the migration. The novelty benefits associated with weak ties

[26] These changes are migrations within a specific dyadic tie: those cases in which A was B's source and eventually became B's friend.

derive from the fact that one's strong ties are likely to be friends with one's other strong ties (thus providing less-novel information) while this is not true of weak ties (Granovetter 1973; Brown and Reingen 1987; Levin and Cross 2004). There would be no reason to expect such a change in the short term following a source-to-friend migration in our dataset. Thus, the primary effect, again, should be attributable solely to the different concerns for reputation one faces as a source becomes a friend.

Table 13 shows the results of this model with the inclusion of the source-to-friend migration variables. Specifically, *LAGFRIENDS* and *LAGSOURCES* represent now those ties who remain friends and sources, respectively, throughout the data window while *PRE-S2F* and *POST-S2F* represent ties that began as sources and became friends. The former (latter) captures the impact of these ties before (after) the migration occurs. Our primary test of reputation is in the comparison of the coefficients on *EXPERIENCE*PRE-S2F* and *EXPERIENCE*POST-S2F*. We again highlight that these variables capture effects from the same people. However, once the source becomes a friend, she is now able to see the focal user's ratings and reviews, creating potential concerns for reputation. Importantly, and consistent with our results throughout the study, these effects are expected to be most salient when the member has higher levels of experience in the community. It is then that one would be expected to care more about reputation. As shown in the table, in both the LIV and IV models, the mean coefficients on *EXPERIENCE*PRE-S2F* are higher than those on *EXPERIENCE*POST-S2F*: while the information I acquire from a source does not degrade when she becomes my friend, I nonetheless appear to adjust my ratings downward to reflect concerns for my reputation. This is consistent with previous literature (Schlosser 2005; Moe and Schweidel 2012).

Table 13. Sources turn into friends

Description	(4) Source turned into Friend	(5) Source turned into Friend
Method	IV	LIV
VALENCE	1.106*	1.091**
	(0.355)	(0.356)
VARIANCE	-0.370*	-0.375*
	(0.161)	(0.147)
EXPERIENCE	-0.078	-0.092
	(0.112)	(0.106)
TIME	0.000	0.000
	(0.000)	(0.000)
ORDER	0.000	-0.001*
	(0.000)	(0.000)
LAGFRIENDS	0.003*	0.004*
	(0.001)	(0.002)
POST-S2F	0.053*	0.018
	(0.028)	(0.026)
LAGFOLLOWERS	-0.023	0.015
	(0.019)	(0.027)
LAGSOURCES	-0.017*	-0.011*
	(0.008)	(0.006)
PRE-S2F	-0.034*	-0.031^
	(0.022)	(0.020)
EXPERIENCE*LAGFRIENDS	-0.014*	-0.004*
	(0.007)	(0.002)
EXPERIENCE*POST-S2F	0.000	0.008*
	(0.000)	(0.005)
EXPERIENCE*LAGFOLLOWERS	-0.008*	-0.026*
	(0.003)	(0.011)
EXPERIENCE*LAGSOURCES	0.004*	0.006*
	(0.002)	(0.003)
EXPERIENCE*PRE-S2F	0.009*	0.012*
	(0.006)	(0.007)
Number of Individuals	5,389	5,389
Number of Books	16,595	16,595
MAD	0.770	0.768
RMSE	1.247	1.250

Notes. The posterior standard errors are reported in the parentheses

^ Indicates that 0 is not contained 90% Bayesian Credible Interval

* Indicates that 0 is not contained 95% Bayesian Credible Interval

** Indicates that 0 is not contained 99% Bayesian Credible Interval

While not our primary motivation for the analysis, it is nonetheless interesting to note several other comparisons. First, the mean estimates on *EXPERIENCE*LAGFRIENDS* are lower than those on *EXPERIENCE*POST-S2F*. This represents a comparison of two groups of friends: those that began (and remained throughout the window) as friends and those that began as sources. This may either be because the negative reputational concerns driving the former are not quite as pronounced in the latter or the beneficial information delivered by the previous sources (since they are expected to provide more novelty) is higher. In addition, we note that the estimates on *EXPERIENCE*PRE-S2F* are positive and much larger than those for *EXPERIENCE*LAGSOURCES*. This is a comparison of two groups of sources: those that remain sources and those that eventually become friends. We attribute this to learning and interpretability: those who will eventually turn into friends may have evolved a stronger tie characterized by more learning interaction with the reader than other sources, resulting in better interpretability of opinions. Returning to our earlier theoretical arguments, these ties should give rise to the highest improvements in decision quality for experienced members. On one hand, they began as weak ties, or sources, delivering novel and highly-useful information. Over time, they evolved into a valued source of information which would ultimately lead to a strong-tie relation and, we might expect, higher levels of interpretability than average for sources. Moreover, as this variable captures the pre-migration period, it also reflects the period before any reputation concerns play a role. Indeed, for experienced users and as expected, in both the LIV and IV models, the *EXPERIENCE*PRE-S2F* is the highest among the experience interactions.

Extension 2: Experience

Our main results in Table 12 are derived using the number of rating sessions as a proxy for experience. Given the central role played by this construct, we discuss what this variable represents and explore other potential proxies. Our operationalization of experience represents the learning process through which individuals go when making actual decisions. Each day that an individual posts a rating, we assume that she is using the platform to make real decisions (i.e. ratings sessions reflect decision-making experiences). In so doing, she learns about the norms of the community and of other members and evaluates her decisions, which we would expect will further improve learning. Of course, this is an imperfect proxy for actual learning and, here, we consider other possible proxies.

To create an alternative proxy we collect all book posting activities that an individual engages in. On Goodreads, in addition to posting ratings, individuals can also post the books they have read or are planning to read (i.e. without a rating). Using these book posting observations, we create the number of posting sessions that an individual engages in. Specifically, this variable is 1 on the first posting day and increases by 1 on each successive day with one or more book-posting activities. While this variable also reflects real choices, we expect it to be somewhat less-useful as we assume that these additional activities require less engagement and, thus, may lead to less learning. We also consider other proxies for experience such as the time since the user joined Goodreads and a dummy for whether a user has made a social network tie. Time since joined is calculated as the time between individual *i's* Goodreads sign-up date and time *t*. We calculate the social network dummy as equal to 1 if the individual has made any type of tie at time *t* and equal to 0 otherwise. We expect that these latter proxies, while capturing some measure of experience with the site or with social interactions, may not accurately incorporate decision-making situations and thus the learning that comes with them. This detail is important because we expect that users are most likely to improve as a result of actual decision making.

Table 14 shows the results for the different experience proxies[27]. Model 6 contains no control for experience as either a main effect or interaction. Comparing it with models 6a, 6b, 6c and 6d, we see that controlling for experience yields a better model according to MAD and RMSE criteria[28]. Moreover, operationalizing experience as the number of reviewing or posting sessions yields a better model than when using time since joined or a dummy for social tie formation.

In Models 7a-7d (Table 15), we include the experience interactions using the various proxies under consideration. The interactions are important to consider due to their significance and the improved performance in both MAD and RMSE when comparing to the models without interactions (i.e. comparing Model 7a to 6a). Our results suggest that our findings on learning rely significantly on one's definition and operationalization of experience. When considering actual rating or posting sessions, our main results consistently hold (Models 7a and 7b). We expect that the rating and posting sessions closely represent learning experiences with real decision making because in order for one to post a rating or designate a book being "read," one most likely made a purchase. We find that Model 7a (ratings sessions as experience) performs best according to RMSE and Model 7b (posting sessions as experience) performs slightly better according to MAD. Importantly, the results in these two models are qualitatively equivalent. In

[27] We present the results using the LIV method. The IV method also shows similar results (available from the authors).

[28] To compare the models in Tables 14 and 15, we forecast ratings for a holdout sample, as the deviance information criterion may be problematic for mixture models such as that used in the LIV approach (Spiegelhalter et al. 2002). We use all observations that occur within the first five weeks as calibration and the remaining observations (two weeks) as the holdout sample. We generate a ratings forecast for each holdout observation using the posterior from our model estimates. This is done for each set of posterior estimate across 5,000 iterations. We use the forecast to calculate mean absolute deviation (MAD) and root mean squared error (RMSE) and present the average MAD and RMSE across each of the 5,000 iterations and across all observations.

contrast, when operationalizing experience as “time since joined Goodreads” or whether a social network tie was formed, we find the model does not perform as well. These results may stem from the underlying learning process captured by the different experience measures. The time since the user joined Goodreads and the social network dummy may not accurately measure learning (i.e. learning that improves decision quality) because they may not capture very well the heterogeneity in consumers' usage of, or experience with, the platform in making decisions.

Table 14. Experience

Description of Experience	(6) No Experience	(6a) # Reviewing Sessions	(6b) # total posting activity sessions	(6c) time since joined	(6d) 0/1 dummy if user has made a tie
Method	LIV	LIV	LIV	LIV	LIV
VALENCE	**0.959****	**1.095****	**1.083****	**1.084****	**1.003****
	(0.362)	(0.346)	(0.354)	(0.264)	(0.357)
VARIANCE	**-0.457****	**-0.343***	**-0.377***	**-0.291***	**-0.423****
	(0.151)	(0.160)	(0.163)	(0.162)	(0.148)
EXPERIENCE		-0.088	-0.100	-0.086	0.176
		(0.109)	(0.112)	(0.067)	(0.154)
TIME	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ORDER	**-0.001***	-0.001	**-0.001***	**-0.001***	**-0.001***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
LAGFRIENDS	**0.001***	**0.002***	**0.002***	**0.003***	**0.003***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)
LAGFOLLOWERS	**0.033***	-0.019	-0.017	-0.001	0.020
	(0.016)	(0.020)	(0.028)	(0.019)	(0.015)
LAGSOURCES	**0.022***	**-0.009***	**-0.010***	-0.003	**0.010***
	(0.009)	(0.005)	(0.006)	(0.004)	(0.005)
Number of Individuals	5,389	5,389	5,389	5,389	5,389
Number of Books	16,595	16,595	16,595	16,595	16,595
MAD	0.781	0.777	0.774	0.782	0.789
RMSE	1.271	1.266	1.263	1.287	1.273

Notes. The posterior standard errors are reported in the parentheses
^ Indicates that 0 is not contained 90% Bayesian Credible Interval
* Indicates that 0 is not contained 95% Bayesian Credible Interval
** Indicates that 0 is not contained 99% Bayesian Credible Interval

Table 15. Experience with interactions

Description of Experience	(7a) # Reviewing Sessions	(7b) # total posting activity sessions	(7c) time since joined	(7d) 0/1 dummy if user has made a tie
Method	LIV	LIV	LIV	LIV
VALENCE	**1.088****	**1.073****	**1.123****	**1.03****
	(0.349)	(0.347)	(0.342)	(0.369)
VARIANCE	**-0.378***	**-0.369***	**-0.312***	**-0.403***
	(0.160)	(0.159)	(0.148)	(0.148)
EXPERIENCE	-0.072	-0.084	-0.083	0.244
	(0.109)	(0.113)	(0.107)	(0.144)
TIME	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
ORDER	0.000	0.000	**-0.001***	**-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
LAGFRIENDS	**0.007***	**0.006***	**0.003***	-0.001
	(0.003)	(0.003)	(0.001)	(0.001)
LAGFOLLOWERS	0.033	0.035	0.007	**0.056***
	(0.025)	(0.026)	(0.019)	(0.027)
LAGSOURCES	**-0.040***	**-0.041***	0.008	0.015
	(0.019)	(0.021)	(0.007)	(0.012)
EXPERIENCE*LAGFRIENDS	**-0.004***	-0.004	-0.001	**0.046***
	(0.002)	(0.003)	(0.001)	(0.018)
EXPERIENCE*LAGFOLLOWERS	**-0.029***	**-0.032***	**-0.005***	**-0.145***
	(0.013)	(0.014)	(0.002)	(0.070)
EXPERIENCE*LAGSOURCES	**0.015***	**0.016***	**0.019***	0.000
	(0.006)	(0.007)	(0.011)	(0.036)
Number of Individuals	5,389	5,389	5,389	5,389
Number of Books	16,595	16,595	16,595	16,595
MAD	0.771	**0.770**	0.780	0.778
RMSE	**1.250**	1.255	1.284	1.272

Notes. The posterior standard errors are reported in the parentheses
^ Indicates that 0 is not contained 90% Bayesian Credible Interval
* Indicates that 0 is not contained 95% Bayesian Credible Interval
** Indicates that 0 is not contained 99% Bayesian Credible Interval

Conclusion

In this paper, we study the relationship between one's online social network and decision quality: to what extent does forming new ties, and capturing information from those ties, lead to better decisions? We focus our investigation on two moderating factors: experience and the nature of the tie. After controlling for the ratings environment, individual-level factors, book-level dynamics, self-presentation effects, and, importantly, the potential endogeneity introduced by network variables, our results suggest that increasing the number of social network relationships may, in some cases, be associated with better decisions. Initially, as consumers add more friends ("strong ties"), they may experience higher decision quality while adding more acquaintances ("weak ties") may lead to lower decision quality. In the long run, as one gains more experience in the network, forming more weak ties is associated with better decisions. We expect that this result derives from a fundamental feature of weak-tie relationships. While weak ties may be a good source for novel information, this information may also initially be more difficult to interpret and use, compared with information from strong ties. Only for experienced users, those who have overcome this interpretability hurdle, are weak ties more useful than strong ties in the sense of providing information to improve future decisions

Our results demonstrate significant individual-level learning dynamics. With experience, consumers make better decisions by learning how to best use and interpret opinions from social network ties. Consumers learn how to make better choices after having more experience using information from sources or friends who were originally sources. While opinions from weak ties may initially influence consumers to make worse choices, learning how to use their information may lead to better outcomes in the long run.

We suggest three major implications for our findings. First, we show that consumer learning leads to decision improvement as a result of the formation of online ties. Second, we show that online opinions are not always beneficial to consumers in the short run. While some socially-acquired information may improve decision quality, opinions from others connected through weak ties may initially decrease decision quality. Third, our results suggest that online recommendations are only useful if the consumer can properly process the opinion in order to map them to their own preferences. Otherwise, consumers may find themselves buying a product after incorrectly interpreting a review, and as a result, experience a product that did not meet their expectations.

This research, we believe, is the first to show these dynamic individual-level learning effects with online opinions and to suggest a negative short-term impact that certain online opinions have on decision quality. When shoppers have more weak ties to others in the network, they may make worse decisions if their choices depend on the opinions from those posters. We expect a similar effect when shoppers make choices after reading specific opinions from anonymous individuals. This problem, if not properly addressed, may decrease customer satisfaction over time. Our results show that there may be a simple solution that involves consumers learning how to use reviews from specific posters. We find that consumers who have more experience with the platform have greater satisfaction with their choices if they are following more friends or sources. When shoppers learn how to make better decisions through a larger number of friends or sources, they not only make better choices but also potentially provide higher future ratings for their purchases to reflect that increased decision quality. Therefore, managers who would like to improve consumer decision quality may consider investments targeted at helping consumers to learn more quickly how to process and make use of others consumers' opinions. Such investments may be as simple as building a social network around the review platform, as we observe in Goodreads. It may, as well, take other forms such as better on-line help utilities, more information about other members of the network and better search functionality.

Our research is, of course, not without its limitations. First and foremost, we are only able to estimate aggregate-level parameters for the impact of different social network ties on decision quality. This is due to the relatively-small numbers of social network ties and low within-individual variation, partly due to the fact that we only observe the first 50 days of the platform's evolution. As we expect learning effects to wane over time, this is important for our research design but may overstate the long-term impact of tie formation. A second important limitation with respect to our interpretation of these results is that, while we show that consumers make worse decisions when they increase the number of people they follow, we do not truly know which reviews are viewed, how these reviews are weighted, or the precise source of learning. As a result, for example, we may not be able to separate out learning how to make use of the additional information from learning how to select additional network ties. It might be interesting to use modern eye tracking techniques to identify the specific sources of information that go into the consumer's decision making process as they develop their social network and make decisions based on others' opinions. We also acknowledge that our use of rating as a proxy for decision quality is imperfect. While we have made every effort to control for all factors other than decision quality that may drive one's rating, other possible sources leading to changes in observed ratings may include ratings selection effects (possibly due to a time-varying change in one's ratings cut-off points), book selection effects (consumers could either pick better books over time or be happier by simply having more friends), or scale usage heterogeneity. While we would argue that these effects may not be able to fully explain our non-monotonic results, it would nonetheless be useful to identify both additional measures and alternative research settings which would allow for a more-direct assessment of decision quality. As one example, a lab experiment allowing for direct manipulation of both networks and experience would seem to be a promising future step in this interesting new research stream.

A Market Map of Influence

Introduction

In the age of social media, businesses place heavy emphasis on identifying social media influencers to promote brands and products online (i.e. Forbes' list of the top 50 social media influencers in 2013 (Shaughnessy 2013) and PeerIndex's Social Media King of New York (Taylor 2014). Influencers are individuals whose observed purchase behaviors or recommendations significantly impact the purchase behaviors of others. Thus, firms have an interest in targeting highly influential individuals in their social media campaigns.

Currently, academics and managers define and measure influence as a one-dimensional measure, and these one-dimensional influencer measures have limitations (e.g., Aral and Walker 2012; Trusov et al. 2010; Shaughnessy 2013). One-dimensional measures of influence, such as the probability of being influential (Trusov et al. 2010) or percentage increase in adoption probability (Aral and Walker 2012), assume that an influencer for one product is just as influential for other products in completely different product categories. Take, for example, PeerIndex's selection of Adam Schefter (ESPN NFL Analyst) as the most influential New York social media user. This designation suggests that Adam is universally influential, and he has the same influence in driving sports related choices as consumer electronics related choices.

In reality, we expect that influence is multidimensional, where each individual has separate measures of influence that apply to the different dimensions of a product. With a more flexible measure of influence (i.e. dimension-specific), an individual can be highly influential on one dimension and not influential on the other. For example, an electronic sports watch is related to both sports and consumer electronics. Thus, an individual may be highly influential when promoting the watch's sports related attributes but not influential when promoting electronics related attributes. Dimension specific influence also means that an individual can be highly influential for some products and not for others. Thus, Adam should be heavily influential for choices related to sports but not for choices related to consumer electronics, as we would expect.

However, the identification of dimension specific influence presents three unique challenges. The first challenge is to decompose influence into dimensions that can accurately reflect the attributes of a product. Researchers have used multidimensional scaling and cluster analysis to create market maps that describes products along a smaller set of dimensions (Elrod 1988; Elrod et al. 2002; Lee and Bradlow 2011; Netzer et al 2012). They use choice data or user generated content and plot product locations in m-dimensional space to form a market map. Each dimension reflects latent product attributes, and the location of the product on the map determines the product's weight or values for those attributes. We can use these dimensions to also measure dimension specific influence, allowing one to be more influential for some products and less influential for others.

The second challenge is separating influence from homophily (Hartmann et al 2008; Manski 1993; McPherson et al 2001; Ma et al 2014). Homophily refers to the tendency for individuals who have similar preferences to form ties together. To illustrate the importance of this issue, consider the case where an individual makes a purchase and then, subsequently, her friend makes the same purchase decision. One explanation for this is that the first individual influenced her friend to make the same choice. An alternative explanation could be that both the individual and her friend have the same preferences, and those preferences caused the friend to make that choice. Therefore, we need to separate, explicitly, purchase behavior due to one's baseline preferences for the product from that due to social influence.

The third challenge is to differentiate observational influence from the effects of word of mouth (e.g., Chen et al 2011). Observational influence occurs when an individual makes a decision based on the observed purchase behavior of an influencer. However, positive or negative word of mouth can also affect the impact of influence, above the baseline effect of observational influence. For example, if we observe a highly influential friend make a purchase, then that friend could influence us into making the same purchase. However, if that friend communicates that her recent purchase was a terrible decision (i.e. negative word of mouth), then as a direct result of the negative word of mouth, we would most likely make no purchase. If we do not distinguish word-of-mouth effects from influence measures, then we would erroneously assume that the individual in the previous example has no influence. This is due to current influencer measures (e.g., Aral and Walker 2012, Trusov et al 2010) calculated with the assumption that influence results in others imitating the same behavior as the influencer (i.e. make the same purchase)[29].

In this paper, we propose a new framework and model for identifying dimension specific influentials. Specifically, we model purchase behavior using a proximity model (e.g., Bradlow and Schmittlein 2000) and identify dimension specific influence parameters for each individual based on his/her location, the locations of those who can be influenced, and the locations of the products in the m-dimensional market map. We explicitly model individuals' preferences by estimating their locations on the market map (i.e. assuming consumers prefer products that are closer to their own location on the map). This, in conjunction with dynamic social network data, allows us to disentangle purchase behavior due to homophily from that due to influence.

[29] This assumption is reasonable when measuring behaviors that do not generate much positive or negative word of mouth, such as social network login behavior modeled by Trusov et al. (2010) (Consumers generally do not broadcast how positive or negative they feel specifically for each login decision. However, positive and negative word of mouth is much more prevalent in the context of purchase decisions.

Furthermore, we also separate influence from word-of-mouth effects by explicitly modeling the impact of influence as a function of both observational influence and posted ratings (i.e. 1 to 5 star rating).

For our empirical analysis, we collect and analyze data from Goodreads.com. Goodreads is an online book community where individuals can post the books they have read and provide ratings. In addition, users can also create social ties with other members. The dynamic, social network evolution is critical to help us disentangle homophily from influence, as we assume that one can only influence another after they form a social tie.

Our results show that it is important to estimate dimension specific influence and control for word-of-mouth effects. We show that our proposed model outperforms a model with single measures of influence and a baseline model in which multidimensional scaling is first used to separately identify the locations of the books. We also show that individuals have varying levels of influence across dimensions, and an influencer for one dimension is not always influential on all dimensions.

To further highlight the value of our model, we conduct a series of simulations where we test the marketing strategy of using dimension-specific influentials to promote products. Here, we are in essence seeding influentials (e.g., Aral, Muchnik, and Sundararajan 2013; Goldenberg et al 2009; Hinz, Skiera, Barrot, and Becker 2011; Katona et al 2011; Libai et al 2013) by forcing them to adopt and observing the impact of their adoption (and subsequent influence effects) on the adoption behaviors of all other individuals. The results of this simulation show that seeding dimension-specific influentials can greatly increase future purchase behavior above and beyond (a) seeding influentials determined by single influence measures, (2) seeding random individuals and (b) seeding no one. This highlights the importance of managers in choosing the proper influentials in their social marketing campaigns.

The rest of this paper proceeds as follows. In the next section, we provide a brief overview of our Conceptual Framework. Then, we fully specify our Model in detail. Next, we describe our data in the Empirical Analysis Section and show the results of our proposed analysis. Finally we conclude with a discussion and managerial implications.

Literature Review and Conceptual Framework

Before we discuss our conceptual framework, we first motivate our model by discussing relevant work in two streams of research. The first is the impact of influentials. Given that our research aims to estimate dimension specific influence measures, we first discuss how existing research models influence and assesses the impact of influence. The second is literature on developing market maps. These maps decompose products into different dimensions and may aid us in uncovering dimension specific influence measures.

Influence

Research in offline environments have clearly shown that social effects and influencers play a significant role in the diffusion process (Bass 1969; Bell and Song 2007; Nair et al 2010; Shriver et al 2013; Van den Bulte and Joshi 2007). For example, Bass (1969) theorize that word-of-mouth effects from early adopters influence the adoption decisions of later consumers. Bell and Song (2007) and Shriver et al. (2013) show that neighborhood effects can future buyers who observe the actions of previous buyers. To isolate social effects within a smaller group of influentials, Van den Bulte and Joshi (2007) and Nair et al. (2010) show that there is a segment of influencers that help drive the diffusion process. Therefore, certain types of individuals may have significantly more influence in driving future sales than others.

Given the importance of this segment of influencers, researchers have developed methods to identify influencers and understand the impact of influentials on consumer behaviors (Aral and Walker 2012; Iyengar et al. 2011; Trusov et al. 2010). For example Aral and Walker (2012) use a randomized field experiment to measure how one's purchase observations impact others' purchase decisions and find that certain demographic characteristics are linked to one's ability to influence. In addition, Trusov et al. (2010) model dyadic influence between users of an online social network to assess how one's usage behaviors impact the usage behaviors of other users in the social network. However, these studies have focused on creating single measures of influence. Although these methods are extremely valuable for determining influential users on specific actions (e.g., social media website usage (Trusov et al. 2010) and specific online app purchase (Aral and Walker 2012)), they are unable to assess whether there are different influentials for a wide range of products with different attributes or evaluate an individual’s influence in the various dimensions (or attributes) of a product. The latter may be critical for determining influential users for new products or products that are undergoing significant attribute changes.

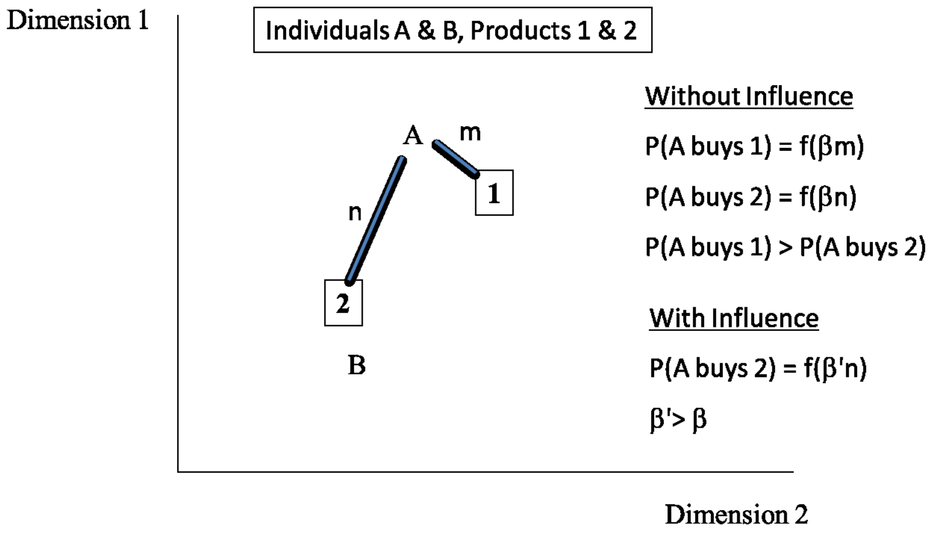
Additionally, most models used to identify influentials do not separate the effect of word-of-mouth from the observational component of influence. This is important, as Chen et al. (2011) find varying effects of observational versus word-of-mouth influence on sales. To illustrate, consider a scenario when an influential buys a product, transmits negative word of mouth, and influences a friend into making no purchases. Without explicitly disentangling the valence (i.e. positive or negative) of the word of mouth, researchers may mistakenly assess no influence for this influential (due to no purchase by the friend after observing a purchase by the influential). To the best of our knowledge, no existing research has examined dimension specific influence and disentangled word-of-mouth effects from influence.

Market Map

In this paper, our main goal is to model dimension specific influence and identify influentials based on multiple influence measures. In a related stream of literature, researchers use market maps to decompose product attributes into a smaller number of dimensions. They plot the locations of products across the market map to assess the substitution and complementary associations between those products. For example, Elrod (1988) and Elrod et al. (2002) use multidimensional scaling on purchase observations to characterize the attributes of products along finite dimensions. Each of these dimensions represents observable or latent attributes related to the set of products. More recently, Lee and Bradlow (2011) and Netzer et al (2012) use text analysis on user generated content to uncover market structure. They tap into the online forums and written reviews to generate product associations from the opinions posted by consumers. While research in market maps does not directly consider the impact of influentials, we use the methods described above to derive product dimensions in order to estimate influence along the same dimensions.

Our research contributes to both the literature in influence and market maps by simultaneously (1) identifying dimension specific influence and (2) estimating the locations of products and the preferences of consumers. Figure 9 presents our conceptual framework for identifying influence in a market map setting. Consider a two-dimensional market map (see Figure 9) with two products (1 and 2) located across the map. We can conceptually think of the axes of the map as different product dimensions (or characteristics), similar to traditional market maps (e.g., Elrod 1988; Elrod et al 2002). A and B on the map are two potential consumers, and the distance between A to the product represents the attractiveness of that product to A. We assume that consumers are more likely to purchase products that are closer to their location (i.e. their preference) on the market map. In Figure 9, individual A's distance to product 1 is relatively shorter than the distance to product 2. Therefore, we assume that A has a greater probability of purchasing product 1.

Figure 9. Two Dimensional Market Map



Impact of influence

When individual B purchases product 2, she may create influence effects for product 2 (e.g., Trusov et al 2010; Iyengar et al 2011; Aral and Walker 2012). This means that, conditional on individual B being influential, B's recommendation of product 2 may increase the attractiveness of product 2, influencing A to extend her reach and purchase a product she may not have otherwise purchased[30]. There are two potential ways that B's influence could increase the attractiveness. First, A may perceive product 2 to be located closer to A. While this is

[30] This is conditional on individual A being susceptible to influence (e.g., Watts and Dodds 2007)

plausible, we expect that this is not the case since B’s purchase would not likely alter the perceived characteristics of the product (i.e. along the dimensions). Instead, we assume that (1) the probability of individual A adopting the product is a function of a coefficient, β, multiplied by the distance between A and the product and (2) influence impacts individual A’s β coefficient on the distance between her and the product. In other words, conditional on the distance between product 2 and individual A, A will be more likely to purchase if she was influenced by B. Of course, this impact of influence depends not only on B’s ability to influence but also on A’s susceptibility to that influence (Watts and Dodds 2007).

The market map framework also allows us to model dimension specific influence along the same dimensions on the market map. In Figure 9, we show the impact of one-dimensional influence. However, we expect that individuals have dimension specific influence, which allows them to be more influential for some products attributes and less influential for others. To incorporate dimension specific influence, we can decompose the distance measure into two components: one for the first dimension and another for the second dimension (in the two dimension case). Then, each individual will have a separate measure of influence that corresponds to each of the dimensions.

Impact of word-of-mouth effects

It is important to highlight that Figures 9 does not include the impact of word -of-mouth effects. As one would expect, in a social setting, many consumers would convey their personal opinions to those they are potentially influencing. Thus, Figures 9 assume that no negative word of mouth is spread. In other words, consistent with Aral and Walker (2012), any purchase observation or social effect resulting from influence always increases the likelihood of others imitating the influencer. This is likely to occur in scenarios where there may not be word-of-mouth effects such as that found in Trusov et al (2010) or in randomized experiments (Aral and Walker 2012). However, one look at many of the online review or social media websites would show that negative word of mouth is very prevalent for many purchase decisions. In theory, when an influential conveys negative word of mouth, this action should reduce the probabilities of others imitating the purchase behavior. However, without explicitly accounting for word of mouth, observers may attribute that the reduction in purchase probability to one's inability to influence. In our model, we will explicitly account for the effects of word of mouth (i.e. a positive 5 out of 5 star rating or a negative 1 out of 5 star rating).

Model Development

In this section, we propose a model estimate dimension specific influence and simultaneously identify the locations of products and individuals in a market map setting. We assume that the products (i.e. book in our empirical analysis) are located across a market map, identified according to purchases made by consumers (Elrod 1988).We also assume that consumers are similarly located across this market map, where their locations reflect their preferences. Furthermore, we assume that consumers prefer products that are located closer to their own location.

To identify preference locations of users simultaneously with the perceptual locations of the products, we model purchase decisions with a proximity model (Bradlow and Schmittlein 2000). Let $y_{ij} = 1$ if we observe individual *i* purchase product *j* at time *t*. We specify the probability of purchase as a function of the location of individual *i* and the location of product *j*:

$$p(y_{ij} = 1) = f(\theta_{im}, \theta_{jm}) = \frac{1}{1+\sum_m d_{ijm}} \tag{1}$$

where θ_{im} denotes the location of individual *i* on dimension *m* and θ_{jm}denotes the location of product *j* on dimension *m*. For the proximity model, we specify d_{ijm} as a function of the Euclidian distance between individual *i* and book *j* on dimension m:

$$d_{ijm} = \beta_{im} \times \left(\sqrt{\left(\theta_{im} - \theta_{jm}\right)^2}\right) \tag{2}$$

In equation (2), the distance between θ_{im} and θ_{jm} represents an individual's preference for that product, such that *i* prefers products that are closer to θ_{im}. In addition, β_{im} represents an individual's sensitivity to that distance. For identification purposes[31], we set $\beta_{im} > 0$. The likelihood for function for (1) and (2) is specified as:

$$L\left(y_{ij} \mid \theta_{im}, \theta_{jm}, \beta_{im},\right) = \prod_i \prod_j {P_{ij}}^{y_{ij}} + \left(1 - P_{ij}\right)^{(1-y_{ij})} \tag{3}$$

Influence

In this section, we build on the proximity model by adding an influence component. Given we expect that some individuals may be more influential for certain books and less so for others, we model influence as an individual-level, dimension specific measure. To model influence, we assume that individual *i* has a certain preference with the product conditional on her location and is more likely to purchase product *j* if *i* is located close *j*. Conditional on observing an influential purchasing product *j*, *i* may be more likely to purchase and thus less sensitive to the distance between $\theta_{im}$and θ_{jm}. Thus, we specify the impact of influence as a covariate on the sensitivity to the Euclidian distance between individual *i* and product *j*[32]:

$$\log(\beta_{im}) \sim N(\alpha_0 - Infl_{im}, \sigma^2) \tag{4}$$

$$Infl_{im} = \alpha_i \sum_{u \in \Xi_{i\tau}} \left[(\gamma_{um} + \delta * Rating_{uj}) \times 1\left(t_{uj} > \tau\right)\right] \tag{5}$$

[31] We also fix the first individual's location to $\theta_{11} = \theta_{12}$= 0, the second individual's location to θ_{21}= 0, and the third individual's location to $\theta_{31} > 0$ in the 2-dimensional model.

[32] We subtract the impact of influence because a smaller β increases the probability of purchase. Thus, this allows us to describe individuals with greater influence estimates as being more influential.

where the impact of influence, $Infl_{im}$, is a function of individual $i's$ susceptibility[33] to be influenced (α_i), an influential user's (u's) ability to influence on dimension m (γ_{um}), and a word-of-mouth effect δ conditional on the rating posted by user u[34]. As a result, a greater impact of influence will increase individual $i's$ preference for product j, conditional on i being susceptible to that influence. We highlight that in order for user u to influence i, user u must be a friend of i and purchase the book before $i's$ purchase time, τ (i.e. $u \in \Xi_{\mathrm{i}\tau}$ and $1(t_{uj} > \tau)$ respectively, where $\Xi_{\mathrm{i}\tau}$ is the set of $i's$ friends at time τ).

Separating Influence from Homophily

In Equations (4) and (5), identification of the individual-level influence parameters may be of a concern. If user u purchases the book before i, we are unable to completely attribute $i's$ purchase decision to the influence of u since these two individuals may have purchased due to similar preferences (e.g., homophily, Ma et al. 2014). To address this in our model, we separate out homophily (i.e. similar purchase behavior due to similar preferences) in two steps. First, our market map approach captures preferences by identifying the locations of each individual across m dimensions. Thus, for each individual, the distance between her location and the location of product j captures her baseline preference for product j. We depict this in Figures 10a and 10b, where individuals A and B both purchase product 1 due to similar preferences but only B purchases product 3.

Second, to identify preferences and separate it from homophily, we take advantage of the dynamics in social network formation by using the point in the time at which two individuals

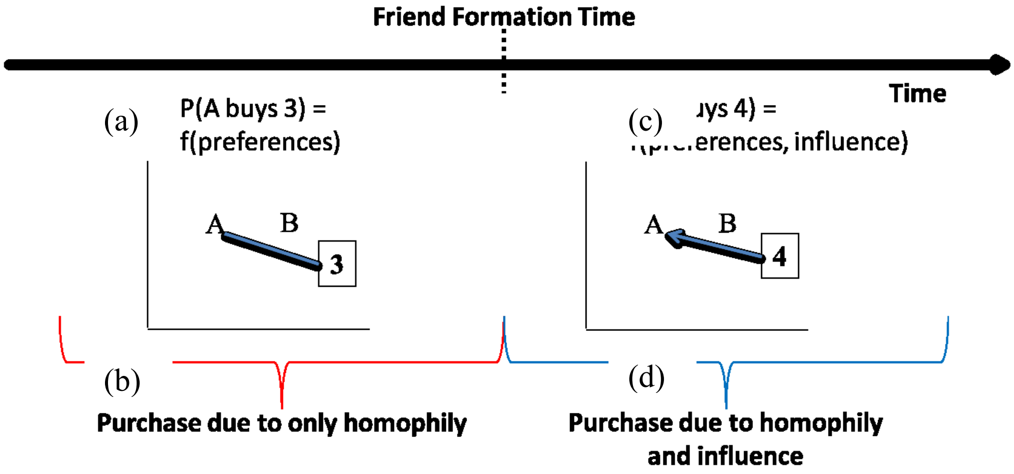
[33] Our operationalization of the impact of influence is consistent with Watts and Dodds (2007) and Trusov et al. (2010), who model the impact of influence as a function of the influencer's level of influence and the affected individual's susceptibility to influence.

[34] 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 Equation 5.

form a connection and isolating the periods before and after that tie formation time. Before two individuals are connected, we assume that they are unable to influence the behaviors of one another and attribute any behavior during that time to homophily. This behavior is reflected in Figure 10a, when individuals A and B are not connected. Thus, the probability of A purchasing product 3 is only a function of A's preferences. After the tie formation, we allow influence to also play a role in affecting behavior. Figure 10b shows that individual B's purchase of product 4 increases the attractiveness of that product to individual A, thus influencing A's purchase as the two have already formed a tie. Therefore, the probability of A purchasing product 4 is a function of both preferences and influence. To incorporate tie formation as our identification strategy, we modify Equation (4) and specify that the influence effect is conditional individuals *i* and *u* being friends (i.e. $g(i,u) = 1$). This allows us to cleanly identify each individual's locations on the market map (i.e. their preferences) using purchases before their tie formation (i.e. $g(i,u) = 0$) and identify an influence affect for purchases that occur after their tie formation.

$$\log(\beta_{im}) \sim \begin{cases} N(\alpha_0 - Infl_{im}, \sigma^2) & if\ g(i,u) = 1, y_{ujt} = 1 \\ N(\alpha_0, \sigma^2) & otheriwse \end{cases} \tag{6}$$

Figure 10: Homophily vs. Influence



To estimate the model, we assume that the locations of all users and books are normally distributed: $\theta_i \sim N(\bar{\mu}_i, \Sigma)$, $\theta_j \sim N\left(\bar{\mu}_j, \Omega\right)$. θ_i represents individual i's preferences for products on the market map and accounts for individual-level heterogeneity in terms of one's preferences for different products. We also capture heterogeneity in individuals' sensitivity to the distance between their location and the book location with individual-level β_i parameters. In addition, θ_j represent the latent attributes of a book and accounts for book-level heterogeneity. To identify the locations (θ_i, θ_j), we fix the first individual's location to $\theta_{11} = \theta_{12} = 0$, the second individual's location to $\theta_{21} = 0$, and the third individual's location to $\theta_{31} > 0$.

We also assume that the measures for influence and susceptibility are normally distributed: $\gamma_u \sim N(\bar{\gamma}, \Psi)$, $\bar{\gamma} \sim N(\bar{\gamma}_0, \sigma_0^2)$, $\Psi \sim InverseWishart(a_0, B_0)$, $\log(\alpha_i) \sim N(\bar{\alpha}, \sigma_\alpha^2)$, and $\sigma_\alpha^2 \sim Gamma(a_0, b_0)$. This specification allows us to account for heterogeneity across individuals in both their ability to influence and susceptibility to influence. Therefore, greater values of γ_u (a vector that measures user *u's* influence across *m* dimensions) means that user *u* is more influential and greater values of α_i means that user *i* is more susceptible to influence. We highlight that for identification purposes, we constrain $\alpha_i > 0$ and set $\bar{\alpha} = 0$. We use a block Metropolis-Hastings algorithm to separately draw each parameter. We run 220,000 iterations and discard the first 200,000 for burn-in. The remaining 20,000 iterations are used to form our posterior results.

Empirical Analysis: Online Book Postings and Ratings

We collect our data from Goodreads.com, an online book community that allows individuals to not only post books that they have read but also create social ties. On Goodreads, consumers can make two separate but related actions. First, an individual can list a book on her profile and publically designate that she has read the book. We refer to this action as *posting* a book. The book is then listed in a virtual bookshelf with all the other books that she has posted as read. Second, an individual can also provide a 1 to 5 star rating for the book. We refer to this action as *rating* a book, and this represents the individual's word-of-mouth opinion. We highlight that Goodreads users do not have to rate every book that they have posted. However, a book with a rating is automatically designated as a post. We make the assumption that all posted books (whether rated or not) have been purchased, providing us with a source of purchase observations[35].

The unique aspect of Goodreads community is the social network formed by the users. An individual on Goodreads can create a friend relationship with any of the other Goodreads members as long as the other member reciprocates that connection. By forming a friend connection with other members, users can easily view their friends' book posts and ratings. For example, Goodreads.com prominently displays the books posted or rated by friends on a user's home page in a "Recent Updates" list and notifies through email. In addition, social network postings are shown first when a reader arrives at the review page for any book. Importantly, while these postings and ratings are easily seen when an individual becomes connected through the friend relationship, they do not stand out or are easily visible when there is no friend relationship[36]. Thus, we assume that before the friend formation time between two individuals, those individuals are not exposed to the postings and ratings by the each other. They are only exposed to postings and ratings – and as a result, potential influence effects – after the formation time[37].

[35]Ideally, we would observe individual level purchase behavior in our dataset. This data is available to the firm, as Goodreads.com tracks and records all referred purchases from third party websites (e.g., Amazon.com) by their members. However, due to privacy concerns, we are unable to obtain the data.

[36] In the case of no relationship between two individuals, the posts are not visible and the ratings are simply one out of the many reviews (on average 3,000 reviews) per book.

[37] Goodreads also allows users to follow other community members. This type of connection allows users to see the posts and reviews of individuals they follow. However, the individuals they follow cannot easily see their own posts

To model dimension specific influence, we collect the book postings/ratings and social network of a sample of individuals and model the effects of dimension specific influence on future purchase behaviors. Our sample consists of a random selection of 200 users who were a member of Goodreads in the beginning of April 2010 and created at least one social network friend and posted at least one book before and after their first tie formation date. We collect each user's social network once every two days over a period of 50 days in April and May of 2010. We also collect all books posted as "read" or rated by not only the original 200 individuals but also all their friends and friend of friends. We use this snowball sampling method in order to obtain all social network connections and posting observations for our final sample of individuals consisting of the original random 200 and their friends. We retain 414 of the original sample's friends who posted at least two or more books during our observation period[38].

Our final sample of 614 individuals (the original 200 and their friends) was responsible for posting or rating 4,617 books during our observation period. We then retain the 50 most posted books to provide sufficient instances where multiple individuals purchase the same book[39]. This consists of our set of products. For each individual i in our sample, we set $y_{ij} = 1$ for all j books that they have posted and $y_{ij} = 0$ for all books that their friends and other users have posted but they have not. Therefore, the no purchase observations consist of books that their friends have read and shared but that they have not. In total, our set of individuals posted a total of 3,198 observations. Furthermore, we also collect 2,405 ratings observations (1 to 5 star rating) posted by our sample of individuals. We use these ratings to create our $Rating_{uj}$

or reviews. For the purpose of our study, we only consider the friends relationship when recording the social network.

[38] We exclude individuals who do not post since the absence of posting means that there is limited (or no) opportunities for influence to play a role in purchase decisions.

[39] In order to estimate influence effects, we need to have multiple individuals purchase the same books. We discard many book observations since, as expected, a large proportion of books (1,613) were only purchased by one individual in our sample.

variable. Table 16 provides the summary statistics of our data (614 individuals and 50 books). The number of books posted/rated, the average ratings, and the number of friends all pertain to our focal set of individuals. We also present the summary statistics for the number of posts and ratings that each individual is exposed to from their friends (i.e. posted by all of their friends). Furthermore, we also provide a book overlap ratio, which is calculated, for each individual-friend combination, the number of similar posted books divided by the number of total number of unique books posted by the two individuals.

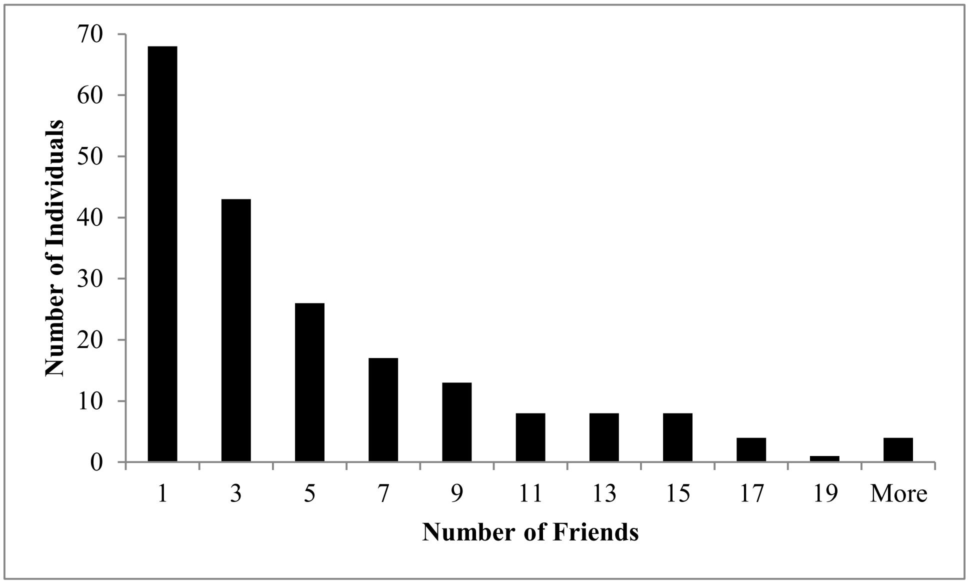
Table 16. Summary Statistics

	Mean	St. Dev.	Min	Max
Number of Books Posted	5.208	6.533	2	39
Number of Books Rated	3.917	4.560	0	30
Average Rating	3.616	1.018	1	5
Number of Friends	10.960	16.845	1	141
Number of Posts by Friends	16.152	6.549	5	34
Number of Ratings by Friends	12.147	6.496	0	29
Book overlap ratio*	0.195	0.198	0	1

*We define the book overlap ratio as the number of similar posted books between a user and a friend divided by the total number of unique books posted by the same two individuals

Figure 11 shows the distribution of the number of ties across users. We highlight that we capture the evolution of these ties over a period of 50 days. This dynamic social network data allows us to categorize books that were posted by the individual when individual *i* is or is not connected to friend *u*. This categorization is important in order to disentangle influence from homophily. As described in the model development section, we use observations before the tie formation time to identify preferences (i.e. the locations of individuals on the market map) and measure an influence effect only for observations after the tie formation time. Again, we assume that there is no shared information before the friend formation time between two individuals, as the postings and ratings are not easily visible to each other. In our data, on average, 30.2% of books were posted before a connection between two individuals and 69.8% of books were posted after.

Figure 11. Distribution of Friends



Model Comparisons

We estimate our model on the Goodreads data. Our proposed model has the following components: the locations of books, the locations of individuals, and dimension specific influence measures. Therefore, we estimate a number of different comparison models to test the value of each component (see Table 17). Our first model (Model 1) is our baseline model and treats the product mapping and individual mapping as two different processes. First, we calculate the locations of the books using multidimensional scaling (e.g., Elrod 1988). Multidimensional scaling (MDS) uses data that describes the perceived similarities of the products to triangulate them onto a market map in multidimensional space. On the map, the locations of the products represent their characteristics along the dimensions of the map, and products located near each other are more similar than products located far from each other. Similar to Elrod (1988), we use consumers' purchase (posting in our case) data to create the similarity matrix. Each row and column of the similarity matrix designates a unique book and the values inside the matrix, for row n and column m, corresponds to the number of individuals who post both books n and m. Using MDS on the similarity matrix, we estimate the locations of the books on the market map. Then, with the book coordinates as given (i.e. MDS coordinates) we separately estimate the individual and influence components of the proposed three dimensional model.

For Model 1, we also assess the value of dimension specific influence by estimating three variations of Model 1 that differs in the influence component. Model 1a assumes no influence (and thus no influence component), Model 1b assumes one-dimensional influence (i.e., each user has one parameter for influence that is the same across all dimensions), and Model 1c contains heterogeneity in influence across dimensions (i.e. dimension specific influence). Table 17 shows the corresponding components for each of the models. For comparison purposes, all models denoted with (a) contains no influence, models denoted with (b) contains one dimensional influence, and models denoted with (c) contains multiple dimensions of influence.

In addition to Model 1, we also evaluate our proposed model by jointly estimating the locations of books, locations of individuals, and the influence estimates. Models 2, 3, and 4 (see Table 17) refer to our proposed model under different market map dimensions specifications (one dimensional, two dimensional, and three dimensional, respectively). Again, we also include variations (a, b, and c) to evaluate the performance of our proposed dimension specific influence specification. Specifically, Models 2a and 2b are one-dimensional models without and with influence, respectively. Models 3a, 3b, and 3c are two-dimensional models (in the market map component) without influence, with one-dimensional influence, and with two dimensional influence parameters, respectively. Finally, Models 4a, 4b, and 4c are three-dimensional models (in the market map component) without influence, with one-dimensional influence, and with three-dimensional influence parameters, respectively.

Table 17: Model Comparisons

Model	No Influence	1 Dimensional Influence	Multidimensional Influence
1) MDS fitted book locations	1a	1b	1c
2) One dimension	2a	2b	N/A
3) Two Dimension Market Map	3a	3b	3c
4) Three Dimension Market Map	4a	4b	4c

We compare the models described in Table 17 by first computing posterior estimates resulting from the model applied to a calibration dataset (a random 80% of our total observations). We then evaluate model fit (in Table 18) by using the posterior estimates to forecast the remaining 20% of the data (holdout sample). We compute three separate metrics to evaluate model fit: (1) posterior likelihood, (2) model based purchase probabilities for both purchase and non-purchase observations, and (3) hit rates.

First, the posterior likelihood forecast uses the posterior estimates from the calibration dataset to calculate the likelihood of the remaining 20% of the data. This allows us to assess how well our posterior estimates fit out of sample data. Table 18 suggests that including additional dimensions on the market map improves fit (comparing Models 2, 3, and 4). We also find that a two dimensional market map with two dimensional influence (Model 2c) and all three dimensional market maps (Models 3a, 3b, and 3c) fits better than Model 1a, 1b, and 1c, which separately identifies the locations of books using multidimensional scaling. We also choose the three dimensional market map specification (Model 4c) as our final model because the posterior likelihood on the holdout sample of the four dimensional map (with four dimensional influence) does not fit as well as the three dimensional specification (-1741.712 versus -1,736.697, respectively).

Second, we calculate the probability of purchase by using the posterior estimates from the calibration data set and forecasting the probability of purchase for each individual-book observation in the holdout sample. Table 18 separately reports the probability of purchase for both actual purchase observations (y_{ij}=1) and non-purchase observations (y_{ij}=0). This allows us to assess whether the models are accurately predicting purchases when the actual observation is a purchase. In addition, we calculate the probability of purchase for all the data in the holdout sample (labeled [All] in Table 18) and a portion of the data that only includes the individual-book combinations for which the individual was previously exposed to the book through a social network tie (i.e. a friend previously posted the same book, labeled [Social] in Table 18). We find that the probabilities of purchase for purchase observations are greater than that for non-purchase observations, suggesting that our model can accurately predict purchase occurrences. This is the case for the forecasts using all the data or the social observation data. In addition, the differences between the two values are greater when using our proposed three dimensional model with dimension specific influence. This suggests that our proposed model can better predict purchase activities compared to other baseline models.

Finally, we also calculate a hit rate for each model as the percentage of the observed purchases that we accurately predict. Specifically, we use our posterior estimates to compute the probability an individual will purchase a given product. Then we simulate purchase using

Bernoulli draws with the calculated purchase probability. The hit rate is the average percent of accurate purchase predictions across 10,000 simulation iterations. We also separately calculate hit rate for all data as well as a portion of the data that only includes individual-book combinations for which the individual was previously exposed to the book through a social network tie. Table 18 shows hit rates for all models. Again, we find that increasing the market map dimensions improves hit rate (for both all data and for social observation data). Furthermore, models that include dimension specific influence outperform the corresponding models that do not (i.e. comparing Model 3c with Models 3b and 3a). These results clearly show the value for each additional dimension and influence parameter in our proposed model specification. Next, we discuss the results associated with the full model specification (Model 4c).

Table 18: Model Comparisons (Posterior Likelihood and Hit Rate)

	Posterior	Probability of Purchase		Hit Rate	Probability of Purchase		Hit Rate
Model	Likelihood	[All]*		[All]*	[Social]*		[Social]*
		y=1**	y=0**		y=1**	y=0**	
4c	-1736.697	0.4881	0.1972	0.6629	0.5069	0.1954	0.7858
4b	-1775.272	0.4837	0.2083	0.6501	0.4368	0.2035	0.7735
4a	-1786.229	0.4733	0.2908	0.6429	0.4046	0.2863	0.6616
3c	-1809.588	0.4642	0.1899	0.6479	0.4875	0.1893	0.7241
3b	-1987.941	0.4311	0.2647	0.6384	0.3975	0.2454	0.6792
3a	-2013.128	0.3928	0.2112	0.6357	0.3757	0.2521	0.6084
2b	-2063.23	0.4722	0.3390	0.5871	0.4749	0.3380	0.6164
2a	-2091.209	0.4012	0.3562	0.5316	0.3612	0.3479	0.5612
1c	-1910.578	0.4399	0.2789	0.6412	0.4123	0.2490	0.6821
1b	-1932.109	0.4118	0.3021	0.6395	0.4101	0.2824	0.6637
1a	-1981.406	0.4081	0.3102	0.6314	0.3912	0.3014	0.6312

*calculated using either all the data (each individual-book combination observation) or only observations where the book was posted (i.e. recommended) by a friend in the social network

**average model based probability of purchase for all (no) purchase observations (i.e. y=1 or y=0)

Model Results

Table 19 provides the estimates for our proposed model. For the user location, book location, influence and susceptibility, Table 19 reports the mean effect of the individual level parameters and the standard deviation of the distribution of those parameters. Both the individuals and the books are more diffusely spread out across dimensions 2 and 3 (standard deviation of 0.188, 0.181, 0.144, and 0.134 for θ_{i2}, θ_{i3}, θ_{j2}, θ_{j3} respectively). First, to visually show these parameter estimates on the market map, we present three two dimensional maps – Figures 12, 13, and 14 – representing dimensions 1 and 3, 1 and 2, and 2 and 3 respectively. In these market maps, the small dots represent the locations of the individuals and the larger colored dots represent the books. Each color is a unique genre of that book as determined by the most commonly designated genre for that book by the Goodreads community. The results suggest that our proposed method produces some clusters of books that are similar in terms of content. We highlight a few that are more distant from the center and are closely related to other books in our dataset. As seen in Figure 12, *The Lord of the Rings* and *The Hobbit*, two books in the same fantasy series are located closely together. The same is true for the three *Harry Potter* books in our sample. Beyond popular series books, we also have three non-fiction commentary books that are located close to each other (*Fast Food Nation, America: A Citizen's Guide to Democracy Inaction,* and *Freakonomics*) and two history oriented books that are neighbors (*1776* and *Collapse: How Society Choose to Fail or Succeed*).

Table 19: Parameter Results

	Mean Effect	Std Dev
User Location 1	0.088	0.116
User Location 2	-0.049	0.188
User Location 3	0.269	0.181
Book Location 1	0.088	0.110
Book Location 2	-0.049	0.144
Book Location 3	0.269	0.134
Influence 1	-0.012	0.853
Influence 2	0.009	0.809
Influence 3	0.095	0.893
Susceptibility	0.350	0.528
	Mean	Std Dev
WOM Valence	0.0005	0.0001

Figure 12. Dimensions 1 and 3

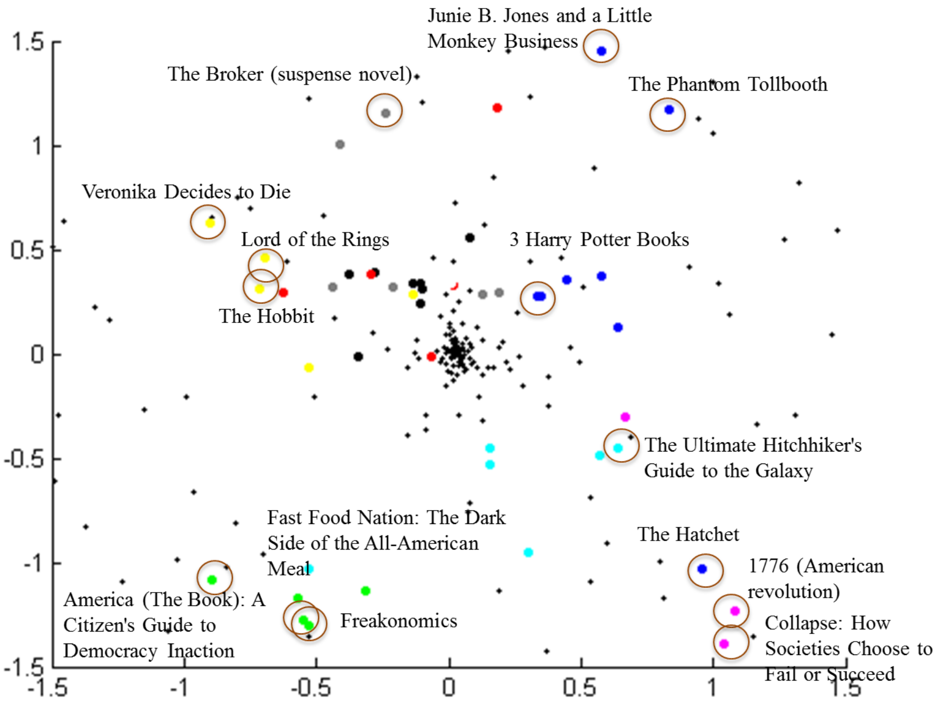


Figure 13. Dimensions 1 and 2

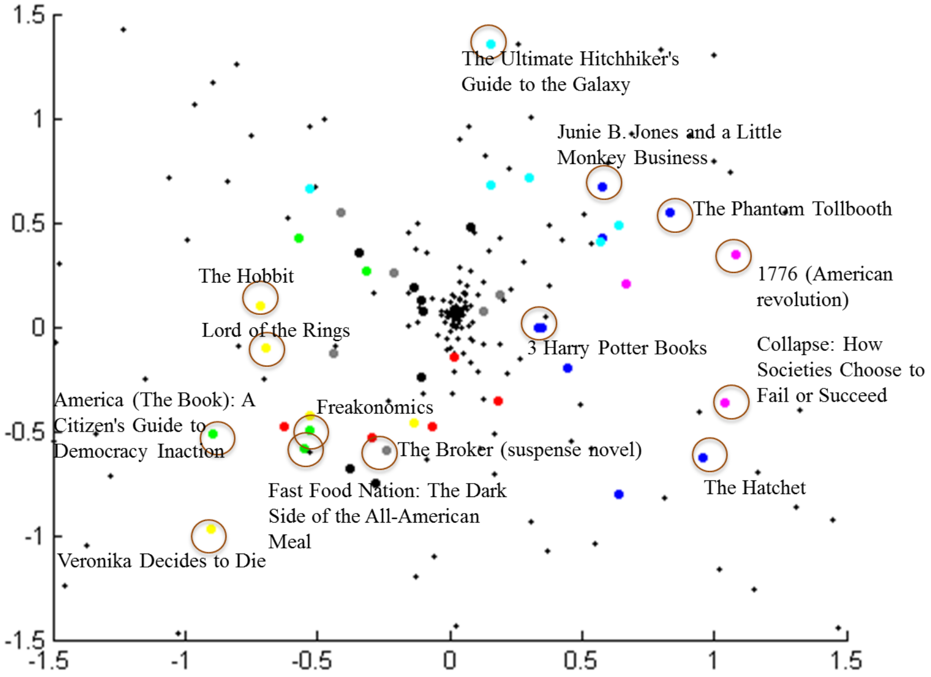
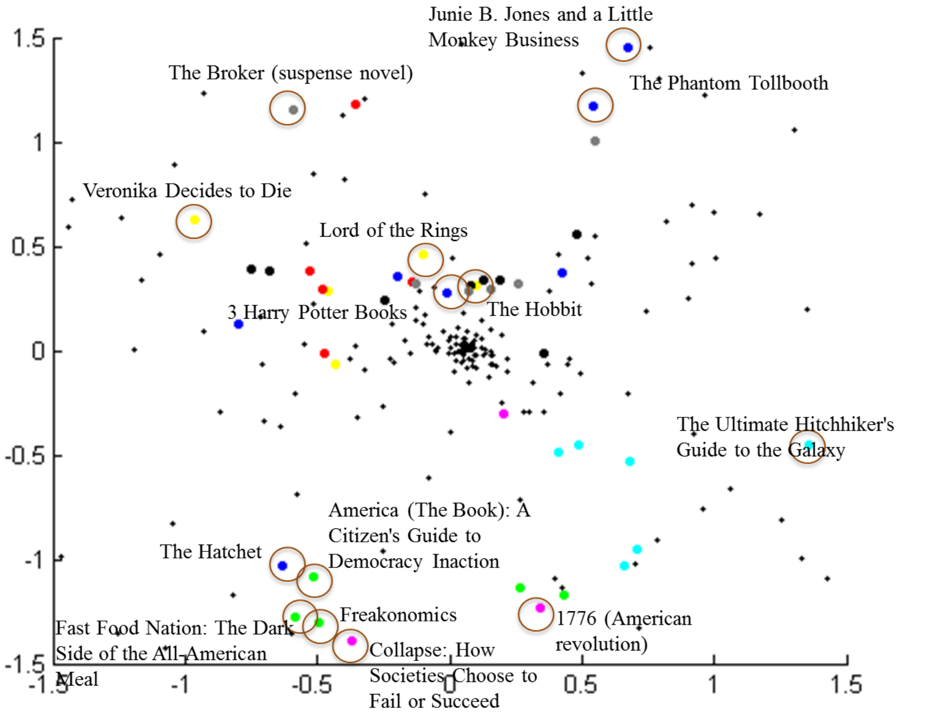


Figure 14. Dimensions 2 and 3



Second, we discuss our results for influence and susceptibility. In terms of word-of-mouth effects, the results in Table 19 suggests that higher ratings increases the impact of influence, thus raising the probability of purchase for those that are influenced (WOM Valence = 0.005). For our influence parameters, the aggregate level influence parameters in Table 19 all contain zero within the 95% Bayesian credible interval (Infl1= -0.012, Infl2=0.009, Infl3=0.095). However, we do find that many individuals have the ability to influence others (i.e. their individual level estimates are positive and do not contain zero within the 95% Bayesian credible interval). Figures 15, 16, and 17 show the distribution of the individual-level influence parameters across the three dimensions. Consistent with Trusov et al. (2010), we find that the majority of our individuals are not influential. However, again similar with Trusov et al. (2010), we do find limited numbers of influentials, with the distribution following an exponentially shaped pattern. Notably, we also find many users who have "negative" influence. This suggests that some users' purchase observations decrease the likelihood that others will purchase. In addition, we also find that many individuals are susceptible to the influence of others ($\bar{\alpha} = 0.38$).

Finally, recent research has suggested that influentials may not be as susceptible to influence as non-influentials (Iyengar et al 2011; Aral and Walker 2012). To assess whether influentials may be less susceptible, we present, in Table 20, the correlation between the influence and susceptibility parameters. Consistent with Aral and Walker (2012) and Iyengar et al (2011), we find that those who are influential are less susceptible to the influence of others.

Figure 15. Distribution of Influentials on Dimension 1

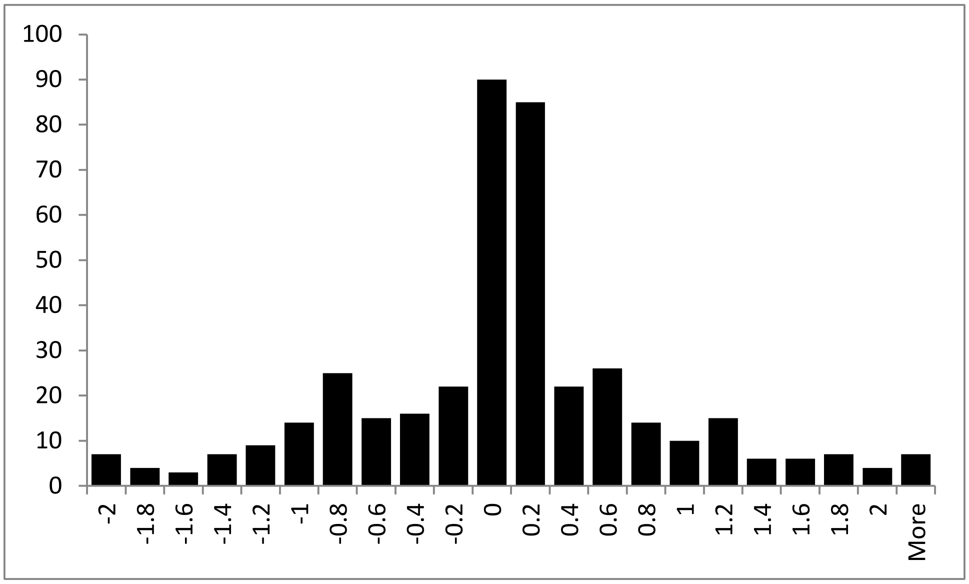


Figure 16. Distribution of Influentials on Dimension 2

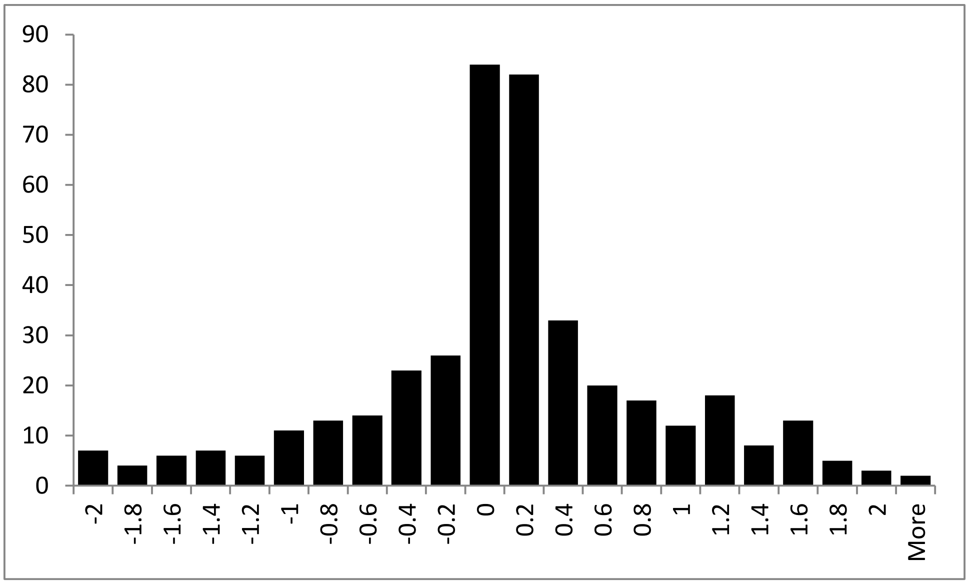


Figure 17. Distribution of Influentials on Dimension 3

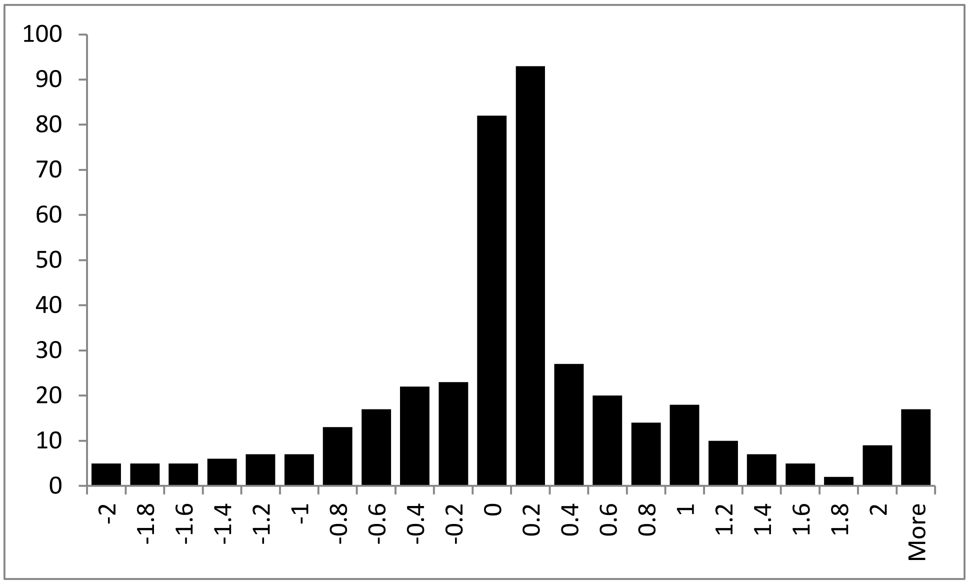


Table 20: Correlation between Influence and Susceptibility

	Infl1	Infl2	Infl3	Suscept
Infl1	1.000			
Infl2	0.198	1.000		
Infl3	0.084	-0.081	1.000	
Suscept	-0.221	-0.139	-0.068	1.000

Simulating Social Media Messaging Strategies

In this section, we present a simulation to assess the extent to which our dimension-specific influentials increase purchase behavior relative to random individuals. Our results suggest that there are unique individuals that are highly influential on one or more dimensions. We expect that for products that are on the extreme of one dimension, it is important to utilize dimension-specific influentials. Therefore, we test strategies that use the following individuals to promote a product (e.g., give them a free product and have them recommend that product to everyone): (1) dimension specific influentials, (2) influentials identified based on a single measure of influence, (3) random, and (4) no individuals.

We simulate purchase behavior related to a specific product by 200 individuals distributed normally (i.e., $\theta_i \sim N(0,1)$) across the three dimensions. First, we specify a product location along each of the three dimensions. We choose products that are located at the center of the distribution of individuals, one standard deviation away on one of the dimensions, and three standard deviations away on one of the dimensions. Next, we decide which users (e.g., influentials, random, or none) to provide a free product to generate influence effects. Then, for each individual i, we compute i*'s* probability of purchasing (P_{ij}) using individual i's and product *j*'s model-based posteriors, the influence (γ_u) of the users that we promote the product to, and susceptibility (α_i) of individual *i*. Finally, we simulate the decision to purchase the book through Bernoulli draws with probability P_{ij}.

We use the posterior estimates from the last 1,000 iterations from the estimation sampler. For each set of posterior estimates, we simulate the purchase behavior that would result from a number of strategic scenarios which we will describe below. We average across 500 simulated iterations for each set of posteriors and then across the 1,000 estimation iterations to obtain our simulation results which we present next.

The results from this simulation are presented in Table 21. In the first row, we show the total number of purchases made without seeding any individuals. Each column refers to a product located in a different area in the three-dimensional market map, with product [0 0 0] located at the center of the distribution of consumers and product [3 0 0] located three standard deviations away from the center for dimension 1. We find that promotion through influential users is significantly more effective at generating future sales than no promotion or promotion with random individuals. Specifically, we find that seeding 5 random users (2.5% of the population) generates about 0.15 to 0.26 times more purchase activity than the no seed condition for products located within one standard deviation of the mean position across consumers. In contrast, our results suggest that seeding 5 influential users (based on a 1 dimensional model) generates 1.12 to 1.6 times more purchase activity. In addition, for products that are located at the extremes (three standard deviations away from the mean position across consumers), we find that seeding influentials produces a marginally greater impact (up to 4.77 times more purchase activity). Therefore, it may be more effective to seed influentials for products located far away from consumers on the market map, as consumers are less likely to organically purchase these products.

More interestingly, our results also suggest that influentials selected using the dimension specific influentials generate greater future sales, for products located at the extremes along a dimension, than those selected using a model with one-dimensional influence. For example, seeding 5 influential users, identified based on a one-dimensional measure, for a book located at [3 0 0] increases purchase activity by 4.77 times above the no seed condition. In contrast, seeding 5 dimension one influentials for the same book increases purchase activity by 5.13 times above the no seed condition. This highlights the importance of a model that allows for dimension specific measures of influence. For products that are located in the extremes of one particular dimension, it is critical to identify individuals that are highly influential in driving purchases along that same dimension.

Table 21: Simulation Results

	Book Location							
Seed:	[0 0 0]	[1 0 0]	[0 1 0]	[0 0 1]	[1 1 1]	[3 0 0]	[0 3 0]	[0 0 3]
No Seed	45.1	37.6	38.3	35.3	30.0	14.4	18.6	17.1
5 Random users	+0.15x	+0.26x	+0.17x	+0.23x	+0.43x	+0.35x	+0.24x	+0.24x
5 Influentials (1D Model)**	+1.12x	+1.48x	+1.42x	+1.6x	+2.08x	+4.77x	+3.53x	+3.95x
5 Dim-1 Influentials*	+0.96x	+1.4x	+1.21x	+1.18x	+1.82x	+5.13x	+2.66x	+1.16x
5 Dim-2 Influentials*	+0.93x	+1.2x	+1.33x	+1.13x	+1.77x	+3.83x	+3.9x	+0.87x
5 Dim-3 Influentials*	+0.84x	+0.74x	+0.96x	+1.27x	+1.53x	+1.63x	+1.72x	+4.21x

*Individuals with the highest dimension 1 (or 2/3) influence measures. May or may not be influential on the other two dimensions

**From a model with one-dimensional influence

Comparing Model Based to Metrics Based Influentials

Using our proposed model, we are able to identify dimension specific influentials. In addition, our simulation shows that seeding strategies targeting these influentials outperform those that target random users. However, there could be other more simple ways to select a target group of influentials. For example, managers could identify users with the most number of friends. Alternatively they could also choose to seed individuals who post frequently.

We use model free evidence to assess our measure of influence compared with other potential selection criteria in order to avoid any potential advantage that our proposed model may place on the influentials identified by the model. First, we identify 10 (and 20) individuals as our target influencers using the following rules. For our simple "influencer" selection rule, we select individuals who have the most (or least) number of friends or who post the most (or least) frequently. For our model based influencers, we select individuals with the greatest influence measure across dimensions 1, 2, or 3. Second, we select all the books in our sample posted by each of the "influencers." Finally, for each dyadic tie between "influencer" and a friend of the “influencer”, we record whether the friend posted the book after the “influencer’s” posting (1 for yes, 0 for no). These consist of our purchase observations. In order to avoid instances where the friend was influenced by multiple individuals, we drop all observations where the friend had at least one other friend post the same book at an earlier time. This isolates all potential influence to the “influencer”.

Table 22 shows the results from this model free analysis. The first column shows the percentage of purchase by friends of our target influencers, calculate by averaging the purchase observations described above. We also present the average number of friends, average number of posts, and the average model calculated influence measures. Our results suggest that the percentage of purchase is significantly greater when we use our model identified influencers compared with the rule of identifying influencers using their total number of posts (both most and least number of posts). However, given the low number of observations when we use 10 influencers (average of 30 observations), a t-test comparing the percentage of purchase between our model identified influencers with the influencers identified by the largest (or smallest) number of friends is not significant. Therefore, we extend our analysis to 20 influencers. The results suggest that the percentage of purchase for 20 dimension 1 and dimension 3 influencers is significantly greater than that for 20 individuals with the most friends at the 90% and 95% level, respectively (also significantly greater than 20 individuals with the least number of friends at the 90% level). While this analysis does not suggest any causal relationship, it provides some confidence that our model identified influentials may be more appropriate for seeding strategies compared to other more simple methods of identifying individuals to seed.

Table 22: Model Based Versus Metric Based Influentials

Focal Influencers	Percent of Common Purchases by Friends	Avg Number of Friends	Avg Dim 1 Infl	Avg Dim 2 Infl	Avg Dim 3 Infl	Avg Number of Posts
10 with most friends	6.9%	131.3	-0.48	0.20	0.46	9.3
10 with most posts	2.2%	108.4	0.16	-0.28	-0.22	19.8
10 with least friends	7.1%	1.0	-0.59	0.10	1.30	11.1
10 with least posts	0.0%	13.1	-0.10	-0.01	0.44	2.0
10 Dim1 infl	**17.1%[2]**	32.3	2.22	0.00	-0.07	7.9
10 Dim2 infl	**14.3%[4]**	46.3	-0.11	1.64	0.03	8.6
10 Dim3 infl	**16%[4]**	23.3	-0.38	0.47	2.70	5.8
20 with most friends	3.2%	88.9	0.11	-0.36	-0.32	7.2
20 with most posts	3.3%	74.4	0.09	-0.08	-0.18	17.6
20 with least friends	5.2%	1.0	-0.33	0.02	0.68	11.7
20 with least posts	3.2%	9.2	0.07	-0.02	0.18	2.1
20 Dim1 infl	**14%[2,3]**	38.1	1.82	-0.02	-0.05	6.8
20 Dim2 infl	9.8%	57.0	-0.11	1.45	0.28	8.1
20 Dim3 infl	**17%[1,2]**	45.4	-0.18	0.17	2.31	6.8

Notes: Focal influencers identified using either metrics (# of friends) or model based measures (dimensions 1-3).

Percent of common purchases calculated using observations of friends who only receive recommendations from the influential

[1,2]Significantly different than influencers selected using the number of (1) friends or (2) posts at the 95% level

[3,4]Significantly different than influencers selected using the number of (3) friends or (4) posts at the 90% level

Conclusion

In this paper, we develop a new method for identifying dimension specific influentials in purchase decision settings using observational data. We take advantage of a dynamic dataset that records purchase behavior before and after social connections are made by individuals and use observations before the connection time as a baseline to determine preferences (i.e. homophily) and after the connection time to estimate an influence affect (in addition to homophily). Overall, our results from our empirical analysis suggest that it is important to separate influence into dimension specific effects and use a unified model to estimate both the locations of individuals and books. We find considerable heterogeneity among one's ability to influence across different dimensions of book characteristics and show that an individual may not be highly influential on all dimensions. Using simulation, we find that for books with extreme characteristic's (e.g., far away from the mean location on one dimension), it is best to target dimension specific influentials, along the same dimension as the most extreme dimension for the book's location, for word-of-mouth promotion. Thus it may be critical for managers to consider individuals who may not be influential on all dimensions but are influential in the appropriate dimension for their products.

Our findings make a number of important contributions. First, we propose a new method to identify dimension specific influentials from observational data using a market map and network tie formation as our identification strategy. This is unique to other researchers who have relied on random experiments or sociometric surveys to infer influence (Aral and Walker 2012 Iyengar et al 2011). We also find evidence of heterogeneity within individuals in their ability to influence across different product dimensions. Second, we explicitly disentangle our influence measure from potential word-of-mouth effects. Given widely prevalent word-of-mouth influence, we directly model this and separate out this effect from our measure of influence. This is critical given that some influentials may spread negative word of mouth, creating situations where others were influenced but made a no purchase decision. Finally, we add to the market map literature (Elrod 1988; Elrod et al 2002; Lee and Bradlow 2011; Netzer et al 2012) by simultaneously estimating the locations of consumers and products and incorporating influence for managerial purposes. Therefore, our framework allows researchers and marketers to not only assess the locations of their products and consumers across different dimensions (i.e. product attributes) but also helps them identify the key influencers for products on a specific dimension.

Our research is also of interests to managers looking to use word-of-mouth marketing strategies and take advantage of influentials in driving purchase decisions. Managers can use our method to identify influentials in their purchase context. Importantly, they can specifically tailor their efforts to individuals who are influential on dimensions that are important to the manager. This is critical since many products may have different features and attributes that appeal to different types of consumers. Managers can also identify areas to launch new products with the aid of word of mouth from certain types of influentials and use our method to determine how these influentials may impact new (or existing) products at a given location.

Our research is not without its limitations. First, while we estimate dimension specific influentials, we are not able to provide clear descriptions of the product characteristics reflected by each dimension. Managers may need to use their judgment to assign certain dimension-specific influentials to a promotion strategy targeted at specific product characteristics. Likewise, an explanation of the dimensions describing the books is not straight forward. While some related books do cluster together on the market map, we cannot pinpoint the exact attribute that impact their location.

Second, for our analysis, we measure influence along the dimensions uncovered by our market map model. However, consumers may be influenced along different dimensions than that uncovered in the market map. For example, consumers may place emphasis on one product characteristic in their purchase decision (e.g., price) but are only influenced along a different characteristic (e.g., quality). We expect that the model identified dimensions are the most relevant for an individual's purchase decision as the market map helps determine the dimensions that consumers rely upon when making purchase decisions. While other dimensions may be a factor, they are not as important in driving purchases as the model identified dimensions. Future research could further examine whether it is beneficial to model influence along additional dimensions beyond that described in the market map.

Third, our proposed model does not account for heterogeneity in consumers' perceived product map, where each individual may use a different map in their decision making process (Desarbo et al 2001). Specifically, while we assume that consumers' perceptions of products along the market map are homogenous, there may exist segments of consumers that view the relationships between products differently. Thus, one consumer may perceive a smaller distance between two products while another may perceive greater differences, and thus a greater distance between the products. Future research could incorporate heterogeneity in the perceived locations into our proposed model.

Finally, our research does not provide explanations to why certain individuals are influential. One potential reason could be the local network characteristics of an individual. For example, individuals with a set of network characteristics, such as a specific betweenness or clustering coefficient among one's local network (e.g. Katona et al 2011), could me more or less influential. Future research can investigate the relationship between dimension specific influence and one's local network characteristics in order to provide firms with a simple method to identify key potential influencers.

Appendices

Appendix 1

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 γ_u =1, user u is considered influential and if γ_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:

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

where p_u is a draw from a Beta distribution representing the proportion of the population that consists of influentials with γ_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 $U \setminus u$ 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:

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

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 2

In this Appendix, we describe our LIV and IV estimation approach. In the LIV approach, we follow Ebbes (2004), Ebbes et al. (2005), Zhang et al. (2009), and Rutz et al. (2012) and estimate a latent discrete instrument to divide the potentially endogenous social network variables into two components: one that is uncorrelated and another that is correlated with the error term. To simplify notation, we discuss a simplified model with one endogenous variable as follows:

$$EVAL_{it}^{*} = \delta_0 + \delta_1 LAGFRIENDS_{it} + \epsilon_{it}$$
$$LAGFRIENDS_{it} = \theta \mathrm{Z_{it}} + \mathrm{\nu_{it}}$$

where Z_{it} is a latent categorical variable with category means of θ. The number of categories needs to be equal or greater than two for identification purposes; however, the model is robust against misspecification of the number of categories (Ebbes 2004). Therefore, we follow Ebbes (2004), Ebbes et al. (2005), and Rutz et al. (2012) and assume that Z_{it} has two categories. We also assume that Z_{it} is orthogonal to $\mathrm{\nu_{it}}$ and ϵ_{it} and, through data augmentation (Tanner and

Wong 1987), follows a bernouli distribution with probabilities (π_1, π_2). Here, π_c is the probability that the c$^{\text{th}}$ latent instrument is one (i.e. belongs to category *c* at time *t*). Furthermore, we assume that the error terms follow a multivariate normal distribution with mean 0 and variance-covariance matrix:

$$\Xi = \begin{bmatrix} \sigma_{\epsilon\epsilon} & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma_{\nu\nu} \end{bmatrix}$$

Thus, the likelihood function is specified as:

$$p(k_{it}|\delta, \theta, \Xi) = (2\pi)^{-1}|\Xi|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(k_{it} - u_{it})'\Xi^{-1}(k_{it} - u_{it})\right]$$

where $k_{it} = (EVAL_{it}^{*}, LAGFRIENDS_{it})$' and $u_{it} = (\delta_0 + \delta_1 LAGFRIENDS_{it}, \theta\mathrm{Z}_{\mathrm{it}})$'. Conditional on known values for θ, Z, and Ξ, we follow Ebbes (2004) and Rutz et al. (2012) and draw δ from a normal distribution with mean $V^{-1}\left[D'\left(\sigma_{\epsilon\epsilon}EVAL^{*} + \sigma_{\epsilon\nu}(D - \theta Z)\right) + \Psi_0^{-1}\delta_0\right]$ and variance-covariance $V = \sigma_{\epsilon\epsilon}D'D + \Psi_0^{-1}$, where $D = [1 \;\; LAGFRIENDS]$. We also draw Ξ from an inverse Wishart distribution with parameters (w_0, R_0) and $(\sum_i \sum_t (k_{it} - u_{it})\,(k_{it} - u_{it})' + \Xi_0)$. θ is drawn from $MVN(a, S)$, where $S = (S_0 + b\sum D'D)^{-1}$, $a = S(S_0 a_0 + b\sum D'LAGFRIENDS)$, $b = \frac{1}{\sigma_{\epsilon\nu}\sigma_{\nu\nu}\sigma_{\epsilon\nu}}$. The latent Z_{it} are categorical variables drawn from the following posterior probability:

$$p(Z_{it} = c) = \frac{\pi_c \times L(k_{it}|u_{it}, \delta, \theta, \Xi, Z_{it} = c)}{\sum_{j=1}^{2} \pi_j \times L(k_{it}|u_{it}, \delta, \theta, \Xi, Z_{it} = j)}$$

where $L(\cdot)$ is the likelihood specified in equation (1) conditional on $Z_{it} = c$. We also draw π_c from a beta distribution: $\pi_c|Z{\sim}Beta(1 + 1[Z = 1], 1 + 1[Z = 2])$. We use the same method for IV estimation, where Z_{it} are observed instruments as opposed to latent.

It is important to note that in the ordered probit model, we set $\sigma_{\epsilon\epsilon} = 1$ for identification purposes. However, this constraint prevents us from estimating Ξ using the traditional Gibbs sampler. Therefore, we relax this constrain and let $\sigma_{\epsilon\epsilon}$ vary and identify valid posterior estimates through post-processing of the posterior draws (McCulloch and Rossi 1994; Edwards and Allenby 2003). While the posterior estimates for δ will not converge because it is not identifiable, the posterior for $\delta/\sigma_{\epsilon\epsilon}$ across all iterations do converge. Thus, after estimation, we set $\sigma_{\epsilon\epsilon} = 1$ and scale all other parameters accordingly. Therefore, we report the values of the post-processed δ as the mean of the converged posterior for $\delta/\sigma_{\epsilon\epsilon}$.

Bibliography

Alba, Joseph W., and J. Wesley Hutchinson (1987), "Dimensions of Consumer Expertise," *Journal of Consumer Research*, 13(4), 411–54.

Albert, James H., and Siddhartha Chib (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88(422), 669–79.

Anger, Isabel and Christian Kittl (2011), "Measuring Influence on Twitter," In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, New York, New York, USA: ACM Press.

Anzai, Yuichiro, Herbert A. Simon (1979), "The theory of learning by doing," *Psychological Review*, 86(2) 124-140.

Aral, Sinan, and Dylan Walker (2012), "Identifying Influential and Susceptible Members of Social Networks," *Science*, 337(6092), 337–41.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan (2013), "Engineering Social Contagions: Optimal Network Seeding and Incentive Strategies", *Network Science* 1(2), 125-153.

Baldwin, Micah (2009), " HOW TO: Measure Online Influence," Accessed December 9, 2012, Available at: http://mashable.com/2009/03/02/measuring-online-influence/

Banerjee, Abhijit V. (1992), "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, 107(3), 797–817.

Barnes, Nora, Ava Lescault, and Justina Andonian (2012), "Social Media Surge by the 2012 Fortune 500: Increase Use of Blogs, Facebook, Twitter and More," Accessed December 9, 2012, Available at: http://www.umassd.edu/cmr/socialmedia/2012fortune500/

Bass, Frank M. (1969), "A New Product Growth for Model Consumer Durables," *Management Science*, 15(5), 215–27.

Berger, Jonah (2011), "Arousal Increases Social Transmission of Information," *Psychological Science*, 22(7), 891–893.

Berger, Jonah, and Eric M Schwartz (2011), "What Drives Immediate and Ongoing Word of Mouth?," Journal of Marketing Research, 48(5), 869–80.

Berger, Jonah and Katherine Milkman (2012) , "What Makes Online Content Viral?," *Journal of Marketing Research*, 49(2), 192-205.

Bell, David, and Sangyoung Song (2007), "Neighborhood effects and trial on the internet: Evidence from online grocery retailing," Quantitative Marketing and Economics, 5(4), 361–400.

Bettman, James (1979), "An Information Processing Theory of Consumer Choice," Reading, MA: Addison-Wesley.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992), "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *The Journal of Political Economy*, 100(5), 992–1026.

Bodapati, Anand V (2008), “Recommendation Systems with Purchase Data,” *Journal of Marketing Research*, 45(1), 77–93.

Borgatti, Stephen P., and Rob Cross (2003), “A Relational View of Information Seeking and Learning in Social Networks*,” Management Science*, 49(4), 432–45.

Bradlow, Eric T., and David C. Schmittlein (2000), “The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines,” *Marketing Science*, 19(1), 43–62.

Brooks, Stephen P., and Andrew Gelman (1998), “General Methods for Monitoring Convergence of Iterative Simulations,” *Journal of Computational and Graphical Statistics*, 7(4), 434–55.

Brown, Jacqueline Johnson, and Peter H. Reingen (1987), “Social Ties and Word-of-Mouth Referral Behavior,” *The Journal of Consumer Research*, 14(3), 350–62.

Chen, Yubo, Qi Wang, and Jinhong Xie (2011), “Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning,” Journal of Marketing Research, 48(2), 238–54.

Chevalier, Judith, and Dina Mayzlin (2006), “The Effect of Word of Mouth on Sales: Online Book Reviews.,” *Journal of Marketing Research*, 43(3), 345–54.

Chintagunta, Pradeep K, Shyam Gopinath, and Sriram Venkataraman (2010), “The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets,” *Marketing Science*.

Clemons, Eric, Guodong Gao, and Lorin Hitt (2006), “When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry,” *Journal of Management Information Systems*, 23(2), 149–71.

De Bruyn, Arnaud, and Gary L. Lilien (2008), “A multi-stage model of word-of-mouth influence through viral marketing,” *International Journal of Research in Marketing*, 25(3), 151–63.

Dellarocas, Chrysanthos, Xiaoquan Zhang, and Neveen F. Awad (2007), “Exploring the value of online product reviews in forecasting sales: The case of motion pictures.,” *Journal of Interactive Marketing (John Wiley & Sons)*, 21(4), 23–45.

DeSarbo, Wayne S., Daniel J. Howard, and Kamel Jedidi (1991), “Multiclus: A new method for simultaneously performing multidimensional scaling and cluster analysis,” *Psychometrika*, 56(1), 121–36.

DeSarbo, Wayne S., Kamel Jedidi, and Indrajit Sinha (2001), “Customer value analysis in a heterogeneous market,” Strategic Management Journal, 22(9), 845–57.

Du, Rex Yuxing, and Wagner Kamakura (2011), “Measuring Contagion in the Diffusion of Consumer Packaged Goods,” *Journal of Marketing Research*, 39(2), 28-47.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008), “Do online reviews matter? -- An empirical investigation of panel data,” *Decision Support Systems*, 45(4), 1007–16.

Ebbes, Peter (2004), "Latent instrumental variables: a new method to solve endogeneity in marketing and economics," PhD thesis University of Groningen, Labyrinth Publications, Ridderkerk.

Ebbes, Peter, Michel Wedel, Ulf Böckenholt, and Ton Steerneman (2005), “Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income,” *Quantitative Marketing and Economics*, 3(4), 365–92.

Edmunds, Angela, and Anne Morris (2000), “The problem of information overload in business organizations: a review of the literature,” *International Journal of Information Management*, 20(1), 17–28.

Edwards, Yancy D., and Greg M. Allenby (2003), “Multivariate Analysis of Multiple Response Data,” Journal of Marketing Research, 40(3), 321–34.

Elrod, Terry (1988), “Choice Map: Inferring a Product-Market Map from Panel Data,” *Marketing Science*, 7(1), 21–40.

Elrod, Terry, Gary J. Russell, Allan D. Shocker, Rick L. Andrews, Lynd Bacon, Barry L. Bayus, J. Douglas Carroll, Richard M. Johnson, Wagner A. Kamakura, Peter Lenk, Josef A. Mazanec, Vithala R. Rao, and Venkatesh Shankar (2002), “Inferring Market Structure from Customer Response to Competing and Complementary Products,” *Marketing Letters*, 13(3), 221–32.

Experian Marketing Services. (2013). Experian Marketing Services Reveals 27 Percent of Time Spent Online is on Social Networking [Press release]. Retrieved from <http://press.experian.com /United-States/Press-Release/experian-marketing-services-reveals-27-percent-of-time-spent-online-is-on-social-networking.aspx?WT.srch=PR_EMS_OnlineTime_041613_gpo>.

Foster, Andrew D., and Mark R. Rosenzweig (1995), “Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture,” *Journal of Political Economy*, 103(6), 1176–1209.

Friestad, Marian and Peter Wright (1994), "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts, Journal of Consumer Research, 21 (June), 1-31.

Gelfand, Alan E., and Adrian F. M. Smith (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85(410), 398–409.

Godes, David, and Dina Mayzlin (2004), “Using Online Conversations to Study Word-of-Mouth Communication,” *Marketing Science*, 23(4), 545–60.

Godes, David, and Dina Mayzlin (2009), “Firm-Created Word-of-Mouth Communication: Evidence from a Field Test,” *Marketing Science*, 28(4), 721–39.

Godes, David, and Jose Silva (2012), “Dynamics of Online Opinion,” *Marketing Science*, 31(3), 448-473.

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.

Granka, Laura A., Thorsten Joachims, and Geri Gay (2004), “Eye-tracking analysis of user behavior in WWW search,” in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR ’04, New York, NY, USA: ACM, 478–79.

Granovetter, Mark S. (1973), “The Strength of Weak Ties,” *The American Journal of Sociology*, 78(6), 1360–80.

Green, Paul E., and V. Srinivasan (1978), “Conjoint Analysis in Consumer Research: Issues and Outlook,” *Journal of Consumer Research*, 5(2), 103–23.

Friestad, Marian and Peter Wright (1994), "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts, *Journal of Consumer Research*, 21 (June), 1-31.

Hartmann, Wesley R., Puneet Manchanda, Harikesh Nair, Matthew Bothner, Peter Dodds, David Godes, Kartik Hosanagar, and Catherine Tucker (2008), "Modeling social interactions: Identification, empirical methods and policy implications," *Marketing Letters*, 19(3-4), 287–304.

Hansen, Morten T. (1999), "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits," *Administrative Science Quarterly*, 44(1), 82–111.

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

Hinz, Oliver, Bernd Skiera, Christian Barrot, and Jan U Becker (2011), "Seeding Strategies for Viral Marketing: An Empirical Comparison," Journal of Marketing, 75(6), 55–71.

Hoyer, Wayne D., and Steven P. Brown (1990), "Effects of Brand Awareness on Choice for a Common, Repeat-Purchase Product," *Journal of Consumer Research*, 17(2), 141–48.

Ho, Teck-Hua, Shan Li, So-Eun Park and Zuo-Jun Max Shen (2012), "Customer Influence Value and Purchase Acceleration in New Product Diffusion," Marketing Science, 31 (2), 236-256.

Hutchinson, J. Wesley and Amitbah Mungale (1997), "Pairwise Partitioning: A Nonmetric Algorithm for Identifying Feature-Based Similarity Structures," *Psychometrika*, 62 (1), 85-117.

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W Valente (2011), "Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30(2), 195–212.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng (2007), "Why we twitter: understanding microblogging usage and communities," in *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, USA: ACM, 56–65.

Johnson, Eric J., Steven Bellman, and Gerald L. Lohse (2003), "Cognitive Lock-In and the Power Law of Practice," *Journal of Marketing*, 67(2), 62–75.

Katona, Zsolt, Peter Pal Zubcsek, and Miklos Sarvary (2011), "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research (JMR)*, 48(3), 425–43.

Kirmani, Amna (2009), "The Self and the Brand," Journal of Consumer Psychology, 19 (3), 271-275.

Kirmani, Amna and Baba Shiv (1998), "The Effects of Source Congruity on Brand Attitudes and Beliefs: The Moderating Role of Issue-Relevant Elaboration," Journal of Consumer Psychology, 7 (1), 25-47.

Kraatz, Matthew S. (1998), "Learning by Association? Interorganizational Networks and Adaptation to Environmental Change," *Academy of Management Journal*, 41(6), 621–43.

Krackhardt, David (1992), "The Strength of Strong Ties: The Importance of Philos in Organizations," in *Networks and Organizations: Structure, Form, and Action*, Boston, MA: Harvard Business School Press, 216–39.

Lakshmanan, Arun, Charles D. Lindsey, and H. Shanker Krishnan (2010), “Practice Makes Perfect? When Does Massed Learning Improve Product Usage Proficiency?,” *Journal of Consumer Research*, 37(4), 599–613.

Lee, Thomas Y, and Eric T BradLow (2011), “Automated Marketing Research Using Online Customer Reviews,” *Journal of Marketing Research*, 48(5), 881–94.

Leetaru , Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook (2013), “Mapping the Global Twitter Heartbeat: The Geography of Twitter, ” First Monday, April 2013.

Levin, Daniel Z., and Rob Cross (2004), “The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer,” *Management Science*, 50(11), 1477–90.

Li, Xinxin, and Lorin M. Hitt (2008), “Self-Selection and Information Role of Online Product Reviews,” *Information Systems Research*, 19(4), 456–74.

Libai, Barak; Eitan Muller; and Renana Peres (2013), "Decomposing the Value of Word of Mouth Seeding Programs: Acceleration Vs. Expansion," *Journal of Marketing Research*, 50(2), 161-176.

Lin, Nan, Walter M. Ensel, and John C. Vaughn (1981), “Social Resources and Strength of Ties: Structural Factors in Occupational Status Attainment,” *American Sociological Review*, 46(4), 393–405.

Liu, Yong (2006), “Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue,” *Journal of Marketing*, 70(3), 74–89.

Luckerson, Victor (2012), “When Colleges Woo Students Through Social Media: Less Viewbooks, More Facebook,” Time. Accessed January 8, 2013, Available at: http://nation.time.com/2012/11/16/when-colleges-woo-students-through-social-media-less-viewbooks-more-facebook/

Ma, Liye, Ramayya Krishnan and Alan Montgomery, (2014), “Latent Homophily or Social Influence? An Empirical Analysis of Purchase within a Social Network,” *Management Science*, Forthcoming.

Manchanda, Puneet, Ying Xie, and Nara Youn (2008), “The Role of Targeted Communication and Contagion in Product Adoption,” *Marketing Science*, 27(6), 961–76.

Manski, Charles F. (1993), “Identification of Endogenous Social Effects: The Reflection Problem,” *The Review of Economic Studies*, 60(3), 531–42.

McCulloch, Robert, and Peter E Rossi (1994), “An exact likelihood analysis of the multinomial probit model,” *Journal of Econometrics*, 64(1–2), 207–40.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (2001), “Birds of a Feather: Homophily in Social Networks,” *Annual Review of Sociology*, 27, 415–44.

Moe, Wendy W, and Michael Trusov (2011), “The Value of Social Dynamics in Online Product Ratings Forums,” *Journal of Marketing Research*, 48(3), 444–56.

Moe, Wendy W. and David A. Schweidel (2012), “Online Product Opinion: Incidence, Evaluation and Evolution,” *Marketing Science*, 31 (3), 372-386.

Morahan-Martin, Janet M (2004), “How internet users find, evaluate, and use online health information: a cross-cultural review,” *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 7(5), 497–510.

Myers, James H., and Thomas S. Robertson (1972), “Dimensions of Opinion Leadership,” *Journal of Marketing Research*, 9(1), 41–46.

Nair, Harikesh S, Puneet Manchanda, and Tulikaa Bhatia (2010), “Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders,” *Journal of Marketing Research*, 47(5), 883–95.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), “Mine Your Own Business: Market-Structure Surveillance Through Text Mining,” *Marketing Science*, 31(3), 521–43.

Newell, Allenl, Paul Rosenbloom, (1981), "Mechanisms of skill acquisition and the law of practice," John R Anderson, ed., *Cognitive skills and their acquisition*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Newman, Andrew (2011), “ Brands Now Direct Their Followers to Social Media,” New York Times. Accessed January 8, 2013, Available at: http://www.nytimes.com/2011/08/04/business/media/promoting-products-using-social-media-advertising.html

Nielson, Jakob, (2006), "Participation inequality: Encouraging more users to contribute," Accessed April 10, 2012, Available at: ,http://www.useit.com/alertbox/participation_inequality.html..

Nokes, Timothy J, and Stellan Ohlsson (2005), “Comparing Multiple Paths to Mastery: What is Learned?,” *Cognitive Science*, 29(5), 769–96.

Oestreicher-Singer, Gal, and Arun Sundararajan (2012), “The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets,” *Management Science*, 58(11), 1963-1981.

Petty, R.E., Unnava, R., & Strathman, A., (1991). Theories of attitude change. In T. S. Robertson & H. H. Kassarjian (Eds.), *Handbook of consumer behavior* (pp. 241-280). Englewood Cliffs, NJ: Prentice-Hall

Porter, Martin, (2006),” The English (Porter2) stemming algorithm”. Available at: http://snowball.tartarus.org/algorithms/english/stemmer.html.

Reagans, Ray, and Bill McEvily (2003), “Network Structure and Knowledge Transfer: The Effects of Cohesion and Range,” *Administrative Science Quarterly*, 48(2), 240–67.

Rossi, Peter E., Zvi Gilula, and Greg M. Allenby (2001), “Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach,” Journal of the American Statistical Association, 96(453), 20–31.

Rutz, Oliver J, Randolph E Bucklin, and Garrett P Sonnier (2012), “A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising,” *Journal of Marketing Research*, 49(3), 306–19.

Schlosser, Ann E. (2005), “Posting versus Lurking: Communicating in a Multiple Audience Context,” *Journal of Consumer Research*, 32(2), 260–65.

Schweidel, David A., Eric T. Bradlow and Patti Williams (2006), “A Feature-Based Approach to Assessing Advertisement Similarity,” *Journal of Marketing Research*, 43 (2), 237-243

Schweidel, David A. and Wendy W. Moe (2014), "Listening in on Social Media: A Joint Model of Sentiment and Venue Format Choice," Journal of Marketing Research, forthcoming.

Shaughnessy, Hayden (2013), “Who Are The Top 50 Social Media Power Influencers, 2013?,” Forbes. Accessed March 1, 2014, Available at: http://www.forbes.com/sites/haydnshaughnessy/2013/04/17/ who-are-the-top-50-social-media-power-influencers-2013/2/

Shriver, Scott K., Harikesh S. Nair, and Reto Hofstetter (2013), “Social Ties and User-Generated Content: Evidence from an Online Social Network,” *Management Science*, 59(6), 1425-1443.

Sinha, Rajiv K., and Murali Chandrashekaran (1992), “A Split Hazard Model for Analyzing the Diffusion of Innovations,” *Journal of Marketing Research*, 29(1), 116–27.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde (2002), “Bayesian measures of model complexity and fit,” *Journal of the Royal Statistical Society: Series B*, 64(4), 583–639.

Sun, Monic (2012), “How Does Variance of Product Ratings Matter?,” *Management Science*. 58(4), 696-707.

Tanner, Martin A., and Wing Hung Wong (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82(398), 528–40.

Taylor, Nick (2014), “Who`s the (social media) King of New York?,” PeerIndex. Accessed March 1, 2014, Available at: http://blog.peerindex.com/whos-social-media-king-new-york/

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

Toubia, Olivier and Andrew T. Stephen (2013), “Intrinsic Versus Image-Related Motivations in Social Media: Why Do People Contribute Content to Twitter?” *Marketing Science*, forthcoming.

Trusov, Michael, Anand V Bodapati, and Randolph E Bucklin (2010), “Determining Influential Users in Internet Social Networks.,” *Journal of Marketing Research (JMR)*, 47(4), 643–58.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84 (4), 327–352.

Uzzi, Brian (1997), “Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness,” *Administrative Science Quarterly*, 42(1), 35–67.

Van den Bulte, Christophe, and Yogesh V. Joshi (2007), “New Product Diffusion with Influentials and Imitators,” *Marketing Science*, 26(3), 400–421.

Watts, Duncan J., and Peter Sheridan Dodds (2007), “Influentials, Networks, and Public Opinion Formation,” *Journal of Consumer Research*, 34(4), 441–58.

Wojnicki, Andrea, and David Godes (2011), “Signaling Success: Strategically-Positive Word of Mouth,” *Working Paper*.

Wortham, Jenna (2012), “Campaigns Use Social Media to Lure Younger Voters,” New York Times. Accessed January 8, 2013, Available at: http://www.nytimes.com/2012/10/08/technology/campaigns-use-social-media-to-lure-younger-voters.html?_r=0

Wu, Fang, and Bernardo A. Huberman (2008), “How Public Opinion Forms,” *Internet and Network Economics*, 5385, 334–41.

Ying, Yuanping, Fred Feinberg, and Michel Wedel (2006), "Leveraging Missing Ratings to Improve Online Recommendation Systems," *Journal of Marketing Research*, 43(3), 355–65.

Zaman, Tauhid, Emily Fox, and Eric Bradlow (2013), "A Bayesian Approach for Predicting the Popularity of Tweets," Working Paper.

Zhang, Jie, Michel Wedel, and Rik Pieters (2009), "Sales Effects of Attention to Feature Advertisements: A Bayesian Mediation Analysis," *Journal of Marketing Research*, 46(5), 669–81.

Zhao, Yi, Sha Yang, Vishal Narayan, and Ying Zhao (2013), "Modeling Consumer Learning from Online Product Reviews," *Marketing Science*, 32(1), 153–69.