The Dissertation Committee for Zhan Shi
certifies that this is the approved version of the following dissertation:

Three Essays on Adoption in Social Networks

Committee:

Andrew Whinston, Supervisor

Jason Abrevaya

Takashi Hayashi

John Mote

Haiqing Xu

Three Essays on Adoption in Social Networks

by

Zhan Shi, B.A., B.S., M.S. Econ.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2012

Dedicated to my parents and my girlfriend Wenxing Liu.

Three Essays on Adoption in Social Networks

Publication No. _____

Zhan Shi, Ph.D.
The University of Texas at Austin, 2012

Supervisor: Andrew Whinston

In the fast growing online social networks, one of the most commonly observed phenomena is the diffusion of information contents, behaviors or products through network members' interactions. In this thesis, I study the diffusion phenomenon by examining the individual-level adoption decision, both theoretically and empirically. In the three essays, I study the effects of the strength of the interpersonal tie and the social network characteristics on a potential adopter's decision-making, and investigate the measurement of network members' influences.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent technological innovations on the Internet, in particular the flourishing social network websites and their mobile applications, have transformed social interactions among people. Services that allow users to connect and share with real-world friends (e.g., Facebook), to track and discuss events as they happen in real time (e.g., Twitter and Sina Weibo), and to interact with others based on physical locations or common interests (e.g., Foursquare and Pinterest) are becoming ubiquitous. Powered by them, online social interaction has reached a level with unprecedented breadth, frequency and pace. As people post their comments on new purchases, review newly opened restaurants, broadcast songs they are listening to and "check-in" attractions they are visiting and share photos, these activities are observed by their "friends" or "followers" in the network, who may later try out the same products, songs, and venues. Every once in awhile, some new topic or behavior becomes "trendy," adopted by a substantial proportion of the social network members.

The sheer size of online social networks and their continuing cooperation and integration with traditional business models have made them

increasingly important in our daily lives. Thus, to understand the social networks' impact on people's social and economic activities is an interesting and timely topic not just for practitioners who operate or want to better harness the power of them, but also for academic researchers who seek to gain insights about the social and economic value of these new technologies. Indeed, there has been a clear call in the literature to bridge the gap between the pure network theory (e.g., Jackson 2008) and the empirical work on how social networks shape behavior (Rauch 2010). In this thesis, I approach this question by examining three examples of human activity in the presence of different types of social network: the dissemination of information, the search for new products, and the propagation of influence. In all the three examples, I focus on the micro-level individuals' decision of adoption, which, perhaps, is one of the most commonly observed user activities in online social networks. Consider, for example,

- a user chooses whether or not to "follow" someone else (the diffusion of one's popularity);
- a user chooses whether or not to "like," comment on, or share an interesting post authored by a friend (the diffusion of informational content);
- a user chooses whether or not to join the discussion on certain trendy topic by adding a "hashtag" in his or her messages (the diffusion of public discussion);

- a user chooses whether or not to try out some new restaurant recommended by many friends (the diffusion of new behavior).

To study individual adoptions in social networks has its unique opportunities and challenges, some of which are shared by social science research of online networks in general. For example, while the phenomenon of user adoptions in networks is commonplace, as exemplified above, few of its instances involve any explicit financial reward or value transfer between the users. In fact, in these online processes, we typically observe absolutely no information about "price," which is perhaps the single most frequently used term in economics. Rather, in this environment, social and technological factors also play a very important role. Hence, it is by nature an area where economics intersects sociology and information systems. In terms of conducting statistical analysis, the datasets collected from the Internet usually contain the social network users' (binary) interpersonal relationships as the *only* observed individual attributes. Detailed personal information, such as the demographic or socioeconomic characteristics that are widely used in empirical economics research, is unavailable. This is a situation faced by many researchers in this area, partly because of the usually huge size of online social networks that makes it extremely hard to record detailed personal information and partly because of the privacy concerns. Yet, such datasets also have their advantages. For example, the data in most cases is machine recorded, which indicates that the data has fewer measurement errors and, more importantly, the

data is produced in a strictly defined technological environment. This could possibly give researchers more structure in building models for interpreting the data. Moreover, the interpersonal relationships are given explicitly, so it provides researchers opportunities (and in the meantime also challenges) to uncover implied personal characteristics embedded in the network graph, either by applying social network analysis techniques or more advanced machine learning toolsets.

The three chapters proceed as follows. I start by looking at the dissemination of information in a social broadcasting environment in Chapter Two. Recent years have seen a tremendous growth of social broadcasting technologies. They have greatly facilitated open access to information worldwide, not only by powering decentralized information production and consumption, but also by expediting information diffusion through social interactions like content sharing. I study users' voluntary information sharing in the context of Twitter, the predominant social broadcasting site, by simultaneously modeling both the technology and users' social exchange on top of it. I collect a detailed dataset about the information-sharing activity on Twitter, called *retweet*, and document the statistical relationships between the users' social network characteristics and their retweeting acts. I estimate a two-stage individual-decision model using the conditional Maximum Likelihood (ML) method. The empirical results convincingly support our hypothesis: Weak ties are more likely to engage in the social exchange process of content sharing. I find that after an

author posts a median quality (as defined in the sample) tweet, the likelihood that a unidirectional follower will retweet is 3.1% higher than the likelihood that a bidirectional follower will.

In the third chapter, I turn to the topic of how location-based social networks help consumers search for venues that meet their needs and tastes. Specifically, I study the micro structure of the endorsement effect of social network neighbors' check-ins on a potential new customer's decision of visiting a venue. The empirical analyses are conducted on a unique panel dataset in which I observe both the explicit interpersonal relationships and the sequential check-ins made by the users. The key result is that the (normalized) number of unique endorsements is a bad predictor of the likelihood of a new visit. I suggest that a more detailed relationship between each connected pair of individuals be considered, for example their "proximity" implied by the network structure. Drawing upon the literature in sociology and computer science, I show that weighting the influencers' endorsements by a parsimonious "proximity" measure can yield a better result. It thus means that an endorsement is expected to have a larger effect if it comes from a "closer" network neighbor. The finding indicates the location-based social networks facilitate people's search for *experience goods*, such as restaurants, by easing the observation learning for their users. Additionally, I find that repeated check-ins have a larger effect, resembling a word-of-mouth effect. In dealing with the endogeneity problem, I apply the machine learning technique

nonnegative matrix factorization to uncover agents' *latent features* from the network graph.

Building upon these results, in particular the result on strong and weak ties' distinct roles in different diffusion processes, I propose a network user-influence measure in the fourth chapter. To make our measure as widely applicable as possible, I require no individual (demographic or socioeconomic) characteristics be available. But I do assume an explicit topic-specific influence network graph is observed, as is the case in the preceding two chapters. I start from a probabilistic model of individual adoption, and then mathematically develop the measure, which has a clear network-level economic interpretation. I then show that our measure of influence admits the famous centrality measure PageRank as a special case. In numerical experiments, I show that our measure has a higher degree of freedom beyond the PageRank algorithm, and this freedom makes our measure a more powerful tool in capturing richer structure of the influence distribution among a population.

Chapter 2

Content Sharing in a Social Broadcasting Environment: Evidences from Twitter

2.1 Introduction

At 10:24 p.m. EST, May 1, 2011, one hour and eleven minutes before the formal announcement of Osama Bin Laden's death by U.S. President Barack Obama, the following message was posted on Twitter by Mr. Keith Urbahn,[1]

> *So I'm told by a reputable person they have killed Osama Bin Laden...*

The post quickly attracted attention and got forwarded by Mr. Urbahn's subscribers on Twitter, and within two minutes, there were already more than 300 reactions to it. In the following hour, tens of thousands more users in the Twitter world were passing this message, and the final number of people who got exposed to the information *before* the formal White House announcement was even higher.

[1] @keithurbahn, http://twitter.com/keithurbahn.

This example not only shows the sheer power of Twitter as a fast-growing *social medium*, but also demonstrates that, the emerging social media can beat even their mainstream competitors in terms of speed, flexibility, and reach, especially in tracking events as they unfold in real time.[2] The unique advantage of websites like Twitter in disseminating news comes from their distinctive technological infrastructure. Although Twitter and a number of other similar online services, such as Tumblr and Sina Weibo, are usually referred to as micro-blogging or social networking sites, these labels fail to capture their whole essence — that these websites each are simultaneously a broadcasting service and a social network. Like content from most traditional mass media, *user-generated content* on these sites is accessible by the public and is broadcasted through directed subscription. The subscription relationships, as the only kind of user relationship, constitute the accompanying social network. The coexistence of a broadcasting service and a social network makes the combination of facets easily distinguishable from each one's respective standalone peers. On the one hand, the broadcasting service differs from traditional mass media like TV or radio in its decentralized structure and its social ingredient; it represents the full spectrum of communications, from headline news to personal and private communications (Wu et al. 2011). On the other hand, the social network, derived from content-subscription re-

[2]Indeed, this capability has been proven again and again during events such as the 2009 Iran election, the 2011 Middle East Revolution, and the 2012 Chinese political scandal.

lationships, also significantly differs from traditional online social networks, which typically map real-world friendships or connections. For example, the social network on Twitter is quite open and loose compared to the social network on Facebook because the follower–following relationship on Twitter can be established unilaterally and usually cuts across long (real-world) social distances. This combination gives these technologies unique advantages in facilitating information diffusion and justifies assigning them to a new category, which we call *social broadcasting networks*. This view is also explicitly or implicitly shared by many computer and information scientists. For example, Kwak et al. (2010) suggested that Twitter more closely resembles an information sharing site than a traditional social network. Bakshy et al. (2011) noted that "unlike other user-declared networks, Twitter is expressly devoted to disseminating information." Social broadcasting networks have blurred the traditional boundary between social networks and news media by adding the "social" ingredient into the cycle of information production, exchange, and consumption (Kwak et al. 2010, Wu et al. 2011, Socialflow 2011).

As exemplified by the Bin-Laden case, information diffusion in social broadcasting networks critically relies on social interactions, such as content sharing. Indeed, without the voluntary relaying of Mr. Urbahn's message by numerous Twitter users, that single post might never have triggered an avalanche of reactions and reached an audience far beyond

Mr. Urbahn's own subscribers.[3] Content sharing is a critical mechanism of information diffusion in social broadcasting networks and is vital to a network's proper functioning and thriving. When interesting or important information does not get passed on, the social broadcasting network fails to reach its full potential as a news medium; meanwhile, excess transmission of redundant or trivial information creates information overload and lowers the value of a social broadcasting network to the users. Understanding the information relaying process is thus both interesting and important. The objective of this chapter is to make an early step in this direction by examing the sharing decision-making process at the individual level. As suggested, one defining feature of social broadcasting networks is that they possess a large volume of weak interpersonal relationships. Thus, our central goal in this chapter is to address the following research question:

Research Question How does the strength of the interpersonal tie moderate people's voluntary content sharing behavior in a social broadcasting network?

Exploring the question might further reveal people's motivation in passing on information.

[3] According to social media company SocialFlow, Keith Urbahn wasn't the first to speculate Bin Laden's death after the news was released about the presidential address. However, Keith Urbahn's tweet proved to be a watershed in people's discussion on Twitter regarding the presidential address.

Users' voluntary content sharing is a social exchange process (Blau 1964) that involves the content's creator, the sharer, and the sharer's subscribers. To develop and test a theoretical model explaining how tie strength moderates people's decisions to engage in the social exchange, we draw on two streams of prior research: the literature on tie strength and the literature on people's pro-social behavior.

Plenty of literature has looked at the implications of tie strength in a variety of social or economic settings. For example, Granovetter (1973) did the pioneering work on the role that weak ties played when people search for jobs, the result of which is famously summarized as *the strength of weak ties* (SWT). The arguments of SWT suggest the importance of weak ties (i.e., ties with acquaintances, rather than close friends) in enabling novel information to flow across two densely knit groups of close friends. Levin and Cross (2004) proposed and tested a model of dyadic knowledge exchange taking into account trust and tie strength between the two parties. Their results also suggested that weak ties provide access to nonredundant information. Bapna, Gupta, Rice, and Sundararajan (2012) studied the link between strength of social ties and trust in an online social network using data from a Facebook application. They found that for the average user social tie strength as measured by actively interacting with someone else is positively linked to trust.

Researchers have also extensively studied people's motivation of sharing knowledge in online environment where explicit financial com-

pensation is often absent (Wasko and Faraj 2005, Bock et al. 2005, Chiu et al. 2006, Olivera et al. 2008). However, most of the previous studies focus on sharing behavior in the form of helping others (often strangers) solve problems by contributing one's own knowledge. Bock et al. (2005) surveyed 154 managers from 27 Korean organizations and found that anticipated reciprocal relationships affect individual's attitudes toward knowledge sharing. Chiu et al. (2006) also found that social interaction ties, reciprocity, and identification increased individuals' quantity of knowledge sharing by surveying 310 members of one professional virtual community in Taiwan. Olivera et al. (2008) developed a framework for understanding contribution behaviors and delineated three mediating mechanisms : awareness, searching and matching, and formulation and delivery. The sharing behavior we study is people's voluntary information relaying decision, which is a quite different type of contribution. Wasko and Faraj (2005) applied theories of collective action to examine how individual motivations and social capital influence knowledge contribution in electronic networks. Using survey data and archival data from one electronic network supporting a professional legal association, they found that people contribute their knowledge when they perceive that it enhances their professional reputations, when they have the experience to share, and when they are structurally embedded in the network. The current chapter can be viewed as an extension of Wasko and Faraj (2005) in the sense that we are also examing people's contribution behavior on

a electronic network. However, this chapter departs from previous literature in two important ways. In terms of data and method, we use micro-level data and a two-stage discrete choice model to study a relatively new form of sharing behavior–relaying information contributed by others–on a social broadcasting network which is also a new form of virtual community. In terms of theory, we integrate SWT with the general framework of social exchange to develop a new theoretical model to examine the relationship between network characteristics and retweeting behavior.

Our theoretical model posits that one's motivation for engaging in the social exchange process of content sharing is the latent benefit of perceived reputation enhancement resulting from consumption of the shared content by one's subscribers. The majority part of the latent benefit comes from the subscribers and thus is positively associated with the perceived novelty of the content to the sharer's subscribers, which in turn is negatively associated with the strength of the social tie between the content's creator and the sharer. Empirically testing our theory in a real-world social broadcasting network is complicated both by the challenge of collecting micro-level data from the Internet and by the specifics of the actual technological environment in which data are produced. To overcome these problems, we deploy 20 servers over a 140-day period to collect a detailed dataset containing information on both the content-sharing activity and social relationships from Twitter, and we develop a two-stage "consumption-sharing" model to help us better understand the

machine-mediated human decision-making process. We then estimate the empirical model using conditional Maximum Likelihood Estimation (MLE) method, the results of which convincingly support our theory.

The remainder of this chapter proceeds as follows. In Section 2.2, we briefly introduce Twitter as an example of social broadcasting networks and describe the technology-mediated information-sharing mechanism on Twitter. Drawing on social and behavioral theories, we develop our hypothesis in Section 2.3. After describing our dataset in Section 2.4, we conduct a series of empirical analyses to test our model in Section 2.5, and we discuss the managerial implications of our findings in Section 2.6. Finally, we conclude and discuss future research directions in Section 2.7.

2.2 Twitter and Retweeting

Designed to be the "Short Message Service of the Internet" at start-up, Twitter was launched in July 2006. During the 2007 South by Southwest (SxSW) festival in Austin, TX, a showcase of Twitter impressed the highly tech-savvy attendees. Since then, Twitter has entered a phase of rapid growth and gained popularity far beyond the technology industry insiders. As of March 2011, Twitter had more than 200 million registered users worldwide, who in total post an average of 150 million updates a day.[4] Twitter is now one of the most vibrant online communities in the

[4]See http://blog.twitter.com/2011/03/numbers.html and http://en.wikipedia.org/wiki/Twitter for more statistics.

world.

Twitter: A Social Broadcasting Technology

Twitter is an example of a social broadcasting site, where a broadcasting service and a social network organically constitute the technological infrastructure. On top of that, Twitter users produce and consume informational content by authoring and reading *tweets*,[5] which are text-based updates/messages of up to 140 characters. Like content on most traditional mass media, tweets are by default open to the public, and there is no restriction on consumption. Powered by its service, every Twitter user can be a content broadcaster and/or a content consumer.

Twitter users are networked to each other through a *following-follower* relationship. A user's *followers* are those who subscribe to receive his or her tweets, and a user's *followings* are the users whose tweets he or she subscribes to receive.[6] This following-follower relationship is the sole interpersonal link in the Twitter network. It is not only the pathway through which broadcasted content traverses the Twittersphere but also the channel of person-to-person communications, such as public reply and direct message. This relationship differs from *friendship* on Face-

[5] Tweet can also be used as a verb, meaning to post. So "to tweet a tweet" means "to post an update."

[6] A user *A* does not have to follow *B* to consume *B*'s tweets. *A* can access *B*'s Twitter webpage at any time to consume *B*'s tweets, which, like everyone else's, are always publicly available. But if *A* follows *B*, *B*'s tweets will be "pushed" to *A* in real time.

book or some other social network site in two respects: (1) the following-follower relationship on Twitter is relatively open in the sense that A following B does not require B's consent, and they usually do not map to real-world friendships as the ones on Facebook do;[7] and (2) perhaps more importantly, the following-follower relationship is directed (A's following B does not imply B's following A) while friendship is undirected (A's being a friend of B implies B's being a friend of A). The existence of a large volume of (loose and directed) subscription relationships is thus a distinctive characteristic of a social broadcasting network.

Retweeting: Content Sharing on Twitter

Content sharing is an integral part of the Twitter experience. In addition to composing and posting tweets themselves, Twitter users can also rebroadcast — or *retweet*[8] in Twitter's terminology — other users' (most likely their followings') tweets that they find are of particular (informational, entertaining, etc) value.[9] Retweeting spreads information by exposing new audience to the content. Meanwhile, retweeting is a special kind of sharing because a retweet is simply a copy of the original

[7]The fact that users who are connected in a social broadcasting site are usually neither friends nor even acquaintances in the real world allows us to narrow our focus just to the online context in studying their interactions. For instance, we do not have to worry that a favor A does for B online would be reciprocated offline.

[8]Retweet is both a verb and a noun, just as tweet is. When user A retweets a tweet t, we call the reposted copy of t a retweet and call A a retweeter of t.

[9]Posting others' tweets simply by copying and pasting their tweets without mentioning the original author is technologically possible but is not considered retweeting. Rather, it is a highly criticized misbehavior in the Twitter community.

tweet, and thus the author, content, and format of the shared information stay exactly the same as the original tweet. Retweeting can also display a "chain effect": not only a tweet's author's followers, but also sharers' followers, and so on, can further retweet, spreading the content onto their respective networks and amplifying the audience of the content to a potentially massive scale (Socialflow 2011). Thus, retweeting is evidently a critical mechanism of information diffusion on Twitter. Since it was introduced, retweeting has been extremely popular on Twitter because of the straightforward idea and the easy-to-use official retweet button.[10] Therefore, we use retweeting in the Twittersphere as the primary real-world example of content sharing activity.[11]

The mechanism of retweeting is graphically illustrated in Figure 2.1. Hereafter, we call the user who writes the original tweet the *author*, and the *author* is denoted *R* in the figure. The other nodes represent other users who are linked to each other via the following-follower relationship, together forming a tiny community inside the Twitter world. If two users mutually follow each other, the edge between them is drawn in solid (e.g.,

[10]The official retweet function is built into most mobile applications, as well as Twitter's official website. There is no publicly available statistic on the popularity of retweeting vs. other ways of information sharing. For example, another widely adopted way is to quote a tweet and add "RT" in front. An off-the-record interview with a Twitter employee confirmed that the official retweeting button had been the more popular mode of sharing.

[11]In addition to Twitter's dominance in the social broadcasting domain, another important reason we focus on it is that the openness of Twitter allows us to collect a detailed, micro-level dataset to complete our study. Section 2.4 describes our data collection in detail.

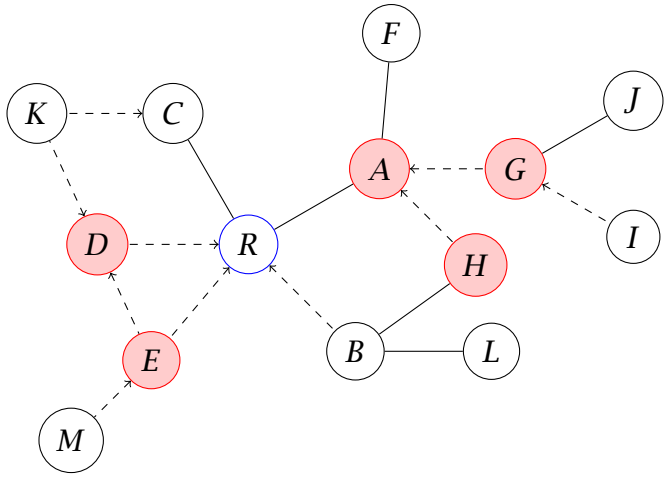

Figure 2.1: An Illustration of Retweeting

R and A, and we call A a bidirectional follower of R). Otherwise, if only one of them follows the other, the edge between them is a dashed line, with an arrow pointing to the user followed (e.g., B follows R but R doesn't follow B, so that we call B a unidirectional follower of R). After R posts an update, if no one retweets it, only R's followers A, B, C, D, and E would receive it. But now assume that after reading the message, users A, D, and E retweet (retweeters are shown in filled circles), thereby making F, G, H, K, and M, who are not immediate followers of R, receive a copy of the tweet. Then the new receivers could also retweet (as G and H do in the Figure 2.1 example), circulating the information more broadly around the network. One thing to note is that a retweet is also a content broadcast; because of the technology, a sharer cannot select a subgroup of his or her

followers and only retweet to this subgroup.[12]

Using the graphic example in Figure 2.1 as the context, we emphasize a few things related to our research question. First, we do not consider network dynamics (the formation and destruction of personal relationships among the users). In this research, we take a snapshot of the network structure, consider it as fixed and exogenous, and study user behavior on top of it. Second, in later econometric analyses, we model potential retweeters only in the first order (i.e., R's immediate followers A, B, C, D, and E), but not those in the second and higher orders (i.e., F, G, H, I, J, and K). As we explain in the data section, the reason is that we do not have the network graph data for higher order potential sharers. Third, the variation of user behavior we exploit is different users' different reactions to a single tweet (e.g., A, B, C, D, and E's reactions to a tweet authored by R), rather than one single user's different reactions to different tweets (e.g., B's reactions to different tweets authored by R, H, and L).

2.3 Theoretical Model

In this section, we develop the hypothesis on how the strength of the interpersonal tie moderates people's decision of relaying others' message. Although we often refer to Twitter as we develop our hypothesis, our theoretical arguments are applicable to other social broadcasting net-

[12] In *non-broadcasting* social networks, such as Facebook, users typically can post messages only to a chosen subgroup of his or her "friends."

works as well.

Content sharing is a social exchange process (Homans 1958; Blau 1964) that involves three parties: the sharer, the content's creator, and the group of individuals to whom the content is shared. By choosing to relay the information, the sharer incurs the cost of sharing[13] without being rewarded in any explicit way. However, the other two parties explicitly benefit: The subscribers can consume the shared information, and the content's creator reaches a larger audience.

Social exchange theory posits that people engage in social exchange in expectation of getting returns. When no explicit material or financial gains are received, the latent benefit of a social exchange process can be emotional comforts or social rewards (e.g., reputation). Indeed, "people's positive sentiments toward and evaluations of others, such as affection, approval, and respect, are rewards worth a price that enter into exchange transactions" (Blau 1964, p 112). Certain acts conducted by members of a community, such as sharing knowledge, benefit the collective but do not generate any immediate financial returns to the actors. Such behaviors are often referred to as "pro-social," because social rewards have been identified as an important incentive. For example, perceived reputation enhancement is identified as an important factor in motivating sharing in the information system and management literature (Wasko and Faraj

[13]The cost could be interpreted as the opportunity cost of choosing not to share.

2005).

These early research works suggest that the latent benefit for the sharer to engage in the social exchange process might come from the perception that participation in sharing information enhances his or her reputation either as a connected person in the network or as a person that has the capability to filter large amounts of content and dig out valuable pieces.

How large the latent benefit can be, or the extent to which a user's reputation can be enhanced by sharing a message, is determined by two factors: the number of subscribers who would receive the shared content and the extent to which the subscribers value that piece of content. The subscribers' valuation depends partly on the intrinsic quality of the shared information: The higher the quality is, the more the audience values the content, and hence the greater the latent benefit of sharing.[14] Moreover, different audiences' valuations of the same content (quality) should also differ because they have different preferences and different knowledge sets. For instance, the early tweet about the death of Osama bin Laden should indeed have high informational value to most ordinary Twitter users. However, for anyone inside the White House Situation Room on May 1, 2011, that tweet simply repeated a story he or she already knew and thus was of little additional value. This case shows that

[14]Because of this quality effect, we cluster our observations based on each tweet in our analysis.

information consumers with different backgrounds could attach unequal value to the same piece of content, and, in particular, the novelty of information should affect a particular consumer's valuation.

Earlier works in sociology studied the importance of weak ties in enabling the flow of novel information in a social structure. For example, Granovetter (1973) theorized the relationship between the novelty of information and the strength of the social tie through which the information is transmitted in the context of people finding jobs. Granovetter's results suggested that weak ties — those personal connections linking distant acquaintances — were more likely to provide nonredundant information because strong ties link closely related persons, such as family and friends, who often possess knowledge sets similar to the job seeker's. Following Granovetter's seminal work, subsequent research further demonstrated that, in both real organizations and virtual communities, weak ties are instrumental in connecting diverse groups and enabling a person to access heterogeneous and thus more valuable opinions (see, e.g., Granovetter 1982; Constant et al. 1996; Hansen 1999; Levin and Cross 2004). Adopting this view in the context of information sharing in a social broadcasting environment, we hypothesize that the strength of the social tie between the content creator and a potential sharer mediates the sharer's latent benefit of sharing. Specifically, *on average*, the weaker the tie is, the higher a potential sharer believes the subscribers would value the information and hence the higher the expected reputation enhancement is. The implica-

tion of this line of argument is the following hypothesized relationship between content-sharing probability and tie strength.

Hypothesis 1. *In social broadcasting networks, the latent benefit of sharing content is negatively associated with the strength of the social tie between a potential sharer and the content creator. Thus, given a piece of content, a weak-tie subscriber is more likely to share than a strong-tie subscriber, everything else being equal.*

This hypothesis might look counter-intuitive at first glance for readers who anticipate that, for example in the Twitter world, a Twitter user is more likely to retweet tweets from those who are strongly tied to her.[15] However, as we argued, information sharing in a social broadcasting environment is mainly a social exchange with one's followers. SWT suggests that the followers of a weak-tie follower of the content's creator should on average attach a higher value to the content, which, we argue, serves as a larger incentive for participating in the social exchange of forwarding information. Moreover, although our hypothesis is consistent with SWT, it is not a simple repetition of it. SWT states only that information obtained from one's weak-tie connections is expected to be more valuable; it

[15] Such intuition might have its root in the balance theory in psychology (Heider 1958). Blau (1964, p26) argued that a strain toward imbalance, as well as toward reciprocity, arises in social associations. If we think of the action of retweeting as an endorsement or a favor to the content creator, then a user's retweeting a tweet from someone who does not follow that user represents a greater imbalance than if that tweet were from someone who follows that user. In other words, from the perspective of the social exchange between the sharer and the content creator, a strong tie entails a stronger sense of obligation.

does not say that weak ties actually promote information dissemination in anticipation of the higher value from the information receivers. In this sense, our hypothesis extends the original SWT findings within the social exchange theoretical framework by arguing that in social broadcasting networks, weak ties, in expectation of higher social exchange returns, are more likely to provide the path by which information is relayed. We quote the following paragraph from Friedkin (1980):

> Granovetter's theory, to the extent that it is a powerful theory, rests on the assumption that local bridges and weak ties not only represent opportunities for the occurrence of cohesive phenomena ... but that they actually do promote the occurrence of these phenomena. A major empirical effort in the field of social network analysis will be required to support this aspect of Granovetter's theoretical approach ... It is one thing to argue that when information travels by means of these ties it is usually novel, and perhaps, important information to the groups concerned. It is another thing to argue that local bridges and weak ties promote the regular flow of novel and important information in differentiated structures. One may agree with the former and disagree with the latter.

Our hypothesis suggests that the two things Friedkin tried to disentangle conceptually might after all be indistinguishable practically because people's quest for reputation enhancement motivates them to facilitate the

penetration of novel information into the social network through weak ties.

User relationships in the Twitter environment are apparently not exactly the same as the real-world personal relationships Granovetter initially focused on to study the strength of weak ties. Hence, to adapt our hypothesis in the Twitter world and test it with data, we need to empirically operationalize the strength of social ties in the Twitter network. We do this based on the observed relationship types and assume that reciprocal relationships are *on average* stronger than nonreciprocal ones. This assumption leads to the following assumption, which is key to our subsequent empirical analysis:

Assumption 1. *A unidirectional link between two Twitter users is expected to be weaker than a bidirectional one, in the sense of "tie strength" established by Granovetter (1973).*

For instance in the Figure 2.1 example, ties like *D*-*R* are expected to be weaker than those like *C*-*R*.

Our measure of tie strength looks natural, but it nonetheless needs to be supported by convincing theoretical arguments and empirical evidence. We provide the supporting argument of our assumption in Appendix I for interested readers. Meanwhile, we note here that the emphasis on reciprocity is consistent with a long tradition in the sociology literature. Davis (1970) suggests that mutual choices indicate a strong tie

while asymmetric pairs indicate weak ties.[16] Granovetter also pointed out that the strength of a tie is a combination of several factors, including mutual confiding and reciprocal services (Granovetter 1973). Friedkin (1980) measured tie strength among faculty members in seven biological science departments of a single university based on whether a discussion about current research is reciprocated or not reciprocated.

Based on the assumption, our hypothesis, adapted in the Twitter world, becomes an empirically testable one:

Hypothesis 2. *On expectation, a unidirectional follower is more likely to retweet than a bidirectional follower.*

For instance, in Figure 2.1, *ex ante* we expect D is more likely to retweet R's tweet than C is. We develop our econometric model based on both these theoretical discussions and the technological specifics of the Twitter environment. Before discussing the model, we describe our data in Section 2.4.

[16] Davis measured interpersonal relations on a three-point ordinal scale: mutual positives are the most positive, mutual negatives are least positive, and asymmetric pairs are intermediate. In sociometry, these correspond to mutual choices (i chooses j and j chooses i), mutual nonchoices (i does not choose j, and j does not choose i), and unreciprocated (i chooses j but j does not choose i, or j chooses i but i does not choose j).

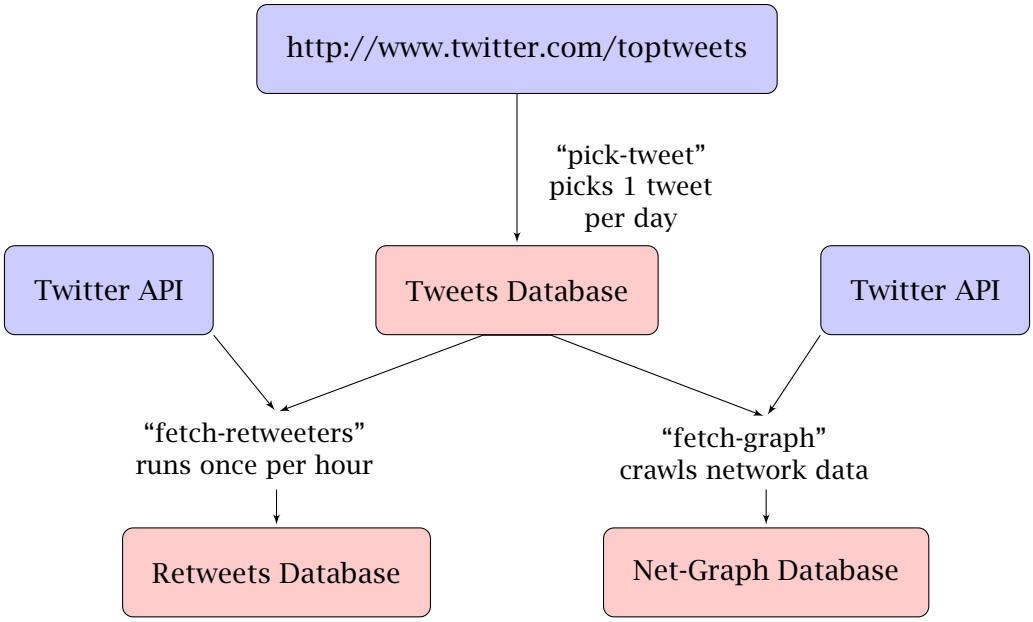

Figure 2.2: Data Collection Workflow

2.4 Data

We deployed 20 servers to collect data by querying Twitter's application programming interface (API).[17]

Data Collection

Figure 2.2 shows the data collection workflow and is a useful illustration for helping readers to understand the details of our data collection process, described in the following paragraphs. From July 22, 2010 to December 2, 2010, at 0:05 each day, our "pick-tweet" program fetched Twitter's *toptweets* webpage, which usually showed 17 to 18 popular tweets in

[17] http://dev.twitter.com

the Twittersphere at the visiting time.[18] Sorting these tweets into chronological order, our program then checked, one by one, the number of followers a tweet's author had and inserted into our tweets database the first one it found whose author had less than 1,500 followers; the rest were discarded. If all the authors had more than 1,500 followers, the program wouldn't insert any tweet on that day. In other words, our program picked either 1 tweet or 0 tweets every day over this period of time.[19]

After a tweet entered our tweets database, another "fetch-retweeters" program began to track and fetch its retweeting data and would do so constantly during the subsequent five days.[20] At 10 minutes past each clock hour over the 5 days, the program queried Twitter API to get the user IDs of the retweeters (those in filled circles in the Figure 2.1 example). The retweeter IDs were obtained in the order of the time at which the user

[18] *Top Tweets* is an official Twitter account, which is a "new algorithm that finds tweets that are catching the attention of other users." The algorithm is proprietary, so we cannot give a definition for a "popular tweet." Twitter's Chief Scientist, Abdur Chowdhury, explained, "the algorithm looks at all kinds of interactions with tweets, including retweets, favorites, and more to identify the tweets with the highest velocity beyond expectations."

[19] The "pick-tweet" program did not run properly on a few days during our data collection period because technical problems (e.g., server failure) occurred on either the Twitter side or our side. On those days, no tweets were added to our database.

[20] The decision to track retweeting activities for five days was made on the basis of our judgment about how long a retweeting process of one tweet could stay active. The log file written by the "fetch-retweeters" program showed that most retweeting activities of a tweet happened within just one or two days of when it was first posted. Tracking for five days thus seemed conservative enough to ensure that any truncated sample problem (a large number of retweets occurring after our tracking period) was unlikely.

retweeted.[21]

As retweeting data came in, another "fetch-graph" program worked on collecting relevant network graph information. Specifically, for each tweet, we were interested in its author (R in Figure 2.1), the author's followers (A, B, C, D, E in Figure 2.1), and the tweet's (first-order) retweeters (A, D, E in Figure 2.1); we called this set of Twitter users our focal set. For each user in the focal set, our program collected the IDs of both the followings and followers and stored the data in our network graph database. For some users in the focal set, access to their following-IDs and follower-IDs was restricted because they explicitly disallowed third-party access to their data. We used a "protected" flag to indicate this privacy protection status, with the flag = 1 meaning no public data access. With the retweeting data and network graph data in hand, we produced a real-world analog of Figure 2.1 (see Figure B.1 in Appendix II). The figure shows the spread of the first tweet in our database.

We designed our data collection strategy around one important binding constraint: Twitter API allowed only 150 visits/queries per IP per hour,[22] and our computational capacity was limited. One API visit would

[21] One important technical constraint was that Twitter API provided IDs for only the 800 most recent retweeters, so that if more than 800 users retweeted a tweet between two queries, our program was not able to get the complete set of retweeters. In addition, we found no publicly available way to verify the number of retweeters our program had missed. We took a conservative approach to deal with this situation: Unless we were sure we had fetched the complete set of retweeters for a tweet, we discarded that tweet from our database.

[22] This REST API rate limit was as of the second half of 2010:

return only a limited amount of information, so to finish one "job" (e.g., getting the entire set of a user's following-IDs) could require a number of queries (e.g., the actual number of visits required would depend on the number of followings the user had). As discussed in the previous paragraph, we had to collect all following-IDs and follower-IDs for all users in the focal set; moreover, we had to finish collecting the data as quickly as possible to avoid potential significant changes in their following-follower relationships. This 150-visits limit was the reason why we decided to select only one tweet per day, select only tweets whose authors had fewer than 1,500 followers, and track retweeting activity only once per hour, and why we decided *not* to collect network graph data of followers' followers (G in Figure 2.1).[23] Deciding otherwise would have prevented us from finishing the workload for one tweet before the next tweet came into our database.

Data Description and Statistics

We provide a list of notations in Table 2.1.

Tweets, authors, and the number of observations

By the end of the 140-day data collection course, we had successfully completed data collection for 65 tweets. We index the tweets in order

https://dev.twitter.com/docs/rate-limiting.

[23] As a result, we do not have the "second-order" retweeters' network characteristics and we do not include the "second-order" retweeters in later econometric analyses. Studying their retweeting decisions can be a future research topic.

Tweet level	t	index of tweets/authors
	n_t	the number of followers of author t
		the number of observations for tweet t
	v_t	the total number of retweeters of tweet t
Follower level	ti	index of author t's followers, $i \in \{1, 2, \ldots, n_t\}$
	y_{ti}	binary outcome, = 1 if follower ti retweeted tweet t
	w_{ti}	binary variable, = 1 if follower ti is a unidirectional follower of t (weak tie)
	V_{ti}	the number of ti's followings
	W_{ti}	the number of ti's followers
	m_{ti}	the number of times ti's followings retweeted tweet t (before ti did if $y_{ti} = 1$)

Table 2.1: Notations

of posting time by an integer, t, ranging from 1 to 65. The tweets were all authored by different users, so we also denote the author of tweet t author t, for simplicity of notation.[24]

The two plots in Figure 2.3 show the distributions of the tweets by month of post and by hour of post, respectively. The sample frequency of tweets by hour of post is roughly consistent with the distribution of total volume of tweets posted in each clock hour in the entire Twitter world. The left subplot of Figure 2.4 shows the distribution of the number of followers an author had (n_t) and the distribution of the total number of retweeters a tweet gained (v_t). Note that for a tweet, v_t could be larger than n_t because retweeters' followers who were not immediate followers

[24]Among the 65 tweets, 3 are in Spanish, 1 is in Italian, 1 is in Portugese, and the remaining are in English. None of the authors is celebrity, partly because of our 1,500-follower constraint. The textual contents range from breaking news and comments on news to political jokes and witty quotes.

n_t	min	max	$mean$	$median$
total	87	1497	457	370
non-protected	54	1189	375	324

Table 2.2: Number of Observations per Tweet

of the author could also have retweeted. The right subplot of Figure 2.4 is a scatter-plot of the 65 tweets on the n_t-v_t plane. More or less surprisingly, our sample shows no positive correlation between the number of followers an author had and the total number of retweeters her tweet gained (a linear fitting line shows weakly negative slope). However, this simple result is actually consistent with Bakshy et al. (2011), which also finds that the number of an author's followers is in general a poor predictor of the size of the retweet cascade.

Because our objective is to model a follower's binary decision of whether to retweet, n_t, the number of followers that author t had is also the number of observations in cluster t. From this place onward, we exclude users for whom we could not collect following/follower IDs (flag "protected" = 1) and users with zero following/followers (assuming they were either new registrants or inactive members). As a result, the total number of observations ($N = \sum_{t=1}^{65} n_t$) in our sample declined from 29,681 to 24,403, a decrease of 17.78%. Table 2.2 gives the basic descriptive statistics of n_t before and after dropping the observations, and Figure 2.5 shows the number of pre-dropping vs. post-dropping observations in more detail.

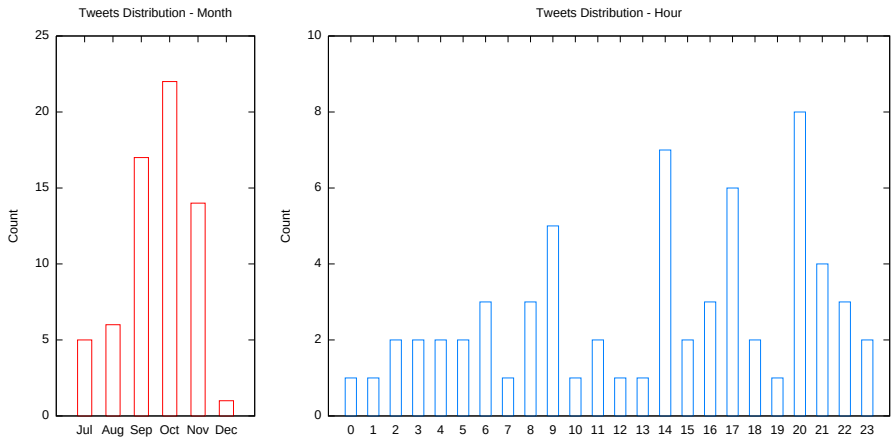

Figure 2.3: Distributions of Tweets by Month of Post and by Hour of Post

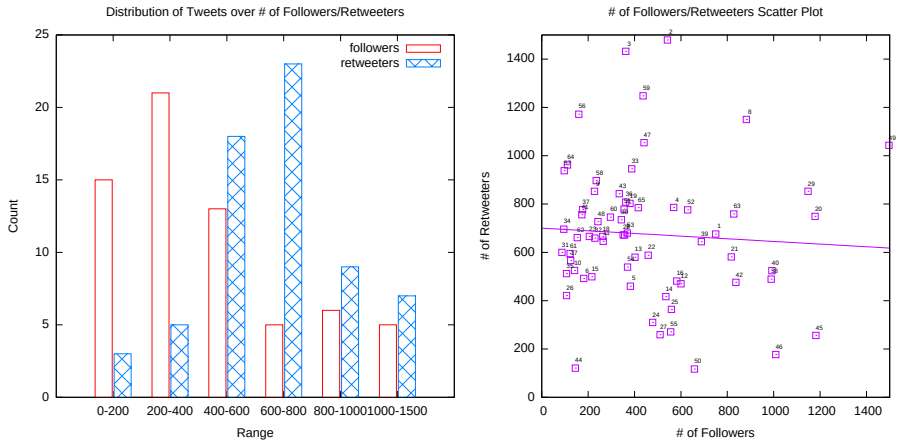

Figure 2.4: Distribution of Number of Author's Followers and Number of Retweeters

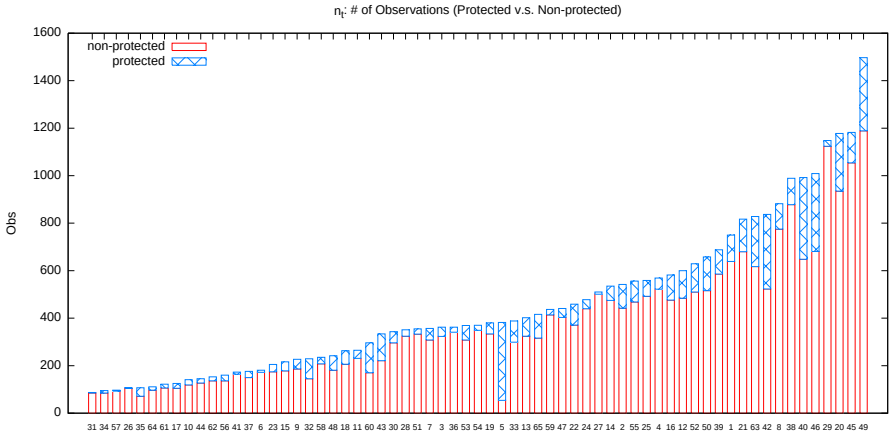

Figure 2.5: Number of Observations per Tweet

		$mean$	std	5%	15%	50%	85%	95%
y_{ti}	retweet dummy	0.0427	0.2022	-	-	-	-	-
w_{ti}	unid'l dummy	0.7598	0.4272	-	-	-	-	-
V_{ti}	# of followings	1574	9046	25	69	347	1714	3297
W_{ti}	# of followers	3304	73124	5	22	190	1117	4970
m_{ti}	# of repetition	3.2845	7.5216	1	1	1	4	11

Table 2.3: Descriptive Statistics

Variables

We now summarize the key variables used in the econometric model. For a tweet t, we use $y_{ti}, i \in \{1, 2, \ldots, n_t\}$ to index whether each of its observations (i.e., author t's followers) retweeted tweet t. The definitions of the key variables can be found in Table 2.1. These variables are either directly observed or constructed from observed ones. We provide the descriptive statistics of these variables in Table 2.3 and the correlations between them in Table 2.4.

Let $y_t = \sum_{i=1}^{n_t} y_{ti}$ be the number of retweeters among author t's

		y_{ti}	w_{ti}	V_{ti}	W_{ti}	m_{ti}
y_{ti}	retweet dummy	1.0000				
w_{ti}	unid'l dummy	0.0072	1.0000			
V_{ti}	# of followings	-0.0225	-0.0921	1.0000		
W_{ti}	# of followers	-0.0065	-0.0338	0.4436	1.0000	
m_{ti}	# of repetition	0.0493	-0.1508	0.2002	0.1400	1.0000

Table 2.4: Correlations

followers (note that $y_t \neq v_t$), and $yr_t = y_t/n_t$ could then be naturally interpreted as the retweeting rate of t. Figure 2.6 shows the retweeting rate across the tweets with a 95% error bar. That the rate varies quite a lot is not surprising given the significant heterogeneity across the tweets (i.e., the intrinsic quality). Hence, we should consider tweet-specific effects when modeling retweeting behavior. Over the whole sample (i.e., tweets pooled together), the retweeting rate is 0.0427, and the 95% confidence interval is $(0.0402, 0.0452)$.[25]

w_{ti} is the binary indicator of unidirectional relationship, which is also our main operationalization of a weak tie in the econometric analysis. The simple correlation of y_{ti} and w_{ti} is positive. $w_t = \sum_{i=1}^{n_t} w_{ti}$ is the number of author t's followers who were not followed back by t. $wr_t = w_t/n_t$ is thus the fraction of t's unidirectional followers. We plot wr_t in Figure 2.7, which shows that for most of the tweets in our sample, wr_t is in the range $(0.5, 0.9)$. Over the whole sample, the fraction is $wr =$

[25]Because we selected popular tweets, this retweeting rate does not generalize to the entire tweet space.

0.7598, and its 95% confidence interval is $(0.7545, 0.7652)$.[26]

Some basic descriptive statistics of the number of followings (V_{ti}) and the number of followers (W_{ti}) can be found in Table 2.3. The $median$ values of both V_{ti} and W_{ti} are much smaller than their respective $mean$ values, so both distributions are positively skewed and have long right tails (i.e., the majority of the users had tens or hundreds of followings and followers, but a handful of them might have had up to hundreds of thousands of followings or even millions of followers). Similar statistics can be found in Kwak et al. (2010) and Wu et al. (2011), but the $median$ numbers are much bigger in our study than in their articles because we exclude observations with zero followings/followers. The Pearson's correlation of V and W is 0.4436, as shown in Table 2.4, and both V and W are negatively correlated with y_{ti}.

m_{ti} is the number of times someone among ti's followings (re)tweeted t (including author t's original tweet). m_{ti} also has a heavily positively-skewed distribution: More than half of the observations received the tweet just once (i.e., none of their followings retweeted). Over the whole sample, the mean is equal to 3.28, and the standard deviation is equal to 7.53.

[26]We also compute the fraction of unidirectional links among all 110,583,366 relationships observed in our database (not only those between authors and their followers); the percentage is 75.2%, which is surprisingly close to wr. In other words, this finding says that, on average, one out of four edges in the Twitter world is bidirectional. Kwak et al. crawled the entire Twitter network in July 2009 and computed this rate to be 77.9%; thus we see more bidirectional links one year after their research. This increment might be an interesting metric for researchers who study network formation.

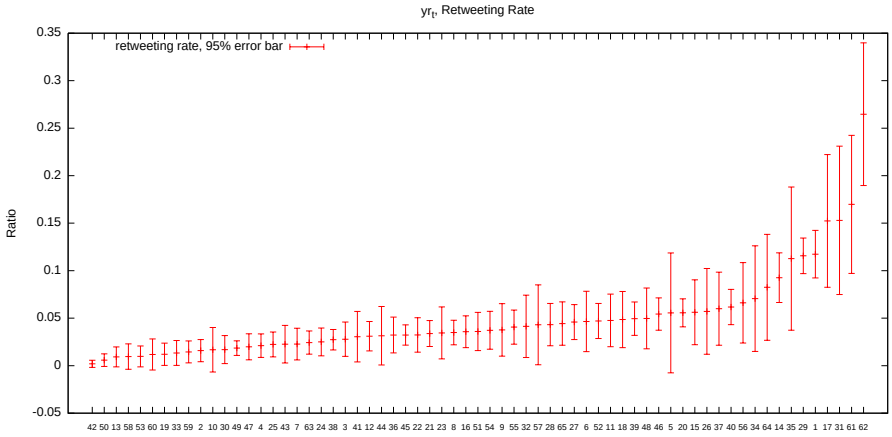

Figure 2.6: Retweeting Rate Across Tweets

We observe that m_{ti} is positively correlated with V_{ti}, the number of followings a user has, because m_{ti} is by definition the size of a subset of followings. m_{ti} is negatively correlated with w_{ti}, meaning bidirectional followers are likely to receive more retweets than unidirectional ones.

2.5 Empirical Model and Results

In this section we use our retweet dataset to perform empirical tests on our hypothesis. Instead of using standard reduced-form econometric methods for binary response (e.g., probit or logit), we take a more structural approach, modeling both the user behavior and special features of social broadcasting technology. We then use MLE technique to estimate the empirical model and present the results.

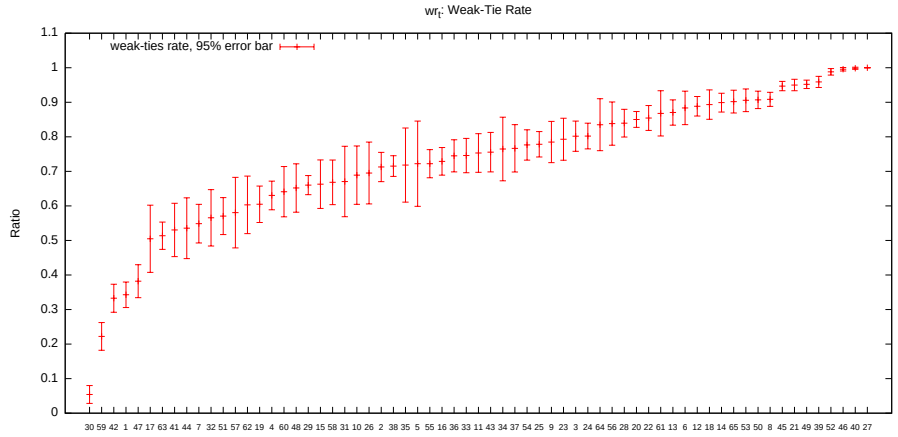

Figure 2.7: Weak-Tie Rate Across Tweets

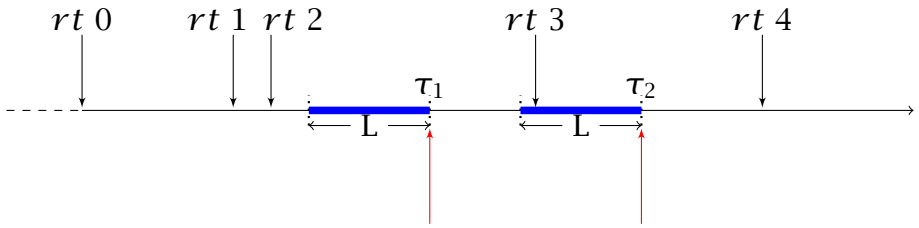

Figure 2.8: (Re)Tweets Entering a Twitter User's Timeline

2.5.1 Conditional MLE

We model a two-stage, consumption-retweeting process, in which consumption is the necessary first step for retweeting. We first describe the two stages and derive the likelihood function that would be used in our final conditional MLE analysis. We then show the results and discuss our findings.

Stage One: Consumption

The first stage models whether a follower of author t, say, ti, after receiving a tweet, actually consumes it. Figure 2.8 illustrates the technological aspect of this stage. The horizontal line stands for ti's *home timeline* or *Twitter feed*, which is a stream of received tweets for ti to consume (read), including retweets, listed in chronological order. Note that not only the original tweet t but also ti's followings' retweets of it, if any, appear in ti's timeline. The downward pointing arrows show the times at which a total of five (re)tweets of t enter the feed. Between these five (re)tweets, other tweets are also posted by ti's followings.

In reality, few Twitter users can or will monitor their Twitter feed continuously. We assume every time they start reading their feeds, they consume only a limited number of tweets. In the example shown in Figure 2.8, the upward pointing arrows indicate the times, τ_1 and τ_2, when user ti launches her Twitter application. Because the tweets are listed in chronological order, tweets posted at times close to the τs are more likely to be consumed. For simplicity, we use a thick horizontal segment to indicate a "period of attention" of length L, inside which tweets posted are consumed. In doing so, we implicitly assume that users do not discriminate between tweets authored by different people. The only factor determining whether a tweet catches the user's attention is whether it enters the timeline during a certain period preceding the time a user checks

tweets.[27]

Therefore, the cognitive limit restricts a user ti from reading every single tweet she receives. In Figure 2.8, tweets that enter into the timeline in the interval $(\tau_1, \tau_2 - L)$ are outside any of the periods of attention and would not be consumed by ti. When a tweet t gets retweeted by ti's followings, it enters the timeline multiple times, thus increasing the likelihood that t falls into one of the periods of attention (e.g., $rt3$ in the figure). If neither the original tweet t nor the retweets fall into some period of attention, then it is not consumed and hence would not be retweeted by ti.

Unfortunately, whether tweet t is actually consumed by ti is unobserved. Our task for this stage is to build a probabilistic model to capture the likelihood that ti consumes t, conditional on observed variables. Based on previous discussions about the technology, whether ti consumes tweet t is determined by three factors: (1) m_{ti}, the number of times t appears in ti's timeline; (2) the frequency with which ti checks her Twitter feed; and (3) L, which is determined by the number of tweets ti can read in each consumption and the number of tweets ti receives per unit of time, which we assume to be a linear function of V_{ti} (i.e., the more

[27] This "random-reading" modeling assumption is only a rough approximation of the real consumption stage. In reality, great variation exists in how people use Twitter and read their Twitter feed. However, because most people receive a large amount of tweets, of which they are "able" to consume only a portion, we believe that without detailed data on individual Twitter usage, "random-reading" is an appropriate modeling approximation for us to use.

people a user follows, on average, the more tweets she receives over a fixed time span). Therefore, we propose the condition for ti to consume t be the following equation:

$$\frac{m_{ti}}{bV_{ti}} > a_{ti}, \tag{2.1}$$

where b is a positive constant and $1/(bV_{ti})$ measures L.[28] The unobserved variable a_{ti} can be interpreted as an inverse measure of the frequency with which ti checks her Twitter feed, and is assumed to be independent of both V_{ti} and m_{ti}. The left side of equation (2.1) can be seen as the scaled frequency with which t appears in the timeline, and the right side as a user-specific threshold. If a user does not check her feed very often, so that she gets a high draw of a_{ti}, then the scaled frequency needs to be high for the tweet to be consumed, and vice versa. To derive the likelihood function, we further assume that a_{ti} is log-normally distributed in the population:

$$\log a_{ti}|t \sim \log a_{ti} \sim N(a, \sigma_a^2). \tag{2.2}$$

So we can rewrite equation (2.1) as

$$\begin{aligned} -\log b + \log m_{ti} - \log V_{ti} &> \log a_{ti} \\ -\frac{a + \log b}{\sigma_a} + \frac{1}{\sigma_a}\log m_{ti} - \frac{1}{\sigma_a}\log V_{ti} &> \frac{\log a_{ti} - a}{\sigma_a}, \end{aligned}$$

[28] Or more generally, we can assume $L = \frac{z_{ti}}{bV_{ti}}$, where z_{ti} is the number of tweets ti can read in each consumption and bV_{ti}, $b > 0$, is the number of tweets received by ti per unit of time. We can still get (2.1) by dividing both sides by z_{ti} and absorbing the unobserved z_{ti} into a_{ti}.

where the term on the right side is a standard normal distribution. Thus, the *ex ante* probability that ti consumes tweet t, conditional on receipt, is

$$\begin{aligned} p_1 &= \mathrm{p}(-\frac{a+\log b}{\sigma_a} + \frac{1}{\sigma_a}\log m_{ti} - \frac{1}{\sigma_a}\log V_{ti} > \frac{\log a_{ti} - a}{\sigma_a}) \\ &= \Phi(-\frac{a+\log b}{\sigma_a} + \frac{1}{\sigma_a}\log m_{ti} - \frac{1}{\sigma_a}\log V_{ti}), \end{aligned} \tag{2.3}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The outcome of this stage is unobserved, so we cannot estimate the parameters in which we are interested just on the basis of equation (2.3).

Stage Two: Retweeting

Recall that a follower ti retweets only if ti consumes the tweet himself. If a user's first stage outcome is a failure (he does not consume t), then his final outcome would automatically be *not retweeting*, $y_{ti} = 0$. In other words, $y_{ti} = 1$ implies success at both stages. Unlike the first stage, where success is determined by the broadcasting technology and chance, the second stage outcome depends on the decision made by the user.

At the second stage, the users who have consumed the tweets each decide whether to retweet. The decision is made on the basis of a subjective cost-benefit analysis. As discussed in Section 2.3, the latent benefit of retweeting depends on both the number of followers the content is retweeted to, W_{ti}, and the mean valuation the followers attach to the

tweet, which we denote α_{ti}. Thus, we write the latent benefit $\alpha_{ti}W_{ti}$. We expect ti's followers' mean valuation, α_{ti}, to be moderated by the strength of the social tie connecting author t and potential retweeter ti. Finally, for the retweeting act to happen, the latent benefit should exceed the user-specific reservation utility or cost, denoted c_{ti}. Therefore, after using logarithmic transformation, the necessary and sufficient condition of retweeting upon consumption can be written (with a slight abuse of the notation α and c):

$$\alpha_t + \delta w_{ti} + \beta \log W_{ti} > c_{ti}, \tag{2.4}$$

where c_{ti}, like a_{ti}, is unobserved, and α, sub-indexed by t, is allowed to differ across the tweets, capturing tweet-specific effect.[29]

Technically, we further assume c_{ti} is distributed normally among the population. We also allow the unobservables at the two stages to be correlated:

$$c_{ti}|t \sim c_{ti} \sim N(c, \sigma_c^2), \quad \text{Cor}(c_{ti}, a_{ti}) = \rho. \tag{2.5}$$

We can rewrite equation (2.4) as

$$-\frac{c}{\sigma_c} + \frac{\alpha_t}{\sigma_c} + \frac{\delta}{\sigma_c} w_{ti} + \frac{\beta}{\sigma_c} \log W_{ti} > \frac{c_{ti} - c}{\sigma_c},$$

where the right side is a standard normal distribution. Therefore, the

[29] α_t also includes the author-specific effect, since in our sample the tweets are all by different authors.

conditional probability of retweeting can be written as follows:

$$p_2 = \mathrm{p}\Big(-\frac{c}{\sigma_c} + \frac{\alpha_t}{\sigma_c} + \frac{\delta}{\sigma_c} w_{ti} + \frac{\beta}{\sigma_c} \log W_{ti} > \frac{c_{ti} - c}{\sigma_c} \Big| \\ - \frac{a + \log b}{\sigma_a} + \frac{1}{\sigma_a} \log m_{ti} - \frac{1}{\sigma_a} \log V_{ti} > \frac{\log a_{ti} - a}{\sigma_a} \Big). \tag{2.6}$$

Two-Stage Model For MLE

At this point, we put the two stages together. Equations (2.2), (2.3), (2.5), and (2.6) represent all the necessary elements for conducting the MLE analysis. The likelihood of observing outcome $y_{ti} = 1$ for tweet t and follower ti is the product of p_1 and p_2, and the likelihood of observing $y_{ti} = 0$ is $1 - \mathrm{p}(y_{ti} = 1)$. In terms of econometrics, not all the structural parameters are identified. For example, we can identify δ/σ_c, but not δ and σ_c separately. Fortunately, for our research purpose, we care most about the signs of the parameters rather than their absolute value. In the example, δ/σ_c has the same sign as δ; thus, identifying the ratio is good enough for understanding w's partial effect. Therefore, for simplicity of notation, we rearrange the terms, rescale the parameters following the standard practices in probit and logit models, and obtain our benchmark

specification:

$$\begin{aligned}
\mathrm{p}(y_{ti} = 1) &= p_1 p_2 \\
p_1 &= \mathrm{p}(e + b_1 \log m_{ti} + b_2 \log V_{ti} > a_{ti}) \\
p_2 &= \mathrm{p}(\alpha_t + \delta w_{ti} + \beta \log W_{ti} > c_{ti} | e + b_1 \log m_{ti} + b_2 \log V_{ti} > a_{ti}) \\
a_{ti}, c_{ti} &\sim N(0,1) \\
\mathrm{Cor}(a_{ti}, c_{ti}) &= \rho \\
\theta &= \{e, b_1, b_2, \alpha_1, \alpha_2, \ldots, \alpha_T, \delta, \beta, \rho\},
\end{aligned} \tag{2.7}$$

where θ is a vector of parameters to estimate. α_t — with t ranging from 1 to T — absorbs the constant term and captures the tweet-specific effects. δ is the coefficient of the weak-tie indicator, which is of our primary interest. b_1, b_2, and β determine the partial effects of the other social network characteristic variables.

Results

With equation (2.7) in hand, we estimate the parameters using the conditional MLE method. We report the results in Table 2.5.[30] We estimate a total of six different specifications, the first five of which are described in detail in the following paragraphs. The last one is discussed in the next subsection. In all specifications, we use dummy variables to capture tweet-specific effects,[31] α_ts, and we do not report these fixed effects be-

[30] *, **, and *** indicate 0.1%, 1% and 5% significance levels, respectively.

[31] Technically, we can directly use dummy variables to control for fixed effects without appealing to more sophisticated econometric specifications because we have a large number of observations for every tweet. See Figure 2.5.

cause they are less important in our analysis.[32] All standard errors are computed to be robust to tweet clustering.

Model 1 is a simple probit of y_{ti} on the four key variables: w_{ti}, m_{ti}, V_{ti}, and W_{ti}. Model 2 corresponds to equation (2.7), with an additional restriction that $a_{ti} \perp c_{ti}$, which implies $\rho = 0$. Model 3 strictly follows the benchmark equation (2.7), allowing correlation between a_{ti} and c_{ti}. Models 4 and 5 slightly modify model 3: Model 4 includes the interaction term of w_{ti} and W_{ti} in the retweeting equation; model 5 includes w_{ti} in the consumption equation.

We observe that the fitted likelihood increases from model 1, 2 to 3, $\{4, 5\}$, as we gradually relax the model restriction by adding richer structures and more variables. Across the five columns, we find consistent support for a positive m_{ti} coefficient (repetition of retweets) and a negative V_{ti} coefficient (the number of followings). All estimates are significant with 99.9% confidence level. Therefore, the results are consistent with the model prediction described in Section 2.5.1, and in particular with equation (2.3).

The unidirectional-relationship/weak-tie indicator is found to have a significantly positive effect on the (conditional) retweeting probability. In the benchmark model (model 3), its coefficient is positive at the 0.1%

[32] We do not control for follower fixed effects because, for each tweet, all followers/observations are by definition distinct, and when we pool tweets together, among all the 24,403 observations, 24,002 are unique.

significance level. The w_{ti} coefficient becomes less significant, but is still positive at the 5% significance level, when we allow an interaction effect of tie-strength and the number of followers (model 4) or when we put the weak-tie indicator into both the consumption and retweeting equations (model 5). These results show that, in the retweeting equation, the positive sign of the weak-tie coefficient is robust; thus, they support our hypothesis: Weak ties are more likely than strong ties to relay information to their social network neighbors.

In model 4, where we include the weak-tie dummy w_{ti} in both the consumption and retweeting equations, we find that, although its effect on retweeting probability is positive and significant, its effect on consumption probability is negative but insignificant. This result shows that messages generated from stronger ties *might* be more likely to be read than those from weaker ties. However, the difference in likelihood is not statistically significant. It supports our assumption that users generally do not discriminate between tweets received from strong ties and tweets received from weak ties. We believe the separation of the different effects that weak ties have on the two probabilities, as model 4 reveals, shows the merit of our two-stage econometric model. It indeed uncovers more structure in the retweeting process than a reduced-form probit regression.

In all models, the number of followers has a significantly positive coefficient. This revelation by our econometric models is a new one because, as shown in Table 2.4, the simple correlation between y_{ti} and W_{ti}

is negative. This result thus supports our argument in the theory section that the number of subscribers is positively associated with the latent benefit of retweeting.

2.5.2 Theoretical Model Revisited

From model 1 to model 5, we consistently find that, conditional on the consumption of a piece of information, weak-tie users are more likely to share information with their social network neighbors. In the theory section, we argued the reason is that a weak-tie follower's followers would *on average* value the information more than a strong-tie follower's followers; thus, the latent benefit from the social exchange of content sharing is greater for a weak-tie follower than for a strong-tie follower, everything else being equal.

In a social broadcasting environment, two possible explanations remain for the higher mean valuation of the shared content from a weak-tie follower's followers:

1. *New audience effect*: Because of the social broadcasting technology (in which whatever is posted or shared is broadcast to all followers), the possibility exists that the information has already been circulated to more of a strong-tie follower's followers than to a weak-tie follower's followers.[33] Holding the total number of a potential

[33] One important observation is that a strong-tie follower's followers are more likely

sharer's followers constant, the expected number of followers who are new to the information is larger for a weak-tie follower. Therefore, a weak-tie follower can reach a larger new audience, and hence the sharing gives a greater social exchange benefit.

2. *Informational value effect*: The information to be shared is *intrinsically* more valuable to a weak-tie follower's followers than to a stronger-tie follower's followers. Therefore, a weak-tie follower is more willing to share it because the sharing is expected to yield higher social exchange benefit.

3. A third possibility is that both of these two effects exist.

We test the three possibilities in model 6 by adding two empirically constructed followers-overlap measures into the second-stage retweeting equation. Mathematically, we define two versions of an overlap index of followers:

$$OI_{ti}^{W1} = \frac{\bar{W}_{ti}}{\sqrt{W_t}\sqrt{W_{ti}}} \quad OI_{ti}^{W2} = \frac{\bar{W}_{ti}}{\min\{W_t, W_{ti}\}},$$

where $\bar{W}_{ti}$, W_t, and W_{ti} are the number of mutual followers author t and user ti shared, the number of followers author t had, and the number of followers ti had, respectively. OI_{ti}^{W1} and OI_{ti}^{W2} basically measure how "similar" user ti's followers and author t's followers are: The larger the index is, the more similar the two sets of followers are. The indexes are

to be simultaneously following the author than a weak-tie follower's followers. Readers can refer to Appendix I for an empirical test.

also used in Appendix I, where we test whether unidirectional relationships are weaker than bidirectional ones. Readers can refer to Appendix I to see more discussion on the indexes.

We include OI_{ti}^{W1} and OI_{ti}^{W2} to capture the *new audience effect*, the first explanation. If it is indeed a driver of the result, we expect OI_{ti}^{W1} and OI_{ti}^{W2} collectively to have a negative effect on retweeting probability: If a user has a large number of followers who also follow the author, then he or she should be less willing to share the information. Moreover, if the *new audience effect* is the sole driver, then the weak-tie indicator w_{ti} should have no effect on retweeting probability once we include the two indexes.[34] If we find the two indexes have negative coefficients *and* the weak-tie indicator still has a positive coefficient, then we should conclude that both the *informational value effect* and the *new audience effect* exist.

The result of model 6 shows that the coefficients of the two indexes are indeed negative. Although the second version of the overlap index, OI_{ti}^{W2}, separately is insignificant, collectively they are significant with 99.9% confidence level. The magnitude of the coefficient of w_{ti} decreases from model 3, but, it is still positive at 0.1% significance level. These two findings together support the third possibility: Both the *informational value effect* and the *new audience effect* exist.

[34] Assume the two indexes have perfectly captured the *new audience effect*.

Probability of Retweeting		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
		Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)
			p_1: Probability of Consumption upon Receipt				
$\log m_{ti}$	# of Repetitions	0.174***	0.494***	0.340***	0.340***	0.338***	0.434***
		(8.21)	(4.81)	(4.37)	(4.37)	(4.47)	(5.86)
$\log V_{ti}$	# of Followings	-0.170***	-0.639***	-0.472***	-0.473***	-0.475***	-0.566***
		(-9.19)	(-10.38)	(-5.14)	(-5.05)	(-5.02)	(-7.23)
w_{ti}	Weak tie					-0.076	
						(-0.46)	
			p_2: Probability of Retweeting upon Consumption				
w_{ti}	Weak tie	0.218***	0.284***	0.220***	0.237*	0.249**	0.175***
		(5.13)	(5.57)	(5.52)	(2.14)	(3.22)	(4.11)
$\log W_{ti}$	# of Followers	0.087***	0.115***	0.101***	0.103***	0.102***	0.131***
		(5.41)	(6.32)	(7.46)	(4.95)	(7.21)	(6.83)
$w_{ti} \log W_{ti}$	Weak tie × # of Followers				-0.003		
					(-0.17)		
OI_{ti}^{W1}	Overlap Index of Followers I						-2.131*
							(-2.44)
OI_{ti}^{W2}	Overlap Index of Followers II						-0.429
							(-1.33)
ρ	Correlation		-	-0.836***	-0.835***	-0.834***	-0.606*
	(p-value)			(0.000)	(0.000)	(0.000)	(0.034)
# of Observations		24,403	24,403	24,403	24,403	24,403	24,403
Pseudo Log-Likelihood		-3,953.823	-3,921.876	-3,913.125	-3,913.112	-3,913.010	-3,892.148

Table 2.5: Result of Maximum Likelihood Estimation

2.6 Managerial Implications

Impression vs. Consumption

Internet display advertisement is often priced based on cost per impression or cost per action (e.g., cost per purchase, cost per click). However, it is important to realize that an ad being displayed is not equivalent to an ad being consumed. In other words, between the stage of impression and action is a stage of consumption, which does not necessarily occur after an impression because Internet users are often overloaded with information. The popularization of social broadcasting technologies, or social media as a whole, has greatly facilitated decentralized information production, which further leads to an explosion of user-generated content.[35] Then the question arises: Of the content being produced, how much is actually being consumed? One answer to this practical, important question suggests the possibility of another way of pricing for display advertisement: cost per ad consumption. This approach has largely been ignored in the literature because ascertaining whether an Internet user actually reads or watches an ad is difficult. Indeed, neither the content creator nor any third-party can observe whether an individual has consumed a piece of content supplied to him or her.

What our empirical model can contribute is that the estimation of equation (2.3) provides a simple yet useful way to quantify the consump-

[35] Taking Twitter as an example, as of May 2011, the average volume of tweets posted per day had reached 150 million (i.e., more than 1,700 tweets per second.)

tion probability of a piece of content in a social broadcasting network. Essentially, our model solves this problem in the Twitter context by exploiting the fact that observed acts taken upon content can be used to infer unobserved consumption. A good starting point is to fit the first-stage probability, which we provide below for ease of reference:.

$$p_1 = \Phi(e + b_1 \log m + b_2 \log V).$$

If we set $m = 1$ (i.e., a tweet enters a user's timeline only once as most tweets do), the fitted $\hat{p_1}(V)$ would be the expected probability for a user with V followings to consume one particular tweet received. By the law of large numbers, $\hat{p_1}(V)$ is also the fraction of received tweets that would be consumed by an average user with V followings. The lower/solid curve in Figure 2.9 shows how $\hat{p_1}$ changes with V when m is fixed at 1, using estimates obtained from our benchmark model. Consistent with intuition, $\hat{p_1}$ is a decreasing function of V, reflecting that Twitter users who follow more people, on average, consume a smaller portion of all tweets they receive. We also label in Figure 2.9 the fitted probabilities for Vs equaling 5^{th}, 50^{th}, and 95^{th} percentiles in our retweet dataset (see Table 2.3). We find that, on average, users who follow 25 people (5^{th} percentile in our sample) consume almost every single tweet they receive (94.1%); users with median number of followings (347) probably ignore more than half of the received tweets (54.8%); and users whose number of followings is 3,297, the 95^{th} percentile in our sample, are expected to consume only

5.9% of their followings' tweets. It is an interesting future research question to identify reasons why a significant portion of social broadcasting users follow a large number of people who produce so much more content than the users can possibly consume.

In social media, sharing promotes information diffusion by helping users who otherwise wouldn't be exposed to some content get exposed to and consume it. The second fact we can learn from equation (2.3) is that, in addition to helping information traverse longer social distances and to reach more people, sharing in social broadcasting also helps at the consumption stage of a content cycle (especially when the problem of information overload exists) by creating repeated exposure to a piece of content, which increases the probability of consumption. In Figure 2.9, the upper/dashed curve depicts $\hat{p_1}$ for $m = 10$ (i.e., when 10 retweets enter a user's timeline). The expected probabilities at the labeled V values all increase accordingly, but the increments are different (from 94.1%, 45.2%, and 5.9% to 99.7%, 84.5%, and 33.6% respectively). The expected probability for a user with the median number of followings increases the most — by 39.3%.

Influence

Measuring a user's social influence in an online community is of great interest to managers who want to leverage the power of social media. On Twitter, a user is often regarded as being influential when many

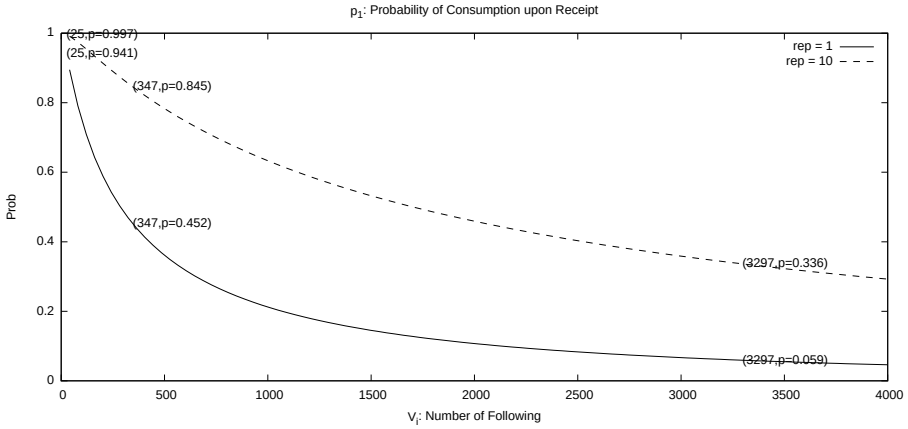

Figure 2.9: The Probability of a Tweet's Being Consumed upon Receipt

people retweet her tweets. Indeed, the depth of penetration and breadth of reach of one's words in an online community are important aspects of social influence. Our model measures the role that social network characteristics play in the information diffusion process. Combined with the probability of consumption, we can compute the expected total number of consumers of a user's tweet based on his or her social network characteristics, which may serve as a starting point for measuring his or her social influence.

One important implication of our study is that having more followers does not directly translate into greater social influence. In particular, the strength of social ties between a user and her followers should have an important moderating role, because it can greatly affect the followers' willingness to forward her messages. To see this more intuitively, we plot the fitted retweeting probabilities in Figure 2.10 for $w = 0$ (solid curve)

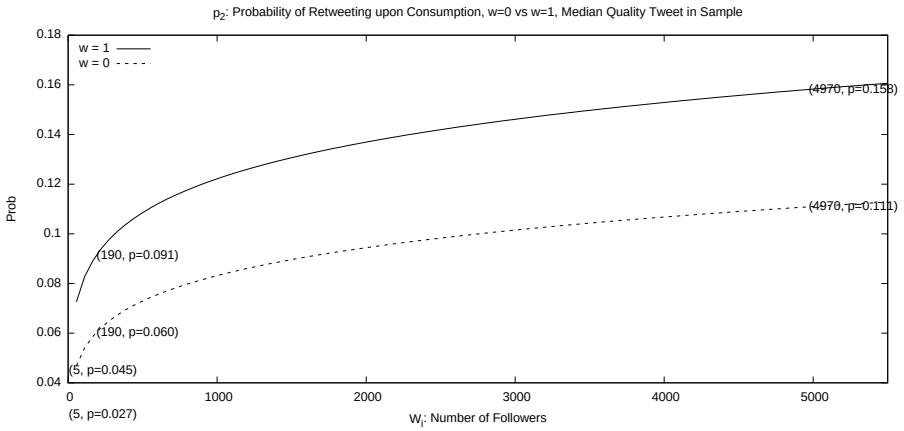

Figure 2.10: The Probability of Retweeting a Tweet upon Consumption

and $w = 1$ (dashed curve), fixing α_t at the median value in our sample. The difference between the conditional probabilities of retweeting for a unidirectional follower and for a bidirectional follower is significant. For example, when W, the number of followers, equals 190, the median number in our sample, the conditional likelihoods of retweeting are 6.0% for a bidirectional follower and 9.1% for a unidirectional follower. The latter is more than 50% higher in percentage.

2.7 Conclusion

An important question in the field of information systems is how information or knowledge is disseminated in an online community (with or without an organizational form). Large-scale empirical studies to address this question have traditionally been challenging because of the dif-

ficulty of obtaining detailed micro-level data. To the best of our knowledge, this chapter is the first such study in the information systems field, where publicly available data from Twitter is used to explore people's voluntary information relay process.

Using a carefully designed data collection process and a series of econometric analyses, we find that information is more likely to be retweeted through weak ties on Twitter. This result is complementary to Granovetter's finding, which advocates for the important role of weak ties in carrying novel information (Granovetter 1973). The implications of our finding are far-reaching. On the one hand, our theory, which is based on two highly influential sociological theories – the social exchange theory and the strength of weak tie theory – and is supported by the latest data from one of today's largest online social networks, reveals the important role that weak ties play in facilitating information dissemination in the social network through people's voluntarily information relay behavior. On the other hand, the interesting connection between tie strength and retweeting behavior indicates the importance of incorporating tie strength when measuring personal influence on Twitter, which is a question of fundamental importance to both researchers and practitioners.

As one of the first in the information systems field to bring together the huge amount of public data on Twitter with sociological theories to study information diffusion in social broadcasting networks, the chapter is not without its limitations. First, the tweets in our dataset

were not randomly sampled. By using this dataset to study the effect of tie strength in information sharing, we implicitly assumed that tweet "quality" changes everyone's retweeting probability only *uniformly*. Relaxing this assumption requires additional work (including obtaining a new dataset) to test whether our results hold when the quality of tweets is moderate or low. Second, we measured tie strength using a binary variable based on whether a link is unidirectional or bidirectional. Measuring tie strength based on the amount of conversation between two Twitter users would be an alternative approach. Third, we used only an author's immediate followers and omitted higher-order potential followers in empirical analyses. As we discussed in the data section, this was due to the difficulty of collecting network graph data for *all* higher-order potential retweeters. In future research one could try to overcome the difficulty by, possibly, sampling these users. Fourth, we observed only one snapshot of the social network and thus modeled it as fixed and exogenous. Future research can examine the interplay of user behavior and the dynamics of underlying network structure. Another possibility for extending the current study is to include more user-specific variables (e.g., demographic information) and tweet-specific variables (e.g., constructed from natural language processing) into the econometric model. Of course, these extensions pose new challenges in terms of data collection and data processing. Nevertheless, they are certainly interesting directions to pursue in the future.

Chapter 3

Shall I Go? The Unequal Effects of Friends' Check-ins

3.1 Introduction

The ongoing innovations of social networking and mobile technologies and their cooperation and integration with both online and offline businesses and services have given users unprecedented ease to share their daily activities with friends. For example, music service Apple iTunes has incorporated a social feature that lets users "ping" the songs they have purchased or are listening to; ticket seller Ticketmaster's application permits users to complete a transaction right from Facebook and easily share with their friends what live events they plan to attend; location-based mobile application Foursquare, by verifying users' GPS coordinates, allows them to "check-in" to physical venues they are currently visiting. Every time a user pings a song, shares a ticket-purchase, or checks-in to a venue, a message containing the information is sent out to her connected friends, who then can read about the activity, probably in real time, and might later decide to try out the song, the live entertainment, or the venue themselves.

As the social layer being gradually woven into real-world businesses and services, nowadays more and more of them have begun to provide incentives to customers for sharing their consumption experience via pings and check-ins. For instance, online file-synchronization services offer free storage space to people who indicate on Facebook or Twitter that they are using their software; restaurants offer eaters coupons when they check-in to the franchises at some specific time. The underlying idea behind these so-called "going-social" promotions is that, presumably, the very simple form of activity sharing by friends, such as the pings, shares, and check-ins, will be perceived as a kind of endorsement; thus, encouraging existing customers to do so can attract potential customers.

However, whether this belief is only an assumption, or it is also a matter of fact is unknown. In fact, many critics say that social networks are overloaded with information that is so "trivial" — for example, what people have just bought and where people are eating for lunch — that others simply will not pay attention to it. Indeed, as compared with the "old-fashioned" form of consumption-experience sharing such as customer reviews, the pings and check-ins convey too little information about the product or business and represent only an implicit and weak endorsement, if at all. Thus, to empirically test whether and how potential consumers react to this very simple form of activity sharing in social networks is important for understanding the economic value of these new social features.

In the present chapter, we approach this question by examining the check-ins that have been enjoying a stellar popularity in recent years. For example, Foursquare, the very first of location-oriented social networks, grew 1,000% annually from 2009-2011.[1] Bigger, all-purpose social networks such as Facebook, Twitter, and Google+ all have introduced similar functionalities. By using them, users can indicate their physical locations, such as restaurants, shopping centers, and movie theaters, through so-called "check-ins", so that their friends in the social network are aware of their activities. In our work, we model the user decision of visiting a venue, taking into account the endorsement effect[2] of friends' check-ins, examining its structure, and testing its existence using observational data collected from the Internet.

Identifying social/peer/network effects is challenging in terms of econometrics. Manski (1993) pointed out the now famous "reflection problem" in a "linear in means" model in which the behavior of an agent is influenced by the mean behavior of some "reference group" of which the agent herself is a member. In our present study, the problem does not occur because (1) our dataset has the advantage of containing the explicit relationships among the agents (i.e., the network graph); hence, we can make a natural assumption that an agent is influenced by her "friends"

[1] See http://en.wikipedia.org/wiki/Foursquare.

[2] In various academic disciplines, it is also often called social/peer/neighbor effect/influence. In this chapter, social effect, endorsement effect, and neighbor/friend influence are synonymous and we use them interchangeably.

observed in the dataset (one is not oneself's friend) instead of defining a "reference group" according to certain common characteristics (Bramoullé et al. 2009); and (2) we assume individual behavior varies with the lagged rather than contemporaneous value of friends' behavior (Manski 2000).

However, even though we do not have the "reflection problem," identifying social effects can still be complicated by the unobserved individual heterogeneity. This problem is related to endogenous relationship formation, also known as the *homophily* phenomenon; that is, agents tend to bond (in our case, form friendships) with others who possess similar characteristics. If some of these unobserved individual characteristics also affect the behavior in consideration, then the regressor that captures social effects will be endogenous. To overcome this problem is not easy, because we have virtually no access to personal information of the network members besides the structure of the network surrounding them. In this study, we approach this problem by adopting an idea developed in the area of collaborative filering and recommendation systems: The structure of the network can be used to infer individual-level characteristics. That is, we assume the unobserved individual heterogeneity is determined by a set of latent features of the agents, which also drive them to form friendships; then, the endogeneity problem can be solved by uncovering the latent features embedded in the network structure. Specifically, we apply the machine learning technique *nonnegative matrix factorization* (Lee and Seung 1999) to uncover the agents' latent features from the network

adjacency matrix and use them in the econometric analysis. As far as we are aware, we are the first to apply the matrix factorization technique in economics.

We model a finite and fixed network of rational agents, who can visit a venue at discrete times. If agents visit the venue, they may send out a check-in status to their network neighbors. In each period, the perceived benefit of visiting for a new visitor is a function of the individual's own characteristics, the period-specific trend, and also the endorsement effect of the past check-ins at the venue by her connected social neighbors. When adopting a seemingly innocent and widely used assumption about individual behavior — Every neighbor's endorsement is equal, the social effect on a potential visitor reduces to the (normalized) number of unique endorsements received (also the proportion of social neighbors who have checked-in). Thus our benchmark model to test resembles the so-called *local threshold* models in social contagion literature (Morris 2000, Watts 2002).

We conduct our empirical analyses on a unique dataset from a major location-based social networking site. We observe both the interpersonal connections and the sequential visits to venues (check-ins). In a sharp contradiction to intuition and theoretical models in the diffusion literature, we find that, in our benchmark model, the normalized number of unique endorsements by neighbors is actually not a good predictor of the likelihood of a new visit. Consider an example in which two non-

visitors each have 10 friends: One of them receives endorsements from three friends, and the other receives five. Our result indicates that the one receiving five endorsements does *not* necessarily have a higher probability to visit the venue, everything else being equal. We suggest that a more detailed relationship between each connected pair of individuals should be considered — for example the "proximity" of users implied by the observed connection patterns, an idea that can be traced back to a series of influential sociological papers (see Granovetter 1978, Burt 1987, Van den Bulte and Lilien 2001 and Centola and Macy 2007). We show that after weighting the endorsements by a parsimonious "proximity" measure (see Granovetter 1973 and Granovetter 1983 in sociology, Liben-Nowell and Kleinberg 2007 in computer science), the social effect becomes significantly positive. This result indicates that the endorsements from different social neighbors have unequal impacts on a focal user; in particular, the impact is expected to be larger if the endorsement comes from a "closer" neighbor. Thus, our work also joins the literature that investigates asymmetric influential roles that individuals play in a diffusion process (e.g., Goldenberg et al. 2009, Nair et al. 2010). However, unlike in the studies such as "innovative hub" or "opinion leader," which concern an individual's global stature, our finding is an evidence of unequal local, person-to-person influences. Additionally, we find that repeated check-ins have a larger effect.

Our findings on the endorsement effect of check-ins shed light on

the economic value of the location-based social networks, one of the most popular genres of online community. Consumers spend a great deal of time and money searching for products and services that meet their tastes and needs. In many markets, the product space is so large that a complete search is very costly; thus, the consumers are often unaware of or poorly informed about a substantial portion of the available choices. For markets of *experience goods*, this problem can be especially severe because consumers are not perfectly certain about their preferences before consumption. Previous literature has suggested that consumers can learn by observing the choices of others (Bikhchandani et al. 1992). Indeed, Amazon posts the ranking of purchases in each category of products; TripAdvisor ranks the popularity of hotels in a certain area. Our finding of a positive social effect indicates that the check-ins may play a similar role in facilitating people's search for venues such as restaurants and nightclubs by making the observational learning more effective. Furthermore, the new technology also deviates from the traditional observational learning paradigm in two important ways. First, it introduces individual identities and social relationships into the learning process, which is particularly important for products or services for which great variation exists in people's tastes. In this case, observing the choices of network friends rather than anonymous others (as is in the case of sales ranking) helps consumers better learn their preferences, because they are likely to know the "similarity" of their tastes and the tastes of their network friends. Our

finding of unequal effects of friends' check-ins support this argument. Second, in the classic observational learning models, each agent takes an action only once, and the outcome is unobserved; in our case, one network member can check-in multiple times, implicitly indicating a positive outcome from earlier visits. In this sense, multiple check-ins resemble the word-of-mouth effect.

The remainder of this chapter is organized as follows. In Section 3.2, we review the related literature. In Section 3.3, drawing upon earlier studies, we introduce our individual-decision model, which leads to the benchmark econometric model in which the effects of friends' check-ins are assumed to be equal. In Section 3.4, we describe the dataset, explain how we discretize the observed history of check-ins, and define the correspondence between the model concepts and data. In Section 3.5, we first present the result of the benchmark model, which indicates that the equal-effects assumption fails to capture the structure of the endorsement effect. Drawing upon the work in sociology and computer science, we show that weighting friends' check-ins by a proximity measure can yield a better result. We then test the robustness of the result with alternative specifications and also discuss the implication of our finding on the economic value of the location-based social networks. Lastly, we conclude and point out potential future research directions.

3.2 Literature Review

Our work mainly draws on two streams of previous literature: (1) the literature on innovation/behavior diffusion in a population, and (2) the literature on matrix-factorization-based recommendation systems in computer science.

The diffusion of a new product/idea/behavior has been an extensively studied topic in the economics, marketing, and sociology literature. One of the most important questions in this area is how earlier adoptions, in different circumstances labeled as *social influence* or *network effect*, affect later adoption decisions. It is well known that different micro-level influence mechanisms lead to completely different cumulative adoption curves (Chatterjee and Eliashberg 1990, and Young 2009, among others) and provide different answers to the question of why some new behaviors become "viral" (Leskovec et al. 2007), whereas others are confined to only a small subset of the population (Rogers 1995, Watts 2002). Moreover, with the increasing popularization of online communities and social media, studying the question also sheds light on identifying network "influencers" (Kempe et al. 2003), which may potentially guide marketers to better harness the power of word of mouth (WOM).

Modeling social influence structure can be dated back to the very early studies on large scale innovation diffusion in a population. These models, especially the ones developed in the era when only aggregate data (e.g., total number of adopters, or total numbers of adopters in different

groups or geographic regions) were available, generally imitate epidemiological models, in which adopters act as "infectors" and non-adopters constitute the "susceptible" group. "Social contagion" models, so they are called (Bass 1969, Mahajan et al. 2000 and references therein). For the most part, aggregate phenomena are the research focus and the population is modeled as more or less amorphous. Without observing the explicit network among the individuals, an encounter between an adopter and a susceptible individual is assumed to be random. The adoption "decision" then either deterministically or stochastically depends on only the *global* ratio of adopters and non-adopters, which is the classical *random mixing* assumption (Granovetter 1978, Van den Bulte and Lilien 1997). The intensity of *social influence* or *social pressure* on individual decisions about whether or not to adopt therefore boils down to the proportion of others in the population who have already adopted (which is also the case in observational learning models; for example, see Bikhchandani et al. 1992; Young 2009).

Although these models provide many insights about a diffusion process at the aggregate level (e.g., equilibrium adoption rate, good/bad herds), most of them deliver little in understanding the fine structure of person-to-person influence: In the majority of realistic scenarios, people usually care much more about the decisions made by family and close friends than by the full population. In recent years, with data that provide explicit interpersonal relationships becoming more available, diffu-

sion studies that consider nonrandom social influence patterns have become increasingly popular. Instead of looking at the *global* ratio of the already adopted, these studies assume that individuals are influenced not by all but by only some relatively small number of *local* influencers (e.g., connected social network neighbors). In particular, observing the "real" network as a graph allows researchers to model and test the finer structure of *social influence*. For example, Hill et al. (2006) found that the direct contact with an existing customer increases the adoption likelihood for a potential telecommunication service customer; using prescription data, Nair et al. (2010) documented that a research-active specialist, or "opinion-leader" can influence a physician's prescription behavior in the same "reference group;" Katona et al. (2011) studied the sequential adoption of a social networking technology and found that the adoption probability is positively associated with the number or proportion of adopted neighbors and with the density of connections among them.

We build our model following this line of "diffusion in a network" literature and use it to test the effectiveness of the new popular social features (in our context check-ins) in influencing individuals' searching for products or services that meet their needs (in our context venues). The lack of individual characteristics makes our empirical analyses vulnerable to the problem of endogeneity caused by the unobserved heterogeneity. As mentioned in the introduction, we uncover the individual-level characteristics from the network graph, a method developed in the area of online

recommendation systems. This area focuses on predicting consumers' future purchases/ratings of products based on the purchase/rating history and, in some cases, the social relationships between consumers. As a class of latent factor models, matrix factorization models have recently emerged as a state-of-art methodology for recommendation systems. The idea is to derive a "high-quality low-dimensional feature representation" of users based on analyzing the social network graph matrix (or user-product matrix) and then use the latent features as the basis for recommendation (Ma et al. 2008, Koren et al. 2009). The methodology's effectiveness has been proven in various applications, and in particular, the Netflix competition. The specific technique we use is *nonnegative matrix factorization* (NMF), which was popularized by Lee and Seung (1999). Lee and Seung (2001) analyzed algorithms for computing NMF.

3.3 Model

Since examining social effects is our research goal, we start by introducing the network structure. We assume that the population is a group of N agents, whose connections/relationships can be described by a $N \times N$ adjacency matrix G, where the i, j element

$$g_{ij} = \begin{cases} 1, & \text{if } i \text{ follows (can be influenced by) } j \\ 0, & \text{otherwise} \end{cases}, \quad i, j = 1, 2, \ldots, N.$$

Note that we do not require G be symmetric. However, it can be, for example, when we are looking at *friendships*, in which case $g_{ij} = g_{ji}$,

$\forall i, j$. The *influencers* of an individual i, denoted $V(i)$, are the set of agents whom i follows, so $V(i) = \{j | g_{ij} = 1\}$. Similarly, we define the *influencees* of agent i to be the set of network members who follow i: $W(i) = \{j | g_{ji} = 1\}$. Obviously if G is symmetric, we have $V(i) = W(i)$ and call them i's *friends*.

Agents choose whether or not to visit a venue v. We assume visits happen in discrete time intervals, indexed by $w = 0, 1, \ldots$ (we use letter w to denote time instead of the more conventional t for consistency with our weekly data, which we describe later). $y_i^w \in \{0, 1\}$ indicates whether i visits the venue at time w. Facilitated by the social networking technology, once an agent visits v, she sends a check-in status (also called an "endorsement" hereafter) to all of her influencees. Call the group of agents who visited v at time w the *w-visitors*, denoted by $A(w)$, and who have not visited v by $w-1$ and can potentially visit v at time w the *w-risk set*, denoted by $R(w)$. $\hat{A}(w) = \cup_{\omega=0}^{w} A(\omega)$ is thus all the visitors up to time w. Let the binary variable $\hat{y}_i^w = 1$ if individual $i \in \hat{A}(w)$. It is easy to see that, for an agent i in $R(w)$, the group of people who have sent her endorsements by time $w-1$ is $\hat{A}(w-1) \cap V(i)$ — that is, the visitors among i's influencers (we also call them i's visitor-influencers).

We assume that the payoff to agent i in the risk set from taking action $y_i^w = 1$ can be written as:

$$u_i^w(y_i^w = 1) = v_i^w(x_i, x^w, \{A(w-1), A(w-2), \ldots, A(1), A(0)\}) - \epsilon_i^w, \tag{3.1}$$

where v_i^w is the perceived benefit of visit and ϵ_i^w is a stochastic disturbance/cost independently and identically distributed across both individuals and times.[3] v_i^w is a function of individual characteristics, x_i; the time-specific effect, x^w; and also the past endorsements received from others, which are determined by $V(i)$ and the visiting history $\{A(w-1), A(w-2), \ldots, A(1), A(0)\}$.[4] As usual, we normalize $u_i^w(y_i^w = 0) = 0$, so i in the risk set visits v at time w if and only if $v_i^w \geq \epsilon_i^w$.

The way by which $\{A(w-1), A(w-2), \ldots, A(1), A(0)\}$ affects v_i^w captures the social effects. In our benchmark model, we assume that the social effect can be captured by the (normalized) number of unique endorsements received from influencers, which is also the number of visitor-influencers divided by the number of the focal individual's influencers. This assumption is widely used in the diffusion literature, and it is also a very natural one in our context because the technology shows to users the group of friends who have checked-in to the venue. We further write

[3] ϵ_i^w includes idiosyncratic shocks, e.g., the difficulty of reserving a seat at the venue at i's specific choice of time.

[4] We model social neighbors' endorsements as a *direct benefit* into the individual preference. We can interpret v_i as individual i's belief or expectation of the benefit of visiting the venue and the belief may be adjusted according to social neighbors' endorsements. This kind of social effect is also called "peer influence," "neighborhood effect," and "conformity" in the literature. The assumption that an agent can "remember" all the past check-ins by friends is based on the fact that the technology allows the users to see the group of friends who have checked-in previously. In addition, we do not allow $A(w)$ to affect v_i^w, so there is no contemporaneous interaction among individuals.

v_i^w as a linear sum (cf. Van den Bulte and Lilien 2001, Valente 2005)

$$v_i^w = \alpha_i + x_i\beta + x^w\gamma + \delta \sum_{j=1}^{N} \hat{g}_{ij}\hat{y}_j^{w-1}, \tag{3.2}$$

where α_i is the unobserved fixed utility component and the last term

$$\sum_{j=1}^{N} \hat{g}_{ij}\hat{y}_j^{w-1} \tag{3.3}$$

is the social effect. $\hat{g}_{ij}$ basically measures how j's endorsement affects the benefit for i. In many cases, the researcher observes only the binary connection patterns, or G, so from a modeler's point of view, $\hat{g}_{ij}$s should be based only upon g_{ij}s. For js that satisfy $g_{ij} = 0$, $\hat{g}_{ij}$s are naturally chosen to be 0. For js that $g_{ij} = 1$ — that is, i's influencers — the form of $\hat{g}_{ij}$ that has been used extensively in earlier models (explicitly or implicitly, e.g., Watts 2002 in social network, Katona et al. 2011 in marketing, Bramoullé et al. 2009 in economics) is

$$\frac{1}{|V(i)|} = \frac{1}{\sum_j g_{ij}}.$$

Mathematically, it leads to the weighted influence matrix, $\hat{G}$

$$[\hat{g}_{ij}] = [\frac{g_{ij}}{\sum_{k=1}^{N} g_{ik}}] \text{ for } \sum_{k=1}^{N} g_{ik} \neq 0. \tag{3.4}$$

Here is a numerical example

$$G = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \hat{G} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

This commonly used specification is straightforward to interpret: Each agent, no matter how many other network members by whom she is influenced, puts equal weights on influencers, and the total weights are normalized to be the same. The value of expression (3.3) ranges from 0 to 1, the same for different individuals. In fact, under modeling choice (3.4), expression (3.3) reduces to a simple proportation

$$r_i^{w-1} = \frac{\sum_{j=1}^{N} g_{ij} \hat{y}_j^{w-1}}{\sum_{j=1}^{N} g_{ij}}, \tag{3.5}$$

that is, the proportion of visitors among i's influencers, or the number of unique endorsements normalized by the total number of influencers. Then, equation (3.2) can be rewritten as

$$v_i^w = \alpha_i + x_i\beta + x^w\gamma + \delta r_i^{w-1}. \tag{3.6}$$

A nonzero δ indicates the existence of social effect.

Such a specification — that is, assuming all influencers are equal, looks innocent, and seems to be a natural first step approximation, especially when the number of individuals in consideration is very large and detailed interactions among them are either unobserved or extremely hard to record by outside researchers. Indeed, many of the local threshold diffusion models (and social interaction models in which individual behavior is influenced by the mean behavior in reference peer group) and some social network analysis techniques — including Google search engine's famous PageRank (Brin and Page 1998) — are implicitly based on

such an assumption. However, as we will see in our subsequent empirical analyses, this seemingly innocent assumption can lead to a quite counter-intuitive result.

3.4 Data

The dataset we use comes from a major location-based social networking website in China. The service is almost always used as a mobile application: It allows registered users to post their location at a venue (called check-in in the application's terminology) via their GPS-enabled mobile devices, typically smart phones, and connect with friends. A check-in requires verification of users' GPS data, so it represents a real visit. Most checked-in places are mainly restaurants, shopping centers, nightlife sites, and tourist attractions. Friendship is mutual, and thus a relationship between two friends is undirected (symmetric G). Points are awarded at check-in, and users can also provide their comments or reviews when checking-in at a venue. The check-in status, possibly with a comment or review attached, is then sent to all friends. Users can choose to have their check-ins posted on other partner social networking sites such as Sina Weibo, Renren, Douban, and so on.

The dataset includes only a subset of the website's members, who were selected by (the company) randomly sampling ids in the population. To protect the users' privacy, we were not given their true online-ids, nicknames, registration dates, or any other demographic information such as

Time	Pseudo-id	Location (Encoded)
2010-08-19 03:46:52	*1803*	*EF6260A6EA3463D6*
2010-08-19 03:48:22	*405*	*866A6700B769EC89550271D60C131D8F*
2010-08-19 03:48:31	*2530*	*5859EE11867F5A6F*

Table 3.1: Sample Check-in Data

age, gender, and city. However, we were indeed provided complete data on friendships among this subset of users. In other words, we know *who* are *whose* friends in our sample. Since a friendship links two users and all users are equally likely to be in our sample, then the observed friendships in the dataset are also random. Therefore, we can conclude that the social network we observe is a representative "sub-network," and our empirical study on diffusion in this "sub-network" can shed light on the underlying diffusion process in the "whole network."

The other key part of the dataset is the users' complete history of check-ins or venue visits. The structure of this part of the dataset is illustrated in Table 3.1. We observe *when who* checked-in and *where*. Again, for privacy concerns, venue names were encoded into human-uninterpretable strings. Apart from the string itself, we know neither the actual geographic location nor the type of the venue (restaurant, shopping center, etc.). However, we do know that a venue was mapped into only one string.

There are three additional concerns regarding the data. The first is that, as in many other empirical applications that use online social network data, we have only one snapshot of the social graph (relation-

ships between users), whereas in reality the network itself is evolving constantly. Following the tradition in the literature, we also assume that the observed network at the end of study period contains true "real-life" relationships among users; these "cyber relationships" are *not* the cause of, but simply a digital mapping of "real-life" ones.[5] The second concern is that the observed diffusion of check-ins at venues may mix with the unobserved diffusion of the platform/application itself. This problem could be especially severe at the very early stage after the application's introduction when its user-base grew the fastest in early 2010. To reduce noises such as this to the largest extent, we choose to focus on venues whose first appearance in the dataset occurred after September 1, 2010, the latter half of the check-in history log. By doing so, we implicitly assume that by then the network operated on this application had entered a relatively stable period and the observed sequential check-ins reflected only the diffusion of venue visits. Third, we observe the diffusion processes for multiple venues. When we pool the venues together for econometric modeling, we need to consider both venue-specific effects and individuals' heterogeneous tastes towards different venues. Hence, we rewrite equation (3.6) here

$$v_i^{v,w} = \alpha_i^v + x_i\beta + x^{v,w}\gamma + \delta r_i^{v,w-1}, \tag{3.7}$$

[5]We essentially assume away the possibility that user i could friend some user j whom she had *not* known in real world, just because they happened to check in the same place at the same time and then get to know each other through the website. In this case, the friendship would be the result of online activities.

where the superscript v means "venue," and the i-sub-indexed fixed component α_i^v allows heterogeneous tastes and can be interpreted as the baseline utility i believes that she can get by visiting v.

Time-Independent Covariates

The total number of individuals in the dataset is 28,740. The observed time-independent covariates of user i include the number of friends ($l_i = \sum_j g_{ij} = \sum_j g_{ji}$) and two other measures of i's network statures: the (local) betweenness ($s_{bw,i}$) and the (local) clustering coefficient ($s_{cc,i}$). They are the social network analysis (SNA) variables typically used in the diffusion in network literature. Betweenness is a graph-based centrality measure first introduced in Freeman (1977). Here, we adopt a local version of betweenness defined in Katona et al. (2011),

$$s_{bw,i} = \sum_{j \neq k \in \{1,2,\ldots,N\}} \frac{g_{ij} g_{ik} (1 - g_{jk})}{\sum_{q \in \{1,2,\ldots,N\}} g_{jq} g_{kq}},$$

which focuses on i's relative importance as a local brokerage (Burt 2005).[6] The local clustering coefficient at node i (Watts and Strogatz 1998, Newman 2003) is defined as

$$s_{cc,i} = \frac{\sum_{j \neq k \in \{1,2,\ldots,N\}} g_{ij} g_{ik} g_{jk}}{\sum_{j \neq k \in \{1,2,\ldots,N\}} g_{ij} g_{ik}},$$

[6] Katona et al. 2011: "for every unrelated pair of users j, k among i's friends, the contribution of the pair j, k to the betweenness of i is inversely proportional to the number of length-2 paths between j and k."

		l	s_{bw}	s_{cc}	$l \cdot s_{cc}$	$c1$	$c2$	$c3$	$c4$	$c5$
# of friends	l	1.00								
betweenness	s_{bw}	0.79	1.00							
clustering	s_{cc}	0.03	-0.00	1.00						
	$l \cdot s_{cc}$	0.37	0.03	0.64	1.00					
individual	$c1$	-0.03	-0.01	-0.02	-0.05	1.00				
latent	$c2$	-0.04	0.02	-0.03	-0.08	-0.21	1.00			
features	$c3$	-0.02	-0.00	0.01	-0.00	-0.21	-0.22	1.00		
	$c4$	0.06	-0.00	0.05	0.15	-0.20	-0.22	-0.22	1.00	
	$c5$	0.03	-0.01	-0.01	-0.01	-0.31	-0.33	-0.28	-0.30	1.0000

Table 3.2: Correlation between Time-Independent Covariates and Latent Features

which measures the interconnectedness/density of relationships among i's friends.[7] A higher clustering coefficient indicates denser relationships in the local network. We also include the product of l_i and $s_{cc,i}$, because the clustering coefficient decreases quadratically as the size of the network increases while holding the probability of link formation constant. We do not discuss $s_{bw,i}$ and $s_{cc,i}$ more deeply since they are not the focus of the present research. Interested readers can refer to the lengthy social network analysis literature. x_i in equation (3.7) is thus

$$x_i = (l_i \quad s_{bw,i} \quad s_{cc,i} \quad l_i s_{cc,i}).$$

These time-independent covariates can be easily computed based on the network graph G. The correlations between these variables are provided in the upper left part of Table 3.2.[8]

[7] The numerator is the number of links among i's friends and the denominator is the maximum number of relationships possible among them.

[8] In Table 3.2, we drop the sub-index i for cleanness of notation.

Discretization and Time-Dependent Covariates

We break down the check-in history into non-overlapping weekly intervals: For each venue v, we denote the timestamp of the earliest check-in in our history log time 0 (also $w = 0$), the week immediately following the first check-in week $w = 1$, and so on, until we reach the end of the history. If the last week is not whole, we drop it to avoid any censoring issue.

For each venue v, in week w, we can then identify the visitors $A^v(w)$. We define the week w risk set of users to be the individuals who had not visited v by week $w - 1$ and have at least one visitor-friend. Note that a user could stay in the risk set for multiple weeks; once a user first checked-in, she would no longer appear in the risk set in the subsequent weeks. Hence, we focus only on new visitors in this research. Table 3.3 shows the discretized series of the check-in history for two venues. The third and fourth columns are the number of unique visitors and the total number of check-ins up to week $w - 1$. The fifth column is the size of the risk set. The last two columns are the numbers of check-ins in week w made by risk-set members and the users who were neither already-visitors nor risk-set members, respectively.

One thing to emphasize here is that we restrict the risk set to contain only visitors' friends (rather than all users who had not visited the venue). By doing so, we are *not* suggesting that the users outside our risk set have zero probabilities to check-in to the places. They actually did,

as shown in the last column of Table 3.3. Rather, the primary reason of this restriction is that we want to focus on the behavior of the subset of individuals who are more homogeneous: We have a good belief that they at least "knew" the venue and also were most likely to be "able" to visit the venue. China is a vast country and the service from which the data comes was and is used by people living in places, although mostly a handful of top-tier cities, that can be thousands of miles apart. People living in the city of Shanghai hardly know those non-landmark venues in the city of Shenzhen. Indeed, *awareness* is the very important first step to adopt a new behavior, and it is traditionally emphasized in the diffusion literature (Ryan and Gross 1943, Rogers 1995). Moreover, even if individuals were well informed about all the venues in the country, the check-in cost would vary hugely across each user-venue pair. To know the users' demographics and the venues' types and geo-locations can certainly help, but unfortunately we have no access to the information. Thus, we decide to let the risk set be only the visitors' friends. First, because of the functionality of the technology, they must have received the visitor-friend's check-in status, so we have good confidence that they "knew" this venue. Second, two friends tend to be geographically and socioeconomically more close to each other than a pair of random individuals. Thus, a friend of user i who visited venue v is more likely to live within "feasible" distance to and be able to afford v. Therefore, the cost of check-in at the venue should be more homogeneous for the visitors' friends.

Technically, the reason for making this restriction is to ensure the plausibility of our i.i.d. assumption on ϵ_i^w (equation (3.1)), and, in particular, the disturbance being uncorrelated with the number of endorsements received. Without the restriction, as we have argued above, ϵ_i^w would be correlated with r_i^{w-1} (equation (3.6)). On the other hand, the drawback of this restriction is that all observations in the subsequent analyses would have strictly positive r_i^{w-1}s, so we would not be able to compare the likelihoods of visiting for someone who received endorsements and someone who did not; we would only compare the likelihoods for people who received different numbers of endorsements.

There could exist a weekly-specific trend that monotonically changes the visiting likelihood. For example, a promotion campaign might take place at venue v in week w, so everybody's propensity of visiting it is likely to increase in week w. If a campaign lasts for multiple weeks, ignoring this effect would cause the disturbance to be serially correlated. Consequently, $\epsilon_i^{v,w}$ would be correlated with $r_i^{v,w-1}$, so the i.i.d. assumption would fail (see similar discussions in Van den Bulte and Lilien 2001, Nair et al. 2010). One way to solve this problem is to use venue-specific weekly dummy variables, which obviously would substantially increase the number of coefficients we have to estimate. The second way, which is what we do, is to use the number of check-ins made by the users who are neither already-visitors nor risk-set members (denote the number $o^{v,w}$) as a proxy for the weekly trend. In our examples in Table 3.3, this proxy

Week w	Ending Date	# of Visitors up to $w-1$	Total Check-in up to $w-1$	# in Risk Set	Risk C-ins in w	Non-Risk C-ins in w
1	*2010-10-04*	1	1	73	0	0
2	*2010-10-11*	1	4	73	11	33
3	*2010-10-18*	37	51	745	5	0
4	*2010-10-25*	42	58	775	1	1
5	*2010-11-01*	44	61	785	1	1
6	*2010-11-08*	46	63	802	2	0
7	*2010-11-15*	48	65	832	3	2
8	*2010-11-22*	53	70	870	6	2
9	*2010-11-29*	61	78	930	0	0
10	*2010-12-06*	61	78	930	3	2
11	*2010-12-13*	65	83	946	13	4
12	*2010-12-20*	80	104	1227	6	1
13	*2010-12-27*	84	111	1238	2	2
14	*2011-01-03*	88	115	1254	0	0
15	*2011-01-10*	88	115	1254	3	6
16	*2011-01-17*	97	125	1280	14	5
17	*2011-01-24*	113	146	1379	0	3
18	*2011-01-31*	116	150	1395	0	1
19	*2011-02-07*	117	151	1400	0	0
20	*2011-02-14*	117	151	1400	2	0
21	*2011-02-21*	119	153	1399	0	1
				9D3ACE0BC12099CC3F1371656A556B38		
1	*2010-09-30*	1	1	27	0	0
2	*2010-10-07*	1	1	27	0	0
3	*2010-10-14*	1	1	27	0	0
4	*2010-10-21*	1	1	27	0	0
5	*2010-10-28*	1	1	27	0	0
6	*2010-11-04*	1	1	27	0	1
7	*2010-11-11*	2	2	32	0	2
8	*2010-11-18*	4	4	65	2	18
9	*2010-11-25*	22	24	484	8	5
10	*2010-12-02*	33	52	642	0	2
11	*2010-12-09*	35	55	649	11	5
12	*2010-12-16*	47	78	760	12	5
13	*2010-12-23*	61	136	847	2	1
14	*2010-12-30*	64	140	880	2	8
15	*2011-01-06*	73	155	904	0	2
16	*2011-01-13*	75	160	905	17	7
17	*2011-01-20*	97	199	1036	7	4
18	*2011-01-27*	105	214	1056	0	3
19	*2011-02-03*	108	217	1062	7	4
20	*2011-02-10*	117	231	1162	0	2
				FF8FF39D7DF75AEBD705EC853A0F7BF4		

Table 3.3: Sample Discrete Intervals

variable is just the last column. So the time-dependent $x^{v,w}$ in equation (3.7) is

$$x^{v,w} = (o^{v,w}).$$

3.5 Empirical Results

We decide to use a subset of our data because the number of observations is too large. Even though we have only a moderate number of website users in the dataset and the size of the risk set (e.g., the fifth column in Table 3.3) is even smaller because of our restriction, we observe the diffusion processes for a total of 172,217 venues. For each of these venues, we also discretize its check-in history into multiple weeks. Therefore, if we were to include all the venues in the dataset to do the estimation, the number of observations would exceed 1 billion, which we cannot handle computationally. To overcome this problem, we choose to use only the top 50 venues that were checked-in most often in the observed history, which yields us a sample size of 690,896. Since we will consider both venue-specific effects and heterogeneous tastes of different individuals toward the venues, we conclude that using only popular venues will not cause a selection problem.[9] In addition, using popular

[9]In fact, we have estimated the most important specifications using data of top 100 venues, and two sets of 50 venues selected randomly from top 100. In each case, the estimates only slightly change. The signs and significance levels of the estimates are basically the same as reported here. Because of spatial limit, we do not report these results here.

venues also supports our assumption that past check-ins by others are endorsements rather than criticisms.

3.5.1 Results

Our benchmark econometric model analyzed in this section is based on equations (3.1) and (3.7). Following Bell and Song (2007) and Katona et al. (2011), we assume $\epsilon_i^{v,w}$ follows Gumbel distribution, so, after normalization on distributional parameters, the probability that agent i in the risk set visits venue v in week w is obtained as

$$\begin{aligned} \mathrm{P}(y_i^{v,w} = 1) &= 1 - \exp\{-\exp(v_i^{v,w})\} \\ v_i^{v,w} &= \alpha_i^v + x_i\beta + x^{v,w}\gamma + \delta r_i^{v,w-1}. \end{aligned} \tag{3.8}$$

Thus, it suggests that we use the complementary log-log link function to estimate the binary choice model. Parameter estimates are obtained by applying Maximum Likelihood (ML) method and standard errors are computed to be robust to venue-clustering.

Model 1a in Table 3.4[10] shows the result of the benchmark model. It corresponds to equation (3.8), except that for now we ignore the individual specific taste, α_i^v. The most important estimate, coefficient δ, is shown in the first row. Contrary to the general intuition and previous literature, we find that δ is significantly negative, while controlling the time trend and the observed individual network characteristics. If we believe

[10] *, **, and *** mean 5%, 1%, and 0.1% significance levels, respectively.

Probability of Visiting		Model 1a	Model 2a	Model 3a
		Coeff.	Coeff.	Coeff.
		(z-value)	(z-value)	(z-value)
Normalized #	Unweighted: r_i^{w-1}	-1.36***		-1.39***
of Endorsements		(-12.34)		(-11.57)
	Weighted: $\hat{r}_i^{w-1}$		2.06***	2.03***
			(29.11)	(28.31)
Time Trend	Weekly Trend: o^w	0.02***	0.02***	0.02***
		(6.18)	(5.93)	(6.10)
Time-independent	N of Friends: l_i	19.17***	25.35***	15.97***
Covariates	(1/1,000)	(10.33)	(15.49)	(9.90)
	N of Friends^2: l_i^2	-0.20***	-0.36***	-0.23***
	(1/1,000)	(-4.90)	(-7.20)	(-5.49)
	Betweenness: $s_{bw,i}$	0.29***	0.60***	0.38***
	(1/1,000)	(3.48)	(5.99)	(4.45)
	Clustering: $s_{cc,i}$	-0.10	-1.36***	-1.26***
		(-0.86)	(-12.74)	(-10.75)
	N $\times$ Clustering: $l_i s_{cc,i}$	-0.01	0.08***	0.05***
		(-0.64)	(5.28)	(3.32)
N of Observations		690,896	690,896	690,896
Pseudo Log-Likelihood		-17,145.54	-16,944.26	-16,825,66

Table 3.4: Results of Complementary Log-Log Regressions: Part I, α_i^v Unconsidered

Probability of Visiting		Model 1b	Model 2b	Model 3b
		Coeff.	Coeff.	Coeff.
		(z-value)	(z-value)	(z-value)
Normalized #	Unweighted: r_i^{w-1}	-1.00***		-1.06***
of Endorsements		(-9.04)		(-9.41)
	Weighted: $\hat{r}_i^{w-1}$		0.90***	0.94***
			(12.37)	(13.00)
Time Trend	Weekly Trend: o^w	0.02***	0.02***	0.02***
		(6.73)	(6.82)	(6.83)
Time-independent	N of Friends: l_i	27.80***	32.07***	25.09***
Covariates	(1/1,000)	(11.70)	(14.84)	(11.62)
	N of Friends^2: l_i^2	-0.37***	-0.47***	-0.37***
	(1/1,000)	(-6.62)	(-7.67)	(-6.70)
	Betweenness: $s_{bw,i}$	0.62***	0.81***	0.63***
	(1/1,000)	(5.43)	(6.53)	(5.65)
	Clustering: $s_{cc,i}$	-1.00***	-1.61***	-1.52***
		(-8.51)	(-15.98)	(-13.99)
	N $\times$ Clustering: $l_i s_{cc,i}$	0.16***	0.21***	0.19***
		(9.54)	(12.60)	(11.63)
N of Observations		690,896	690,896	690,896
Pseudo Log-Likelihood		-15,937.43	-15,931.93	-15,871.34

Table 3.5: Results of Complementary Log-Log Regressions: Part II, α_i^v Considered

that the social effect of friends' endorsements should be either zero (no effect) or on aggregate positive, considering the fact that we select popular venues, then two explanations exist for the counterintuitive negative sign of the δ coefficient: (1) Because of omitting the unobserved fixed effect α_i^v, our econometric model is misspecified and hence produces an incorrect result; or (2) treating every friend's endorsement as the same (equation (3.4)), the assumption that leads to regression model (3.8), is implausible in capturing the structure of the social effect of past check-ins. We are going to explore both of the two possibilities, propose solutions, and report the new results in the other columns in Table 3.4 and 3.5.

Endogeneity. To see why leaving out the heterogeneous tastes α_i^v invalidates the econometric model (see Nair et al. 2010 for a discussion on physician-specific effect on prescription adoption), recall that an individual may stay in the risk set for multiple weeks. Particularly, the individuals who have lower values of α_i^v are likely to remain for a longer time period. Indeed, a user who believes that she will dislike a venue very much (extremely low α_i^v) may never visit the venue, no matter how many of her neighbors have already visited there and sent her endorsements. Furthermore, it is not hard to see that, for a user i staying in the risk set for multiple weeks, the number of endorsements received by i can only increase as time goes by. Therefore, mathematically, α_i^v and $r_i^{v,w-1}$ are negatively correlated. Leaving the unobserved α_i^v into the disturbance causes the estimates to be inconsistent. A high $r_i^{v,w-1}$ may simply pick up

the effect of a low α_i^v, yielding a negative coefficient. Another aspect of the endogeneity problem is related to the phenomenon of *homophily*: the tendency of individuals to associate and bond with similar others. One may think that two friends are likely to have similar tastes (in our context positively correlated α_i^v) than a pair of random individuals. Therefore, α_i^v could be (positively) correlated with $r_i^{v,w-1}$. Thus, the identification of the social effect is complicated by the unobservability of α_i^v.

To solve the endogeneity problem is difficult, because the unobserved heterogeneity is not individual-specific, but individual-venue specific: Different people have different tastes toward different venues. Hence this problem cannot be solved by using dummy variables. Technically, it resembles a panel/clustered data binary choice model with heterogeneity, where the fixed effect is correlated with some observed covariates (Wooldrige 2001). Here, we innovate to use a machine learning technique to deal with this problem.

Nonnegative Matrix Factorization (NMF). The endogeneity problem is caused by the unobservability of heterogeneity α_i^v. Statistical methods that deal with this problem typically assume certain probabilistic distribution for α_i^v.[11] The approach we explore in this subsection is to find

[11] One existing modeling alternative provided in the econometrics literature is to specify how α_i^v probabilistically relates to the observed covariates. One example is Chamberlain's correlated random effects specification (Chamberlain 1980; Mundlak 1978), which imposes the assumption that the unobserved heterogeneity conditional on the mean of observed covariates follows normal distribution.

a set of individual-level "latent factors" that determine α_i^v by factorizing the adjacency matrix G.

The idea originates from researchers in computer science who design and implement online recommendation systems that use network graph data to predict products that users might be interested in. The key assumption underlying their method is that the relationships between the users and the users' preferences toward the products are simultaneously induced by some hidden lower-dimensional feature space (Ma et al. 2008). Under this assumption, even though individuals' preferences, in our case α_i^v, are unobserved, they can be learned by factorizing the observed network graph, in our case G.

We adopt the idea here. α_i^v is baseline utility user i believes she can obtain by visiting v. We assume it to be

$$\alpha_i^v = \theta_0^v + \theta_1^v c_{i1} + \theta_2^v c_{i2} + \ldots + \theta_K^v c_{iK}, \tag{3.9}$$

where $\{c_{i1}, c_{i2}, \ldots, c_{iK}\}$ are i's latent characteristics, and $\{\theta_0^v, \theta_1^v, \theta_2^v, \ldots, \theta_K^v\}$ are parameters. So the individual-venue-specific α_i^v is modeled as the inner-product of the individual-specific c_i vector and the venue-specific θ^v vector. The vectors of latent features (the c_i vectors) are going to be uncovered by factorizing the social network graph matrix, and the vectors of parameters (the θ^v vectors) are to be estimated in regression.

We use the technique NMF to uncover the c_is. Originally developed by Lee and Seung (1999) for image processing, NMF is popularized in the

area of recommendation systems. Mathematically, the adjacency matrix G is approximated by the product of a pair of matrices C (dimension $N \times K$) and H (dimension $K \times N$):

$$G \approx C \cdot H,$$

where neither of C and H are allowed to have negative elements and K is typically chosen to be much smaller than N. The non-negativity constraint leads to an interpretation that, in row i, elements $c_{ik}, k \in \{1, 2, \ldots, K\}$ is i's loading in the kth "community" or "interest group" (Zhang et al. 2007). Then the preference of each individual may be viewed as being a composition of prototypical preferences in clusters of users bound by interests or community.

Operationally, we choose K to be five[12]. The computation is carried out by applying the standard procedures in Lee and Seung (1999). The correlations among these c_{ik}s and between c_{ik}s and the other time-independent covariates are also shown in Table 3.2. Model 1b in Table 3.5 shows the new result when we control the unobserved heterogeneity by including the c_{ik}s and allowing their slopes to be different across venues. Comparing it with model 1a, we find that although the magnitude and the z-score decrease as expected, the δ coefficient is still estimated to be significantly negative with the 99.9% confidence level.

[12]It is a tradeoff between richness of information and heaviness of computation task. We also tried $K \leq 10$, and the key results did not change.

Weighting by Proximity. As was mentioned earlier in the section, an alternative explanation of the surprising result of a negative δ is that the widely followed assumption that a potential visitor weighting all friends' endorsements equally is not a satisfactory modeling choice.

The sociological branch of the innovation diffusion literature has long pointed out the need of weighting person-to-person influences according to the specific relationships. Granovetter (1978) cautioned that friends' roles might be important in forming collective behavior, saying "the influence any given person has on one's behavior may depend upon the relationship." Burt (1987), in studying medical innovation, formally defined a weight w_{ji} to be "the extent to which person i defines the social frame of reference for i's evaluation" (p. 1295). Even though here we narrow our attention to only friends, a close friend's endorsement may still insert a greater influence than a relatively distant one.

Hence, we explore weighting the endorsements by their senders' "proximity" to the focal individual. If we had more data about user interactions (e.g., online conversations), we would be able to measure the proximity of two users by looking at the frequency and intensity of their interactions. However, we observe only the binary connection patterns, so whatever proximity measure we use should be inferred from the adjacency matrix G. Counting the graphic distances between nodes does not apply here, because all influencers, being friends by definition, have a graphic distance of one to the potential visitor. Instead, we compare

social neighborhoods to infer the closeness between two persons. The proximity between user i and user j is measured by the number of users who are friends of both i and j, divided by the number of users who are friends of either i or j. Mathematically,

$$p_{ij} = p_{ji} = \frac{|V(i) \cap V(j)|}{|V(i) \cup V(j)|} = \frac{|W(i) \cap W(j)|}{|W(i) \cup W(j)|}, \tag{3.10}$$

where the interchangeability of i and j, and V and W results from the symmetry of G. This measure is usually called *common neighbors* proximity measure, and is widely used in social network analysis and link prediction literature (Liben-Nowell and Kleinberg 2007). The measure originates from the sociology concept of *the strength of the personal tie*: The stronger two persons' social tie is, the larger the overlap of their friendship circles (Granovetter 1973).

Adopting this proximity measure, we let the weight of j's endorsement, from the perspective of user i, be proportional to p_{ij}. Then, $\hat{g}_{ij}$ in equation (3.2) can be rewritten:

$$[\hat{g}_{ij}] = [\frac{g_{ij}p_{ik}}{\sum_{k=1}^{N} g_{ik}p_{ij}}] \text{ for } \sum_{k=1}^{N} g_{ik}p_{ik} \neq 0. \tag{3.11}$$

The denominator is the sum of proximity over all influencers, so it still holds that the total weights on influencers is the same across the individuals. With (3.11), the social effect now is captured by a new "weighted" number of endorsements (still normalized)

$$\hat{r}_i^{w-1} = \frac{\sum_{j=1}^{N} g_{ij}p_{ij}\hat{y}_j^{w-1}}{\sum_{j=1}^{N} g_{ij}p_{ij}}. \tag{3.12}$$

$\hat{r}_i^{v,w-1}$, as is $r_i^{v,w-1}$, is in range $[0,1]$. We include $\hat{r}_i^{v,w-1}$ into the regression model and the estimation results are shown in the second (using only $\hat{r}_i^{v,w-1}$) and third (using both $r_i^{v,w-1}$ and $\hat{r}_i^{v,w-1}$) columns in Tables 3.4 and 3.5. Again the a models in Table 3.4 are the ones in which the unobserved heterogeneity is left in the disturbance, and the b models in Table 3.5 are the ones in which we include individuals' latent features.

Comparing the results of models $1b$, $2b$, and $3b$ (which is also true for $1a$, $2a$, and $3a$), we find that the coefficient of the weighted number of endorsements is estimated to be positive at the 0.1% significance level. Moreover, the absolute value of z-score is larger for the proximity-weighted number than for the unweighted number, and the pseudo-likelihood is also larger in model $2b(a)$ than in model $1b(a)$. This result, we hence conclude, supports that the proximity-weighted number of endorsements is a better predictor of the likelihood of visiting. Using the model $2b$ estimates and evaluating the covariates at their median values, we find that increasing $\hat{r}_i^{w-1}$ from 10% to 20% causes the visiting probability to increase from 0.240% to 0.266%, a 10.8% change in percentage.

Across all models, we find consistent support for a positive weekly-specific effect — a trend proxied by the number of venue-visits by the users who are neither already visitors nor risk-set members. All of the time-independent covariates that measure a user's network stature are found to be significant with the 99.9% confidence level. Specifically, we

find a non-monotone number-of-friends effect, although for most of our observations, the result indicates a higher likelihood of visiting for an individual with more connections. The coefficient of the betweenness measure is also positive, meaning that individuals acting as a local bridge between communities are more likely to visit the venue, everything else being equal. As is expected, the signs of $s_{cc,i}$ and $l_i s_{cc,i}$ are estimated to be opposite in Table 3.5.

3.5.2 Robustness

In this subsection, we deviate from equation (3.8) to check the robustness of our result on $\hat{r}_i^{v,w-1}$.

Venue-by-Venue Estimation. By pooling the venues to estimate equation (3.8), we implicitly assume that the parameters β, γ, and δ are the same for different venues. To show that our results on r and $\hat{r}$ are not driven by only a small number of venues in our sample, we now relax this restriction and estimate the model venue by venue, allowing β, γ, and in particular δ to vary across venues. As in Table 3.5, we include individuals' latent features to capture α_i^v.

In each of the two plots in Figure 3.1,[13] we show the 50 z-values of the estimate of δ, the coefficient of $r_i^{v,w-1}$ (upper plot) or $\hat{r}_i^{v,w-1}$ (lower

[13] Again we have tried $K \in \{3, 4, \ldots, 10\}$, the dimension of the latent feature space. In Figure 3.1, we only report the case $K = 5$. The systematic difference we want to show is robust for different Ks.

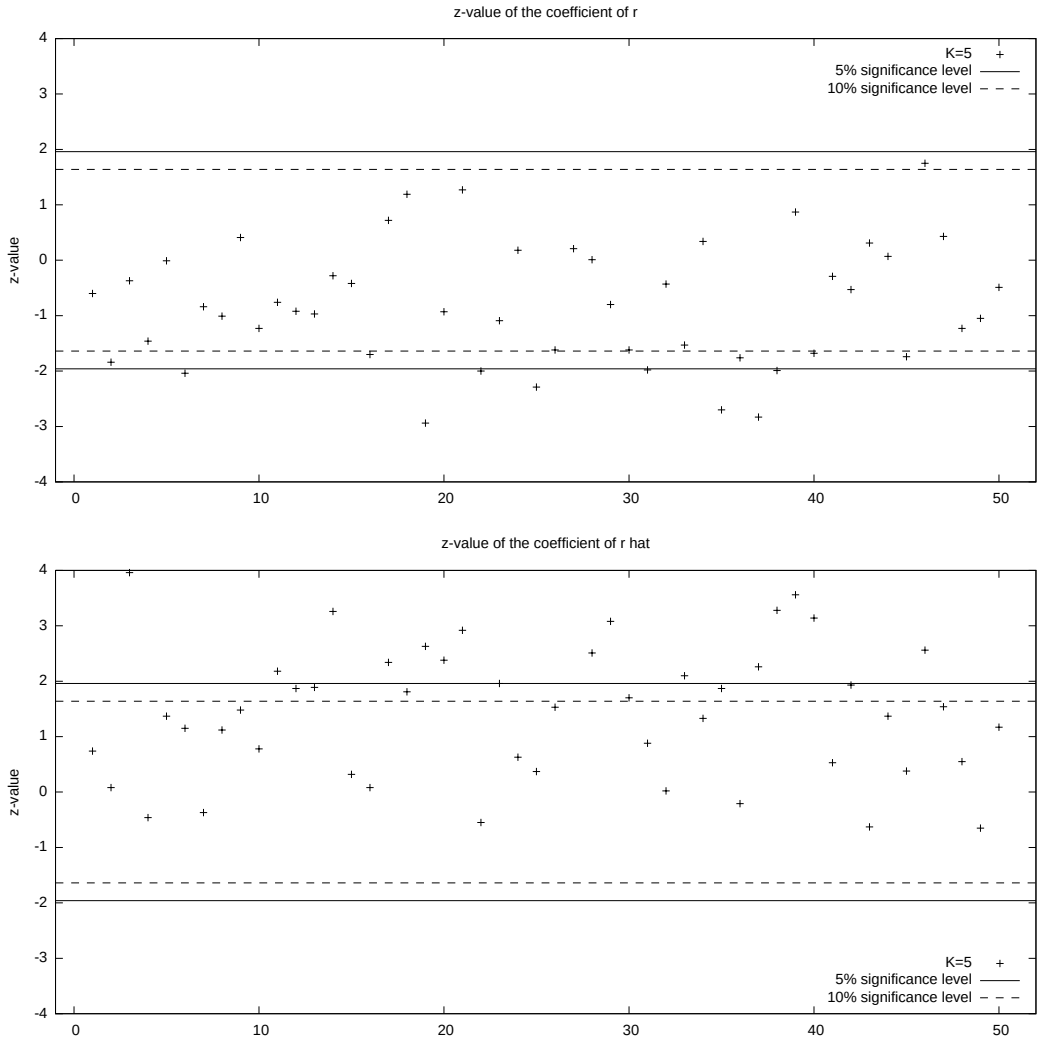

Figure 3.1: z-values of Coefficient δ: r vs. $\hat{r}$, Venue-by-Venue Estimation

plot), corresponding to the 50 venues in our sample. The dashed line is at ±1.64, corresponding to the 90% confidence level; the solid line is at ±1.96, corresponding to the 95% confidence level. We find no evidence of a positive social effect in the upper plot, where we use the unweighted $r_i^{v,w-1}$: 36 of the estimates are insignificant at the 90% confidence level; only one z-value is greater than 1.64; the remaining 13 are smaller than -1.64. However, in the lower plot, where we use the proximity-weighted $\hat{r}_i^{v,w-1}$, none is significantly negative and about a half (23) are positive at the 10% significance level, among which 17 are significant at the 5% level. Thus, the systematic difference shows that our result is not driven by a small number of "abnormal" venues.

Repetition Effect and More Influence Variables. By using $\hat{r}_i^{v,w-1}$ to capture the effect of the whole history of friends' past check-ins, we ignore the fact that a visitor-friend can check-in a venue multiple times. Presumably, that a visitor-friend checks-in more than once indicates positive outcomes from her earlier visits and represents a stronger endorsement to the venue. We call this effect the *repetition effect*. Additionally, although $\hat{r}_i^{v,w-1}$ incorporates the local unequal, person-to-person influences, we do not take into account the visitor-friends' different network statures, which may also lead to different endorsement effects. In this subsection, we extend our regression model by including more variables into equation (3.8) as additive components to test the existence of these effects and the robustness of our key result on the coefficient of $\hat{r}_i^{v,w-1}$.

These additional variables are the total number of check-ins made by friends up to week $w-1$ (*repetition effect*, $m_i^{v,w-1}$), the density of friendships among visitor-friends (*clustering effect*, $c_i^{v,w-1}$), the product of the number of visitor-friends and clustering effect ($a_i^{v,w-1}c_i^{v,w-1}$), and also the average number of friends, the average betweenness and the average clustering coefficient of the visitor-friends. Many of the variables that characterize the influencers' network statures have been discussed and used in Katona et al. (2011). Four different specifications (using either a subset or all of the additional variables) are estimated, and the results are reported in Table 3.6. Again the link function is the complementary log-log function, and standard errors are computed to be robust to venue-clustering. Unobserved heterogeneity is dealt with in the same way as in the models in Table 3.5.

The coefficient of our primary interest, δ, stays significantly positive. The magnitudes of these estimates in Table 3.6 decrease significantly from Table 3.5, indicating a high correlation between $\hat{r}_i^{w-1}$ and the additional variables that measure the *repetition effect* and the effects of the visitor-friends' network statures. Across models 4, 6, and 7, we observe a significantly positive *repetition effect*: More check-ins or endorsements made by friends increase the likelihood of visiting, while holding the number of unique visitor-friends constant. Thus, it is consistent with our intuition that multiple check-ins indicate positive outcomes, resembling a word-of-mouth effect. In model 6, we find an interesting but slightly coun-

Probability of Visiting		Model 4	Model 5	Model 6	Model 7
		Coeff.	Coeff.	Coeff.	Coeff.
		(z-value)	(z-value)	(z-value)	(z-value)
Normalized #	Weighted: $\hat{r}_i^{w-1}$	0.81***	0.23**	0.17*	0.18*
of Endorsements		(11.35)	(2.76)	(2.03)	(2.14)
Time Trend	Weekly Trend: o^w	0.02***	0.03***	0.03***	0.02***
		(7.10)	(7.33)	(7.33)	(7.34)
Time-independent	N of Friends: l_i	28.93***	17.25***	17.42***	17.17***
Covariates	(1/1,000)	(14.38)	(8.52)	(8.46)	(8.45)
	N of Friends^2: l_i^2	-0.45***	-0.30***	-0.32***	-0.31***
	(1/1,000)	(-7.35)	(-6.03)	(-6.16)	(-6.20)
	Betweenness: $s_{bw,i}$	0.78***	0.54***	0.57***	0.56***
	(1/1,000)	(6.28)	(5.05)	(5.22)	(5.24)
	Clustering: $s_{cc,i}$	-1.60***	-1.00***	-0.96***	-0.96***
		(-15.64)	(-8.69)	(-8.29)	(-8.06)
	N $\times$ Clustering: $l_i s_{cc,i}$	0.21***	0.13***	0.14***	0.14***
		(13.01)	(6.38)	(6.70)	(6.74)
Additional	Total check-ins: m_i^{w-1}	5.82***	3.13	3.45*	3.42*
Variables	(1/1,000)	(4.72)	(1.85)	(2.03)	(2.04)
	Clustering: c_i^{w-1}		-1.50***	-1.50***	-1.49***
			(-13.30)	(-13.17)	(-13.11)
	D $\times$ Clustering:		0.43***	0.45***	0.44***
	$a_i^{w-1} c_i^{w-1}$		(6.60)	(6.78)	(6.61)
	Avg. N of Friends			-2.68***	2.13
	(1/1,000)			(-3.87)	(1.01)
	Avg. Betweenness				-0.05*
	(1/1,000)				(-2.10)
	Avg. Clustering				0.01
					(0.02)
N of Observations		690,896	690,896	690,896	690,896
Pseudo Log-Likelihood		-15,905.49	-15,666.20	-15,644.89	-15,635.67

Table 3.6: Results of Complementary Log-Log Regressions: Part III

terintuitive result: The coefficient of the average number of friends for the group of visitor-friends is negative with the 99.9% confidence level, meaning that individuals with more connections have less influential power on a particular neighbor. A similar result is also reported in Katona et al. (2011). However, when we include the average betweenness and the average clustering coefficient (model 7), the coefficient of the average number of friends becomes insignificant.

One of our findings that contradicts Katona et al. (2011) is that the clustering effect is estimated to be significantly negative in all specifications. To interpret this result, a graphic example is given in Figure 3.2. Potential visitors i and j both belong to three communities. In both cases, three friends (red/filled nodes) have sent endorsements. The influencers are otherwise identical except that in the i case, the three belong to three different communities, and in the j case, all of the three belong to one same community. Therefore, in the j case, the visitor-friends are more clustered. Note that both the unweighted and weighted number of endorsements (r_i^{w-1} and $\hat{r}_i^{w-1}$) are the same in the two cases. Our result predicts that i is more likely to adopt. Our interpretation is that when a new behavior is confined to a highly intra-connected community, it might impede the outsiders from adopting it (Burt 2005). An individual, such as i, has a higher likelihood to adopt a new behavior when she can learn about the behavior from more diverse sources.

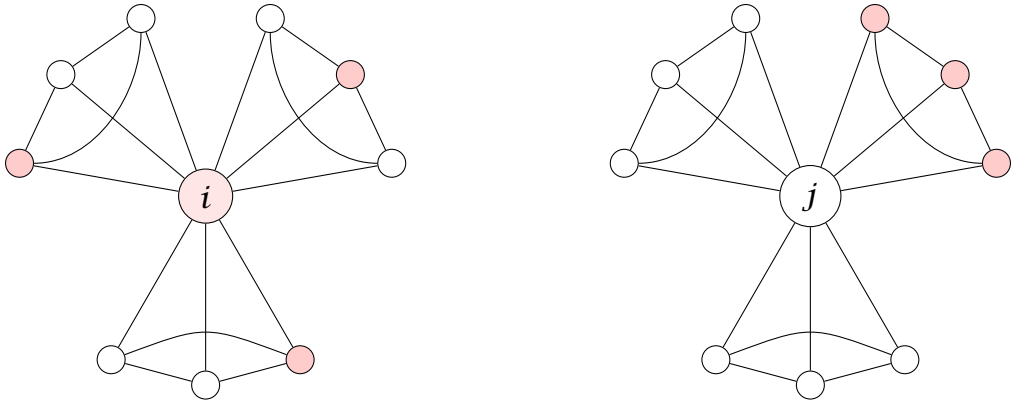

Figure 3.2: An Illustration of Negative Clustering Effect

3.5.3 Implications

Using the location-based social network as our context, we find that earlier check-ins, as a kind of weak endorsement from friends, have a significant social effect on network neighbors' decisions of visiting a venue. Our work sheds light on the economic value of the location-based social networks, one of the most popular genres of online community. Consumers spend a great deal of time and money searching for products and services that meet their tastes and needs. In many markets, the product space is so large that a complete search is very costly; thus the consumers are often unaware of or are poorly informed about a substantial portion of the available choices. For markets of *experience goods*, this problem can be especially severe because consumers are not perfectly certain about their preferences before consumption. We argue that these new technologies facilitate people's search for venues, such as restaurants and night-

clubs, by allowing them to conveniently observe and learn from network neighbors' choices. Our finding serves as evidence that advocates the development of the recently-emerged social features and their integration with traditional business models. Better utilizing these new social features, businesses may to a certain extent access a larger customer-base with lower costs. Unlike much previous research on individual adoption in an electronic setting (e.g., downloading software from the Internet in Duan et al. 2009) in which the influence variable is usually the *total number* of previous adopters, we look at the finer structure of personal relationships. From the observational learning perspective, the finding of unequal social effects of the friends' endorsements indicates that individuals tend to learn from more closely related social neighbors. This kind of learning is socially beneficial if the "similarity" of individuals' private valuations of the activity is positively correlated with their proximity. Otherwise, learning from closely-related neighbors will produce suboptimal outcomes and, presumably, cause the network to restructure. Moreover, the weights on endorsements are constructed according to proximity, which in turn is measured by the strength of the social tie between individuals. Hence, our result also sheds light on the different roles that strong and weak ties play in a social network: Although weak ties — spanning long distance and bridging tightly-knit groups of people into a "small world" — are very powerful to convey awareness of new things (Granovetter 1973), strong ties are often more important in carrying trust (Bapna et

al. 2012), and influencing and shaping certain new behavior (Centola and Macy 2007).

An ongoing trend in the domain of Internet search is to incorporate individual identity and social relationships into the methodology that determines search results. Microsoft's Bing has recently introduced the so-called "social search" feature, which displays a personalized list of Facebook friends' "likes" side by side with the generic, organic list. For example, when a Bing user types in a query "restaurants in Austin, TX," in addition to a list of popular restaurants in the city based on the wisdom of the crowd, she may also see a second, personalized list, generated based on her Facebook friends' "likes." Then the question arises: How should the friends' "likes" be ranked? Our finding of unequal local influences suggests that the effectiveness of the social search results may be improved by incorporating the proximity of individuals into the ranking algorithm — for example, weighting each friend's "like" by his proximity to the focal user. For marketers who want to harness the power of the new social technologies, our finding also has a clear implication: The proximity of users should be taken into account when using statistical models to optimize marketing effort. Admittedly, more data about detailed interactions among individuals should always be helpful. However, our result shows that, when obtaining additional data is too costly, mining out information about proximity and common interests embedded in simple binary connections can prove to be fruitful as well. Moreover, our results on the

effects of a potential adopter's own betweenness and the level of influencers' clustering also have important implications for the allocation of marketing effort by firms. If possible, marketing activity should be best targeted at a group of diverse, well-connected individuals who play important local brokerage roles. Not only do they have a greater likelihood to adopt, but the diverse influencing structure created can potentially increase the overall adoption rate, through a larger multiplier effect.

3.6 Conclusion

In the context of location-based social networks, we studied the micro structure of the endorsement effect of network neighbors' check-ins. We specified our model at the individual level. Heterogeneous rational agents make decisions on whether or not to visit a venue. The perceived utility of visiting is allowed to be affected by their social network neighbors' check-ins. The modeling assumption that a potential visitor puts equal weights on her social neighbors reduces the social effect to a single measure, the normalized number of unique endorsements received. Our empirical results show that this number is a bad predictor of the likelihood of visiting, a result that contradicts both the intuition and early theoretical models in the diffusion literature. This result, although it might be a special case pertaining to our dataset, suggests that in modeling the social effect, even when the researchers observe only the binary connection patterns among individuals, treating every neighbor's influence as

the same can be a dangerous modeling choice. We suggested that a more detailed relationship between each connected pair be considered — for example, the "proximity" of users implied by the observed connection patterns. We showed that weighting the influencers by a parsimonious proximity measure can yield a more accurate result. The weighted number of endorsements better captures the structure of the social effect. Thus, our result supports the tradition in sociology literature that person-to-person relationships should be examined deeply and modeled differently in studying individual adoption. Our findings also shed light on the economic value of the popular location-based social networks. We argue that these new technologies facilitate people's search for venues, such as restaurants and nightclubs, by allowing them to conveniently observe and learn from network neighbors' choices, and our finding serves as evidence that advocates their development and other recently-emerged similar social features and their integration with traditional business models. In dealing with the endogeneity problem caused by unobserved heterogeneity, we innovatively applied the machine learning technique nonnegative matrix factorization to uncover the agents' latent features to proxy the fixed effect. Again, this suggests that marketers should dig more deeply into the network graph, when obtaining more data is too costly.

Our study has a number of limitations. First, as we discussed in the data section, we had only one snapshot of the social network graph, and we assumed it to be fixed over the period of study. If some of the

relevant friendships were formed after the venue-visits were made, then noises would exist in the computation of the time-independent covariates and the measurement of the number of endorsements. Second, we also equated the number of reported visits (check-ins) with the number of "true" total visits, implicitly assuming whether or not a visitor checks-in is random. There are indeed many arguments the reader can employ to dispute this assumption. However, even if the assumption weren't valid, it would not cause a severe problem. After all, we can simply redefine the new behavior to be "visit plus check-in" rather than just "visit." Third, we did not observe any demographic characteristics of the users or information about the venues, which undoubtedly limits our ability to better model the cost of venue-visits and construct more influence measures. Fourth, we did not observe the types of the venues. It would be interesting to investigate whether a systematic difference exists in the structure of the social effects for different types of venues (e.g., restaurants vs. shopping centers). Fifth, we focused only on the endorsements' effect on new visitors. One interesting direction for future research would be to examine whether and how the social effect on already-visitors would differ. Lastly, in modeling user behavior, we did not consider strategic interactions among them, which would definitely behoove researchers to investigate.

Chapter 4

A Graph Based Network Influence Measure

4.1 Introduction

We investigate the measurement of individuals' network influence by modeling their adoption decision-making in the presence of a social network. Individuals' adoption of a certain new product, idea, or behavior is one of the most commonly observed phenomena in social networks. For example, users of social broadcasting networks decide whether to use a "hashtag" in their posts; members of location-based social networks decide whether to "check-in" a particular venue. In making these adoption decisions, people may be affected by the choices of their social network neighbors, because of reasons such as social conformity, network externality, observational learning, and word of mouth. One user's adoption may exist only as an isolated event, or it may cause a large subsequent cascade of adoptions by others. In this chapter, we consider an individual "influential" if a change in her own state is expected to have a significant effect on the overall outcome in the network. Based on this definition, we propose a model-based network influence measure. As is the case in the preceding two chapters, we assume that we do not observe any in-

dividual (demographic or socioeconomic) characteristics. Therefore, the influence measure is purely based upon the social network structure, i.e., the (binary) network graph matrix.

Conceptually, a model-based method of scoring social network members in terms of their influence can serve as a foundation to characterize the distribution of influential power among the population and answer interesting questions such as given the structure of a network, whether the influential power is concentrated to a relatively small set of individuals or shared by a large proportion of the population. Perhaps more importantly, a well-founded ranking method can prove to be useful for practitioners to solve several critical problems in related areas such as advertising and search. For example, the current practice in targeted advertising focuses on the consumers' demographic (e.g., gender) and socioeconomic (e.g., income) characteristics. In the presence of a social network where consumers interact and potentially influence each other, it has been recognized that the targeted individuals' network positions should also be taken into account, in order to maximize the effect of a marketing effort (Kempe et al. 2003). Related, from the network operators' perspective, a methodology of scoring and ranking the network members' influence is important for designing a better pricing scheme for their advertising products. For example, Facebook allows businesses to promote their pages or products by highlighting stories authored by Facebook users about their consumption experiences at the businesses —

so called sponsored stories in the Facebook terminology. Theoretically, everything else being equal, the overall effects of two sponsored stories could be very different if the story authors' network positions are different. Thus, story authors' network influence should be an important factor in determining the price of sponsored stories. In the domain of Internet search, an ongoing trend is to incorporate the information of individual identity and social relationships into the algorithm that determines search results. For instance, Bing, developed and operated by Microsoft, has recently introduced the so-called "social search", which displays a personalized list of Facebook friends' "likes" side by side with the generic, organic list.[1] Google, the top player in the search domain, has employed a similar strategy, called "search plus your world," by promoting its homegrown social network, Google+, and integrating it with Google Search and other products. In light of this change, an influence measure can potentially be used in ranking the social search results, and may even shed light on how to better integrate them with the existing, organic results.

Sociology researchers have developed several methods of measuring the "importance" of members in a network. The earliest and the simplest measure is degree centrality, which is defined as the number of links (or in-links in a directed graph) a node has. For example, in the Facebook network it is the number of friends a user is connected to and on Twitter it is the number of followers a user has. For the purpose of measuring

[1] http://www.bing.com/new?publ=BNPHP&crea=HSC.

influence, the degree centrality is perhaps too simple since (1) it does not embody the transitivity of influence and (2) it treats every connection or follower as the same. In fact, using data of online social networks, empirical researchers have found that great popularity does not necessarily lead to high influence (e.g., Bakshy et al. 2011). As a generalization of degree centrality, Katz centrality (Katz 1953) counts not only the number of immediately connected nodes, but also nodes that can be connected through a path, with the contribution of each node decreasing exponentially with its distance to the focal node. However, Katz centrality still treats the nodes that are equally distant to the focal node as the same. This is not satisfactory because, for example, in Chapter Two we find that in the Twitter context each follower's "attention" to the focal user decays as the follower's number of followings increases. This finding suggests that in addition to the graphic distance to the focal node, the connected nodes' other social network characteristics, such as the number of outlinks, should also be considered in building a model for measuring influence. Betweenness, first proposed in Freeman (1977), measures the likelihood that a node appears on a randomly chosen shortest path between two randomly chosen nodes. The idea is that the communication between two persons in a social network is likely to happen along the shortest path between them, and the individuals located on the shortest path have a great "power" of controlling the communication. While these graph-based measures are intuitive and easy to implement, they all lack a

well-defined behavioral foundation: After all, influence means the ability to affect others' decision making.

A related area in computer science is the ranking of webpages in the World Wide Web (www). Kleinberg (1998) proposed a model by observing that, with respect to a certain topic, two types of webpages generally exist: *authorities*, the webpages that provide quality and authoritative information on the topic, and *hubs*, the webpages that have out-links to many authorities. He developed an algorithm that identifies the two types of pages simultaneously and iteratively scores the relevant webpages as an authority and as a hub. He proved that the authority and hub scores converge to their respective limits and suggested the limiting authority scores be used to rank the webpages. To tackle the same problem, Brin and Page (1998) proposed a model without the distinction between hubs and authorities. In their model, the "quantity of the authority" of a webpage is partly passed onto other out-linked webpages and partly distributed uniformly to the whole www. They showed that an equilibrium of the authority distribution exists and the resulting method is the famous PageRank algorithm, based on which the Google search engine operates. Since the PageRank algorithm relies only on the link relationships between the webpages, i.e. the graph structure, it can also be applied to the social network context. Indeed, Google has apparently adapted it to the problem of ranking social network users and filed a patent application (Green 2008). Weng et al. (2010) suggested applying PageRank to measure

Twitter users' influence. Later in this chapter, we discuss the connection between our measure of influence and the PageRank algorithm.

We organize the rest of this chapter as follows. We start from a parsimonious probabilistic model of individual behavior in Section 4.2, taking into account the *local, person-to-person* influence, as is the case in Chapter Three. We then develop the model and show how the person-to-person influences aggregate to the *global, network-level* influence. In Section 4.3, we show that our measure of influence admits the famous PageRank measure as a special case. In Section 4.4 we conduct numerical experiments by simulating Watts-Strogatz networks and the goal is to demonstrate that the higher degree of freedom of our measure enables it to capture a potentially richer structure of the influence distribution among a population. Lastly, in Section 4.5, we conclude and point out future research directions.

4.2 Model

Suppose that a certain act (e.g., choosing a hotel, going to a new restaurant, purchasing an innovative device, spreading a piece of informational content) can be adopted by the members of a social network. The binary outcome, denoted $y_i = 1$ or 0, indicates whether the act is taken by member i, $i \in \{1, 2, \ldots, N\}$, where N is the size of the network.

We assume that the influence relationships among the network

members are observed. The basic influence structure, corresponding to the topic of interest, can be represented by an $N \times N$ network influence matrix G. The elements of G are specified as follows:

$$g_{ij} = \begin{cases} 1, & \text{if there is a directed link from } i \text{ to } j; \\ 0, & \text{otherwise,} \end{cases} \qquad i, j = 1, 2, \ldots, N.$$

g_{ij} indicates whether or not member j can influence i on the act of interest. G does not have to be symmetric, since the influence relationship can be nonreciprocal. We write $V(i) = \{j | g_{ij} = 1\}$, the *influencers* of i; and $W(i) = \{j | g_{ji} = 1\}$, the *influencees* of i. In practice, the influence matrix G should be constructed according to the specific act in consideration and not necessarily coincide with the network adjacency matrix derived from the actual "friendships" or "following-follower relationships."[2]

We adapt Richardson and Domingos (2002)'s model of viral marketing to formulate a parsimonious probabilistic model of individual adoptions. Suppose that, from the perspective of the researchers, there exists a baseline probability for member i to take the act, $i \in \{1, 2, \ldots, N\}$. We also call this baseline probability the internal probability, since it would be the likelihood of i adopting the act if i lived in isolation. We allow the internal probability to vary across the individuals to take into account the fact that they could possess heterogeneous characteristics that are related to adopting the act, but cannot be captured by the network structure (e.g.,

[2] So ideally, even for the same group of individuals, when we consider different acts, the corresponding influence matrices should be different.

the individual-level demographic or socioeconomic characteristics used in binary choice econometric models), and also that they might be informed differently about the act (e.g., different prior knowledge sets, different levels of ad exposure). These factors are unobserved by the researchers, so we use a single parameter p_i^0 to capture the internal probability.

In the presence of a social network, i's decision may be affected by her influencers, through the mechanisms such as social conformity, network externality, observational learning, word of mouth, and so on. We abstract away from these specific micro-level mechanisms, and assume that i's influencers' acts y_js, $j \in V(i)$, affect the probability that i adopts through a moderating function δ_{ij} — that is, δ_{ij} determines the structure of the local, person-to-person influence. The primitive δ_{ij} is indexed by both i and j, so the person-to-person influence can vary across different pairs of individuals. We further assume that the external influence to be an additive component in the adoption likelihood function. To sum up, the key assumption of the network members' adopting behavior is given by the following conditional probability equation

$$\begin{aligned} \mathrm{p}(y_i = 1|y_{-i}) &= \mathrm{p}(y_i = 1|y_j, j = 1, 2, \ldots, N, j \neq i) \\ &= \mathrm{p}(y_i = 1|y_j, j \in V(i)) \\ &= \beta p_i^0 + (1 - \beta) \sum_{j=1}^{N} g_{ij} \delta_{ij}(y_j), \end{aligned} \tag{4.1}$$

where β and $1 - \beta$ are the relative weights network members put on their internal probability and the external influence. To be complete, we also

require the following regularity condition hold, $\forall i$,

$$\sum_{j=1}^{N} g_{ij}\delta_{ij}(\cdot) \leq 1,$$

so that equation (4.1), as a probability, is always well-defined.

To obtain the unconditional probability of adoption, we take expectations on both the sides of equation (4.1). Note that the expectation is taken with respect to y_{-i}, which is a random vector from the researchers' perspective. Then we have

$$\begin{aligned}
\mathrm{p}(y_i = 1) =& \beta p_i^0 + (1-\beta)\sum_{j=1}^{N} g_{ij}\mathrm{E}[\delta_{ij}(y_j)] \\
=& \beta p_i^0 + (1-\beta)\sum_{j=1}^{N} g_{ij}[\delta_{ij}(1)\mathrm{p}(y_j = 1) + \delta_{ij}(0)(1 - \mathrm{p}(y_j = 1))] \\
=& \beta p_i^0 + (1-\beta)\sum_{j=1}^{N} g_{ij}\delta_{ij}(0) \\
&+ (1-\beta)\sum_{j=1}^{N} g_{ij}(\delta_{ij}(1) - \delta_{ij}(0))\mathrm{p}(y_j = 1), \forall i.
\end{aligned}$$

For simplicity of notation, we write

$$p_i = \mathrm{p}(y_i = 1), \quad \bar{\delta}_{ij} = \delta_{ij}(0), \quad \delta_{ij} = \delta_{ij}(1) - \delta_{ij}(0).$$

Hence, the $p(y_i = 1)$ equation becomes

$$p_i = \beta p_i^0 + (1-\beta)\sum_{j=1}^{N} g_{ij}\bar{\delta}_{ij} + (1-\beta)\sum_{j=1}^{N} g_{ij}\delta_{ij}p_j. \tag{4.2}$$

So the unconditional probability for i to adopt the act is determined by other network members' unconditional probabilities as well as her internal probability. Note that g_{ij} is determined by the influence structure, so

once we have chosen the behavioral primitives β and $\delta_{ij}(\cdot)$, the second term on the right hand side of equation (4.2) is just an i-specific constant, which without loss of generality can be absorbed into βp_i^0. Now suppose p_i^0 can be somehow changed (e.g., through targeted advertising), and we examine how it would affect the outcome of the whole system. Replacing the subscript i with k in equation (4.2) and taking the partial derivative of it with respect to p_i^0 gives

$$\frac{\partial p_k}{\partial p_i^0} = \begin{cases} \beta + (1-\beta)\sum_{j=1}^{N} g_{kj}\delta_{kj}\frac{\partial p_j}{\partial p_i^0} & \text{if } k = i, \\ (1-\beta)\sum_{j=1}^{N} g_{kj}\delta_{kj}\frac{\partial p_j}{\partial p_i^0} & \text{otherwise.} \end{cases} \tag{4.3}$$

Thus, (4.3) is a system of equations that determine the relationships among the $\frac{\partial p_i}{\partial p_j}$s. Because of the additive form of equation (4.1), the system does not depend on the p^0s. Note that even if individual k is not directly influenced by i ($g_{ki} = 0$), a change of p_i^0 can still cause p_k to change as long as some of the $g_{kj}\delta_{kj}\frac{\partial p_j}{\partial p_i^0}$s are nonzero. The transitivity of influence is hence taken into account in this model.

At the aggregate level, the total number of adopters is $S = \sum_{i=1}^{N} y_i$, whose expectation, $\mathrm{E}(S)$, equals $\sum_{i=1}^{N} p_i$. The partial derivative

$$NI(i) := \frac{\partial \mathrm{E}(S)}{\partial p_i^0} = \sum_{k=1}^{N} \frac{\partial p_k}{\partial p_i^0} \tag{4.4}$$

thus measures how a small change in member i's internal probability would affect the expected total number of adopters in the whole network and we define it to be our measure of network influence. Since equation (4.3) does not depend on p_i^0s, the NI measure, as is defined by (4.4), is

also independent of p^0s. It can be computed based only upon the influence structure (g_{ij}s) and the behavioral primitives (β and δ_{ij}s).

Another interpretation of the behavioral primitive δ_{ij} is the weight that individual i attaches to j's influence, $j \in V(i)$, which is unobserved to the researchers. Yet, its specification is important in determining NI. As we have found in the previous two chapters, the strength of the relationship between two individuals, or the proximity between them, can have an important moderating effect on the person-to-person influence. In Chapter Two, we found that in information diffusion, voluntary content sharing happens more frequently through weak ties; by contrast, in Chapter Three we found in new behavior diffusion — for example, venue visits — strong ties carry larger social effects. Therefore, we believe δ_{ij} should be left as a free parameter in the NI measure. In practice, the researchers should specify δ_{ij} carefully according to the topic of interest.

4.3 Mathematical Examination

In this section, we examine the analytic properties of the key equations that are used to derive our measure. We also show that our measure admits a version of Google's PageRank centrality measure as a special case.

Existence and Uniqueness

Here we show that given $p_i^0, i \in \{1, 2, \ldots, N\}$, and that behavioral primitives β and δ_{ij} satisfy a certain regularity condition, a unique solution $\{p_i; i = 1, 2, \ldots, N\}$ of equation (4.2) exists, and when the p_is are viewed as functions of the p_i^0s, the partial derivatives in equation (4.3) are well-defined.

We rewrite equation (4.2) here, absorbing the second term into p_i^0,

$$-p_i + \beta p_i^0 + (1-\beta) \sum_{j=1}^{N} g_{ij}\delta_{ij}p_j = 0, \forall i. \tag{4.5}$$

(4.5) specifies a system of N equations with N unknown variables $p_i, i \in \{1, 2, \ldots, N\}$, so, informally, as long as the N equations are mutually compatible, we can solve the p_is, in terms of the primitives and p_i^0s. Moreover, if none of the equations are redundant, then the solution is unique.

We formalize it in the framework of Implicit Function Theorem. Given β and δ_{ij}, the left hand side of equation (4.5) defines a continuously differentiable function $f : R^N \times R^N \to R^N$ and $f(p_1^0, \ldots, p_N^0; p_1, \ldots, p_N) = 0$. If the Jacobian

$$[\frac{\partial f_i}{\partial p_j}] = -I + (1-\beta)G^\delta, \text{ where } G^\delta = \begin{pmatrix} g_{11}\delta_{11} & g_{12}\delta_{12} & \cdots & g_{1n}\delta_{1n} \\ g_{21}\delta_{21} & g_{22}\delta_{22} & \cdots & g_{2n}\delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1}\delta_{n1} & g_{n2}\delta_{n2} & \cdots & g_{nn}\delta_{nn} \end{pmatrix} \tag{4.6}$$

is invertible,[3] then there exists a unique continuous differentiable func-

[3] A sufficient condition is $0 < \beta \leq 1$ and δ_{ij}s are nonnegative.

tion $p : R^N \to R^N$ that

$$f(p_1^0, \ldots, p_N^0; p(p_1^0, \ldots, p_N^0)) = 0$$

and the partial derivatives $\frac{\partial p_i}{\partial p_j^0}, i, j \in \{1, 2, \ldots, N\}$, are well-defined.

Relationship with PageRank

PageRank (Brin and Page 1998), the founding stone of the Google Internet search engine, is a link analysis algorithm. It assigns a "numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of 'measuring' its relative importance within the set." PageRank can be interpreted in the following way: All web surfers start on a random page in the set of documents; at any moment, they either choose to follow a random link from the page they are currently visiting or jump to a random page in the whole set. The PageRank value of a webpage i is then the probability that a random surfer is on webpage i. Like our measure, the algorithm of RageRank is based purely on the graph, i.e., the structure of links among the documents. In this subsection we show that if we directly apply PageRank to our context, it is actually equivalent to our measure NI with a specific choice of β and δ_{ij}.

The most widely used version of PageRank should satisfy the following equation:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \to i} \frac{PR(j)}{o(j)}, \forall i, \tag{4.7}$$

where $PR(\cdot)$ is a webpage's numeric PageRank value and constant N is the size of the set of webpages. The second term on the right hand of the equation is a summation over all webpage js which have an out-link to i. $o(i)$ is the total number of out-links of i. d is a parameter called *damping factor* in Page and Brin's original paper, and is usually set to 0.85 in practice. It has been shown that, for $d \in [0,1)$, there is a unique vector of PRs that satisfies (4.7). The larger the value of $PR(i)$ is, the more important webpage i is relative to other ones.

When we abstract both webpages and network members as nodes, and both hyperlinks and influence-relationships as directed edges, a hyperlinked set of webpages and an influence network of human users are structurally the same: Both of them are graphs, i.e., a set of nodes linked by directed edges. From this perspective, the PageRank measure of importance can be directly used here in our social network context, even though its original interpretation of a random surfer clicking links no longer applies. To make the definition (4.7) consistent with our notation, note that $o(i) = \sum_j g_{ij}$ and hence we can rewrite

$$PR(i) = (1-d)\frac{1}{N} + d\sum_{j=1}^{N}\frac{g_{ji}}{\sum_k g_{jk}}PR(j), \forall i. \tag{4.8}$$

Now we show that when we set the behavioral primitives $\beta = 1-d$ and $\delta_{ij} = \frac{1}{\sum_j g_{ij}}$ for is such that $\sum_j g_{ij} \neq 0$, our measure of network influence, NI, and PageRank, PR, are mathematically equivalent. We continue using the notation G^δ in the previous section. With the specific choice

of δ_{ij}, G^{δ} is just a very simple transformation of the graph matrix G: For each row i that $\sum_j g_{ij} \neq 0$, we divide each element g_{ij} in the row by $\sum_j g_{ij}$. Simply call the transformed matrix $\hat{G}$. Below is an example of G-$\hat{G}$ pair.

$$G = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \hat{G} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

It is easy to see that, with this specification, we are assuming each influencer j of individual i imposes equal influence on i's decision of adopting y.

We represent the relevant systems of equations in matrix form. Let $E(n,m)$ be an $n \times m$ matrix whose elements are all ones. Let $PR = (PR(1), PR(2), \ldots, PR(N))'$. Then we can write equation (4.8)

$$PR = d\hat{G}' \cdot PR + \frac{(1-d)}{N} E(N,1), \tag{4.9}$$

where operator $\cdot$ is matrix multiplication and $'$ is transposition. Similarly (4.3) can be rewritten as

$$PA = \beta I + (1-\beta) G^{\delta} \cdot PA, \tag{4.10}$$

where PA is an $N \times N$ matrix whose element i,j is $\frac{\partial p_i}{\partial p_j^0}$. Replacing the primitives, equation (4.10) becomes

$$PA = (1-d) I + d\hat{G} \cdot PA. \tag{4.11}$$

Lastly the relationship between our network influence measure NI and matrix PA, as implied by (4.4), can be written in matrix form

$$NI = PA' \cdot E(N,1). \tag{4.12}$$

The key result we want to prove is $PR(i) = \frac{1}{N} NI(i), \ \forall i$. From equation (4.11) we have

$$(I - d\hat{G}) \cdot \frac{PA}{1-d} = I.$$

So $\frac{1}{1-d} PA = (I - d\hat{G})^{-1}$ (invertibility is assumed in the previous section). By the property of invertible square matrix, we have also

$$(I - d\hat{G}') \cdot \frac{PA'}{1-d} = I.$$

Multiplying both sides by $\frac{(1-d)}{N} E(N,1)$ gives

$$(I - d\hat{G}') \cdot \frac{PA'}{N} E(N,1) = \frac{(1-d)}{N} E(N,1).$$

Comparing the equation above with equation (4.9), we have

$$PR = \frac{PA'}{N} E(N,1) = \frac{NI}{N}.$$

Thus the NI values only differ from the PR values by a constant scaling factor, which is just the size of the social network in consideration. By doing this exercise, we essentially show that our model can serve as a behavioral foundation for applying the classic PageRank algorithm, which was originally developed for ranking the importance of hyperlinked documents on the Internet, to this new task of ranking network members in terms of their influence. However, as we have found in the preceding two chapters, the "equal-influence" assumption ($\hat{G}$), which is really the basis of PageRank, should be modified for different acts in consideration and we

should leave the option of choosing specific forms for δ_{ij} open for different applications. As we deviate from the "equal-influence" assumption, our model can provide more structure of influence than the PageRank measure.

4.4 Numerical Experiment

In this section, we use the Watts-Strogatz network model (Watts and Strogatz 1998) to simulate pseudo influence networks as the context to compute our measure. The goal of the numerical exercises is to show how the distribution of the individuals' network influences, as captured by the measure NI, changes, when we use different choices of δ_{ij}. Specifically, we relate δ_{ij} to the strength of the social tie between i and j, or the closeness between i and j. For each simulated network, we compute three different variants of NI. The first is based on the "equal-local-influence" assumption, under which we have shown that NI is equivalent to PageRank. For the second variant, we assume that strong ties or close relationships carry a larger person-to-person influence, as is in the venue-visit case in the third chapter. For the third variant, we assign greater weights of local influences on weak ties, to incorporate situations like information diffusion discussed in Chapter Two.

The network model we adopt here is popularized by Watts and Strogatz (1998) and we illustrate it in Figure 4.1. The population (size

N) lives on a one-dimensional lattice and each individual is connected to a small number of nearest neighbors (denoted K, following Watts and Strogatz (1998) we require $N > K > \ln(N)$) by undirected edges. So these edges represent strong relationships: Each individual influences and is influenced by his or her close neighbors. Next, for each edge, with probability p_w, we disconnect one end of the edge and reconnect it to an individual chosen uniformly at random in the population, creating the "shortcuts" in the figure. And these relationships are weak ties. Thus, p_w determines the density of these "shortcuts". In Figure 4.1, we show two examples with $p_w = 10\%$ and $p_w = 50\%$ respectively.[4] The Watts-Strogatz model, often labeled "small-world" network model, has a fairly simple structure. Yet, it resembles the real social network in possessing two important properties *when* p_w *is small*: "short average distance" between a pair of nodes (Milgram 1967) and the "high clustering" effect (Newman 2003).

To compute our NI measures, we still need to decide on β and δ_{ij}. We set $\beta = 0.15$, following the standard practice of PageRank. For the more important primitive δ_{ij}, we apply three specifications based on three different schemes of weighting influence relationships according to the strength of the relationship. We use the method of comparing the social neighborhoods to measure tie strength, as we did in Chapter Two. Specifically, our metric is the first version of the overlap index developed

[4]In both cases, we use a small $N = 50$ for clearness of visualization. In later simulations, we shall use much larger Ns.

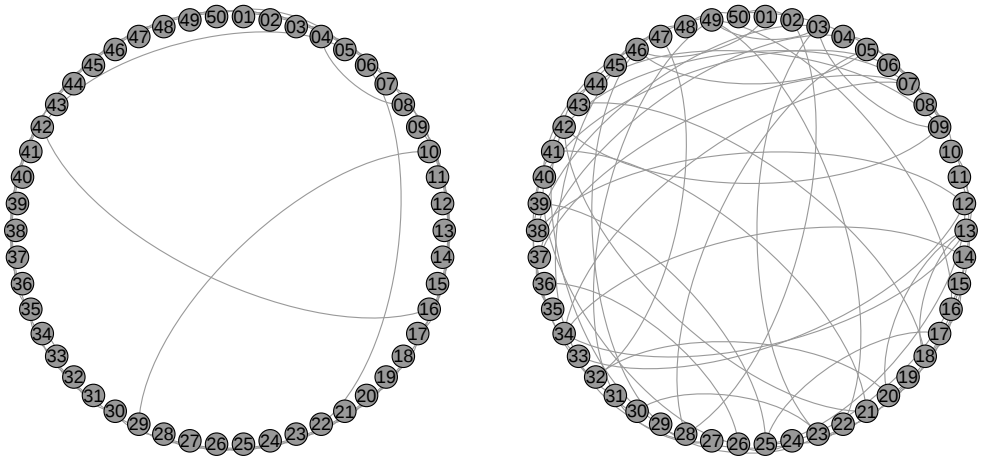

Figure 4.1: Watts-Strogatz Social Network Model

there

$$o_{ij} = \frac{\sum_k g_{ik} g_{jk}}{\sqrt{\sum_k g_{ik} \sum_k g_{jk}}}.$$

o_{ij} lies in the interval of $[0, 1]$. The higher index o_{ij} is, the more similar i and j's social neighborhoods and the stronger their relationship. The rationale of the index is the strength of the social tie theory in sociology (Granovetter 1973). Interested readers can refer to Chapter Two and its appendix. Building on index o_{ij}, our three specifications of δ_{ij} are:

1. *Equal-influence Case*, denoted N,

$$\delta_{ij}^N = \frac{1}{\sum_k g_{ik}} \text{ for } \sum_k g_{ik} \neq 0;$$

2. *Large-weight-on-strong-tie Case*, denoted S,

$$\delta_{ij}^S = \frac{o_{ij}}{\sum_k o_{ik}} \text{ for } \sum_k o_{ik} \neq 0;$$

3. *Large-weight-on-weak-tie Case*, denoted W,

$$\delta_{ij}^{W} = \frac{1 - o_{ij}}{\sum_k (1 - o_{ik})} \text{ for } \sum_k (1 - o_{ik}) \neq 0.$$

We first show that, given an influence structure, the three specifications can indeed produce different orderings of network influence. A simple influence-network example is given in Figure 4.2, where $N = 20$, $K = 4$, and $p_w = 20\%$. The three variants of the network influence measure, NI^N, NI^S, and NI^W, computed based on (4.4) and the corresponding orderings (the columns labeled "ID") are shown in Table 4.1. We find that the three variants *do* suggest different influence rankings for the 20 individuals. Take user 07 for example. In the N case, 07 is ranked at about the median; in the S case where strong-ties are designated to carry larger influence, she is ranked at the 70th percentile; and in the W case in which larger weights are put on weak ties, she is ranked at the 25th percentile. The relative high ranking of her in the S case and the low ranking in the W case should be consistent with the intuition, since user 07 influences and is influenced by four close social neighbors 05, 06, 08 and 09, and does not have any weak ties.

We now investigate how the distribution of network influence NI in a population changes with different specifications of δ_{ij}. We choose N to be 5,000 and K to be 10. Figure 4.3 shows the distribution of NIs in four typical networks corresponding to $p_w = 5\%$, 10%, 45% and 50% respectively. In the upper two subplots, p_w is very small, so the simu-

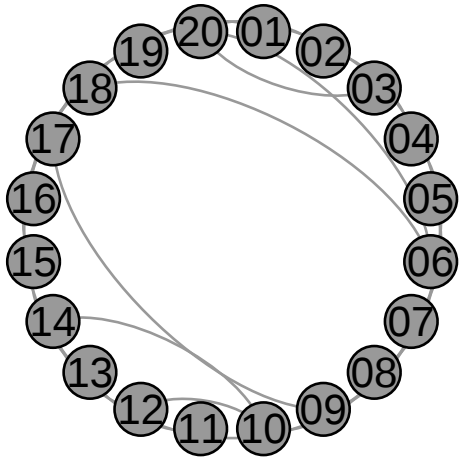

Figure 4.2: An Example of NI Orderings: Influence Network

lated network structurally resembles the "small world" networks in human society. We find that NI^S and NI^W both are more dispersed than NI^N. Moreover, while the distribution of NI^N is more or less symmetric around its mean (i.e., 1), the distribution of NI^S is negatively skewed and the distribution of NI^W is positively skewed. The difference indicates that when considering acts that more frequently transmit through weak ties, for example the dissemination of new information, in a small world where weak ties are a scarce resource, the distribution of influence skews heavily towards a small group of network members who have relatively more weak ties; by contrast, for acts that more readily diffuse through strong relationships, such as the adoption of venue-visits which we discussed in Chapter Three, the bulk of influence is shared by a majority part of the population. In the bottom two subplots, p_w is relatively large. In these networks, weak ties are no longer a scarce resource, and the distributions

Rank	N		S		W	
	ID	NI^N	ID	NI^S	ID	NI^W
1	06	1.458	06	1.359	06	1.610
2	10	1.280	10	1.255	20	1.515
3	17	1.277	17	1.159	17	1.457
4	20	1.276	02	1.144	10	1.311
5	02	1.085	20	1.122	18	1.100
6	14	1.067	07	1.098	09	1.099
7	15	1.062	15	1.072	14	1.085
8	04	1.044	08	1.063	04	1.057
9	18	1.020	14	1.061	15	1.053
10	09	1.013	04	1.038	19	1.046
11	07	1.010	18	0.973	02	0.991
12	08	1.004	12	0.959	08	0.900
13	19	0.848	03	0.958	13	0.869
14	11	0.842	09	0.950	11	0.865
15	13	0.841	16	0.939	07	0.862
16	12	0.833	05	0.875	01	0.666
17	03	0.819	13	0.827	05	0.663
18	16	0.809	11	0.827	16	0.621
19	05	0.793	19	0.720	12	0.619
20	01	0.621	01	0.600	03	0.609

Table 4.1: An Example of NI Orderings: Rank

of NI^S and NI^W become closer to each other.

For a given p_w value, we generate a large number of networks. For each of these networks, we compute the second and third moments[5] of the distributions of NI^N, NI^S and NI^W. We then calculate the mean standard deviation and the mean skewness across all simulations for a given p_w. In Figure 4.4, we plot these mean moments against p_w. From the upper subplot, we can see that weighting ties of different strength enlarges the variance of the distribution and our NI^W variant has a even larger variance than NI^S. The lower subplot shows our two different weighting

[5]The first moment, mean, is always 1, whichever variant of NI we use.

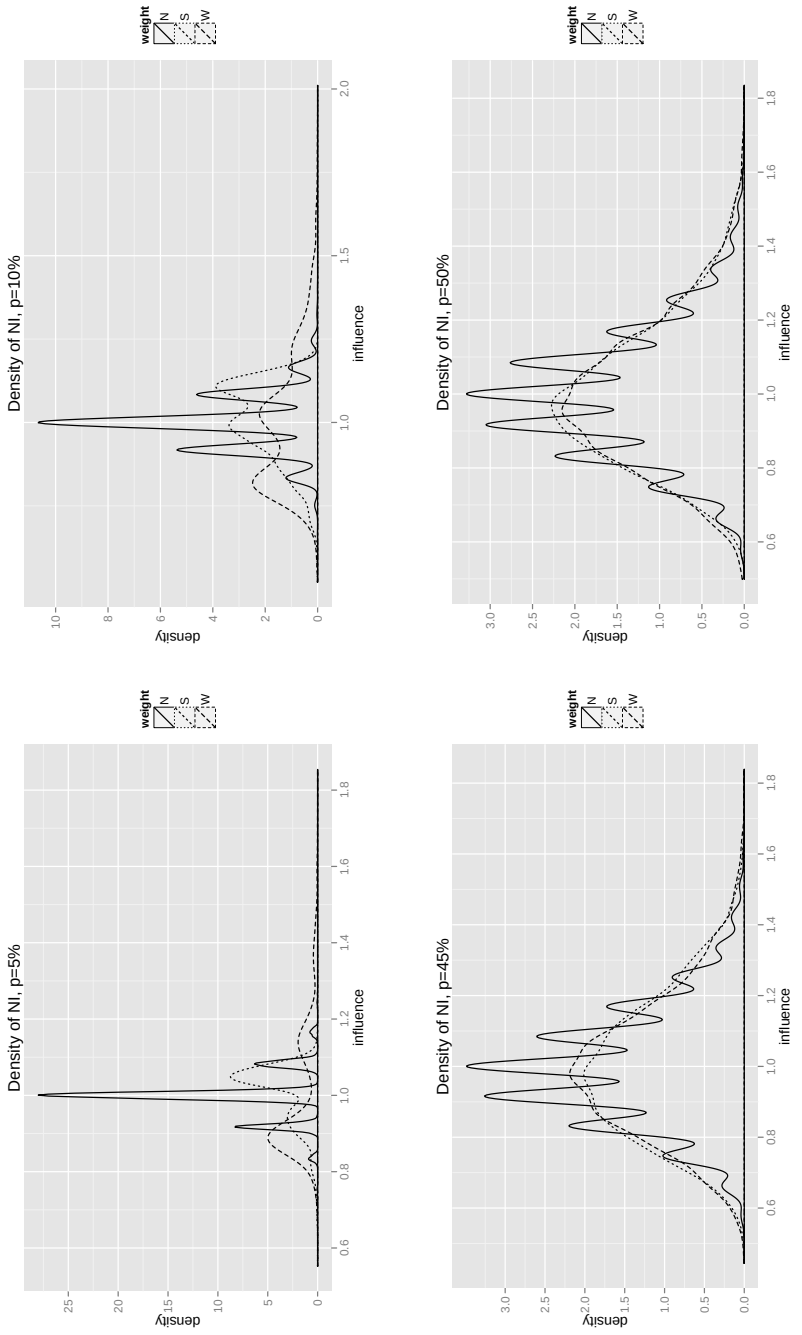

Figure 4.3: Distribution of *NI*

schemes produce opposite results on skewness. The difference of skewness is specially significant for small p_w. As p_w approaches 0.5, the skewness of the three distributions become closer. These results again show that the freedom of choosing δ_{ij} gives our measure NI more power to capture the rich structure of the distribution of influence than PageRank.

4.4.1 Extension to Item Ranking

The NI measure ranks the individuals of a social network in terms of their influence. The measure can then be used as an *influence weight* to rank items. Suppose that we have computed NI regarding the choice of movies and a total of J movies are discussed and endorsed by the network members. Let matrix A be an $N \times J$ user-movie matrix that has entries specified as follows: A_{ij} denotes the fraction of the endorsement from individual i that goes to movie j. Then, a weighted total endorsement for movie j can be obtained by $T(j) = \sum_{i=1}^{N} A_{ij} NI(i)$, based on which the J movies can be ordered.

4.5 Conclusion

In this chapter, we propose a methodology to score and rank the individuals in a social network in terms of their relative influence regarding a certain act. The measure builds on an individual-level model of adoption

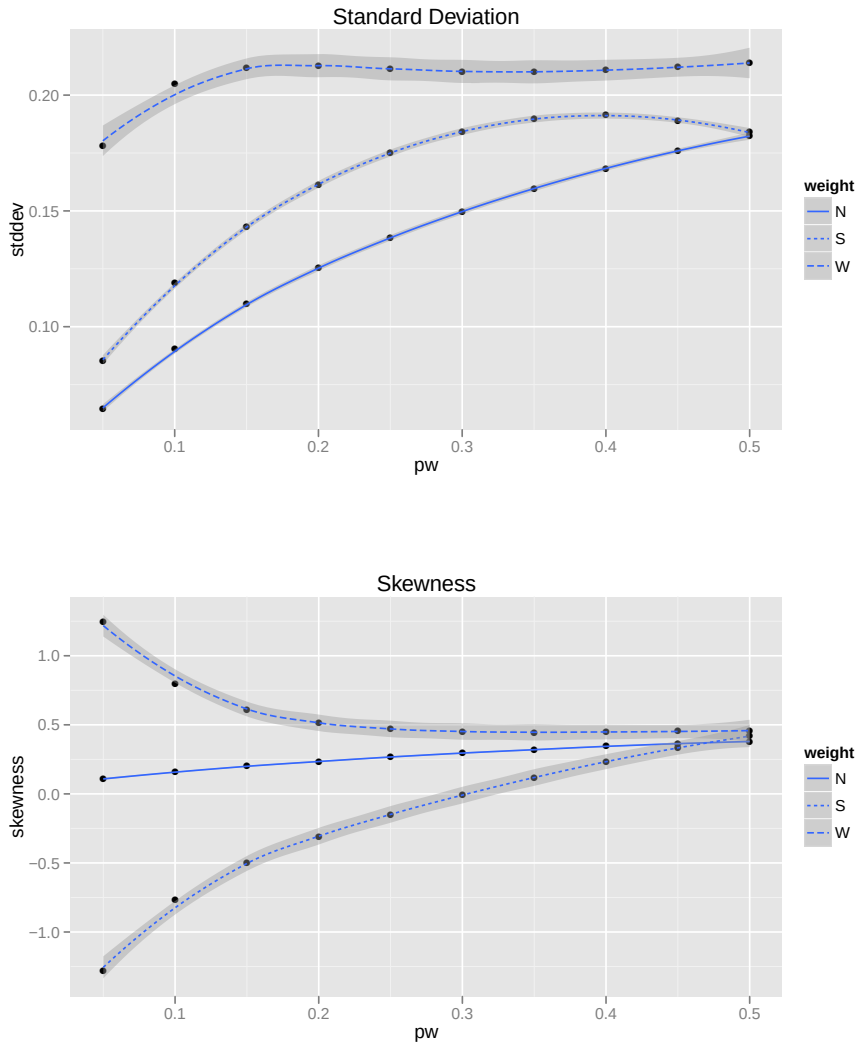

Figure 4.4: The Second and Third Moments of NI Distribution

decision-making and relies only upon the influence network structure. The proposed measure depends on two behavioral primitives, which we argue should be best determined on an application-to-application basis. We also show that with one specific choice of the primitives, our measure is equivalent to the popular PageRank algorithm. We conduct numerical experiments by simulating Watts-Strogatz networks. For each of the simulated networks, we compute the influence measure using three different specifications of the free primitives, which are inspired by our findings in the preceding two chapters. We show that our measure is indeed able to capture a richer structure of the influence distribution in a networked population.

Our study has a number of limits and they also shed light on possible future research directions. First, in our model, the primitive β is assumed to be exogenous and identical across all individuals. One interesting research direction is to explore the possibility of relaxing this restriction and endogenizing β, probably by allowing β to be correlated with the number of one's influencers. Indeed, individuals who put a larger weight on themselves (i.e., larger β) may *choose* to have fewer influencers. Second, we assume the primitives are given by the researcher or practitioner who implements the measure. One future research question would be, if a researcher observes the outcomes y_is, then can the primitives be estimated from the data? Third, the three specifications of δ_{ij} we have experimented are constructed based on the strength of the relationship

between i and j, which is operationalized by evaluating the overlap of i and j's social neighborhoods. Researchers can explore other specifications of δ_{ij} in the future. Lastly, the effectiveness of our measure should be tested against existing methodologies by using observational data produced in real world.

Appendices

Appendix A

Unidirectional Relationships as Weak Ties

In this appendix, we discuss our operationalization of weak ties used in our empirical analyses. We define tie strength based on the following-follower relationships observed in the Twitter network, and specifically, we claim that unidirectional relationships are *on average* weaker than bidirectional ones. We want to stress a few points regarding this assumption. First, we are not claiming that a bidirectional relationship in the Twitter world is a strong tie in the *absolute* sense. Twitter users, even if they are mutually connected online, often barely *know* each other in the real world, so to a certain extent, the claim that almost all ties on Twitter are weak is a fair one to make. The hypothesis only emphasizes the ordinal strength of the two tie types, and the comparison is carried out in the sense of *probabilistic expectation*. The reason why reciprocity makes a difference is that frequent learning or regular interaction is more likely to happen when a reciprocal relationship exists. By reading each other's posts, a pair of users can more easily develop mutual understanding about each other's topics of interest and expertise, and sometimes even about detailed aspects of each one's personal life. Over time, even

though the pair are unknown to each other in the real world, they could probably become very familiar with each other's activities and habits in the online community. Of course, reciprocal following does not guarantee such relationship development (which is why we emphasize the probabilistic nature of the hypothesis). However, without it, the relationship development is unlikely. Moreover, our operationalization is consistent with the previous sociological literature. Granovetter (1973) pointed out the importance of reciprocity by defining that "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie." In Friedkin (1982), asymmetrical contact between college professors was classified as a weak tie, and a reciprocal connection was classified as a strong tie. Marlow et al. (2009) also applied similar definitions in analyzing friendships on Facebook.

We perform an empirical test on the hypothesis, using the network graph data we collected. Note that we know not only the number of followings (followers) a user has, but also whom the followings (followers) are (i.e., we observe the IDs of the user's immediate social neighbors in our database). This information should give us more knowledge about, and in the meantime the ability to build important metrics of, a user's network characteristics. In particular, knowing the IDs of two users' social neighbors, we can compare how "similar" their social neighborhoods are. In deriving his theory, Granovetter in his 1973 paper claimed that

the stronger the social tie between two persons, the larger the overlap of their friendship circles. Applying this statement in the Twittersphere, under our assumption, we would expect that two users who mutually follow each other, on average, have a larger overlap in their followings (followers) than two who don't. Our test is based on this prediction. Operationally, we do so by empirically verifying whether $w_{ti} = 0$ positively correlates with a higher "similarity" between user ti and author t's followings (followers). We measure "similarity" by computing two overlap indexes of followings (followers) of author t and user ti:

$$OI_{ti}^{V1} = \frac{\bar{V}_{ti}}{\sqrt{V_t}\sqrt{V_{ti}}} \quad OI_{ti}^{V2} = \frac{\bar{V}_{ti}}{\min\{V_t, V_{ti}\}}, \tag{A.1}$$

where $\bar{V}_{ti}$, V_t, and V_{ti} are the number of mutual followings author t and user ti shared, the number of followings author t had, and the number of followings i had, respectively (Onnela et al. 2007 defined a similar "neighborhood overlap"). Similarly, we can define and compute overlap indexes of followers (OI_{ti}^{W1}, OI_{ti}^{W2}) by changing V to W in equation (A.1). Note that the two numerators in equation (A.1) are the same: $\bar{V}_{ti}$. The difference between OI_{ti}^{V1} and OI_{ti}^{V2} is in the denominators, or in the way by which we scale down $\bar{V}_{ti}$ based on the number of followings ti has. Both indexes are in the range $[0, 1]$ because $\bar{V}_{ti} \leq min\{V_t, V_{ti}\}$. The larger the indexes are, we say the more "similar" the two sets of followings are. When t and ti have no mutual followings shared, both indexes equal 0. When t and ti have exactly the same sets of followings, $OI_{ti}^{V1} = 1$. When ti's followings represent a subset/superset of t's followings, $OI_{ti}^{V2} = 1$.

	OI^{V1}	OI^{V2}	OI^{W1}	OI^{W2}
w_{ti}	-0.042***	-0.069***	-0.034***	-0.064***
F	(2322.21)	(1476.43)	(3837.34)	(2158.65)
p-value	0.00	0.00	0.00	0.00

Table A.1: Results of ANOVA Tests

We investigate wether different w_{ti} values lead to significantly different overlap indexes by running a series of ANOVA tests, the results of which are given in Table A.1. In all four tests, we control tweet-specific effects. As the regression coefficients in the first row show, we find that a unidirectional relationship ($w_{ti} = 1$) is indeed associated with a smaller overlap in social neighborhoods. The F statistics and p-values indicate this difference is significant at 0.1% level, no matter which index we use. Therefore, bidirectional relationships are associated with higher transitivity in social neighborhoods. The results thus support our hypothesis that unidirectional relationships are, on average, weaker than bidirectional ones.

Appendix B

An Graphic Example of Retweeting

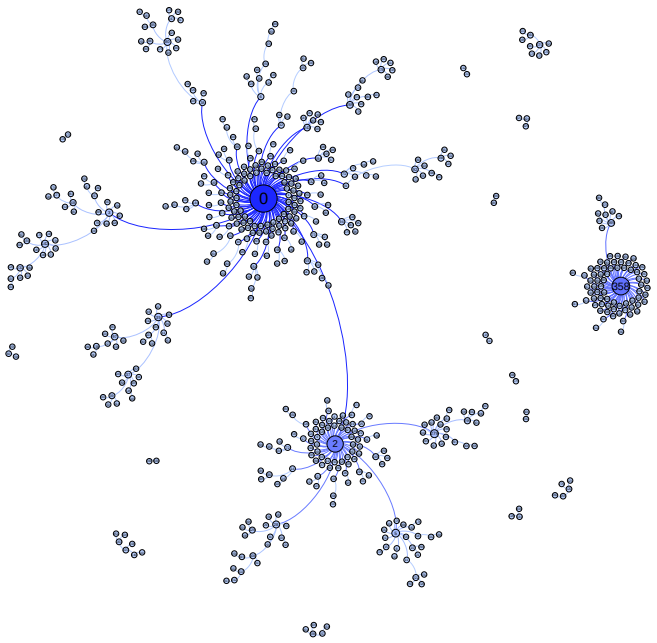

Figure B.1: The Spread of a Single Tweet (idx=1) in Our Sample

Bibliography

[1] Bakshy, E., J. Hofman, W. Mason, D. Watts 2011, "Everyone's an Influencer: Quantifying Influence on Twitter," *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*.

[2] Bapna, R., A. Gupta, S. Rice, A. Sundararajan 2012, "Trust, Reciprocity and the Strength of Social Ties: An Online Social Network based Field Experiment," *working paper*.

[3] Bass, F. 1969, "A New Product Growth for Model Consumer Durables," *Management Science* 15(5), 215-227.

[4] Bell, D., S. Song 2007, "Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing", *Quantitative Marketing and Economics* 5(4), 361-400.

[5] Bikhchandani, S., D. Hirshleifer, I. Welch 1992, "A Theory of Fads, Fashion, Custom and Cultural Change as Information Cascades," *Journal of Political Economy* 100, 992-1026.

[6] Blau, P. 1964, *Exchange and Power in Social Life*, Transaction Publishers.

[7] Bock, G., R. Zmud, Y. Kim, J. Lee 2005, "Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate," *MIS Quarterly* 29, 87-111.

[8] Bramoullé, Y., H. Djebbari, B. Fortin 2009, "Identification of Peer Effects through Social Networks," *Journal of Econometrics* 150, 41-55.

[9] Brin, S., L. Page 1998, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Proc. 7th International World Wide Web Conference*, 107-117.

[10] Burt, R. 1987, "Social Contagion and Innovation: Cohesion versus Structural Equivalence," *American Journal of Sociology* 92(6),1287-1335.

[11] Burt, R. 2005, "Closure, Trust and Reputation," *Brokerage and Closure*, 93-166, Oxford University Press.

[12] Centola, D., M. Macy 2007, "Complex Contagions and the Weakness of Long Ties," *American Journal of Sociology* 113(3), 702-734.

[13] Chamberlain, G. 1980, "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225-238.

[14] Chatterjee, R., J. Eliashberg 1990, "The Innovation Diffusion Process in a Heterogeneous Population: A Micromodeling Approach", *Management Science* 36(9), 1057-1079.

[15] Chiu, C., M. Hsu, E. Wang 2006, "Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories," *Decision Support Systems* 42, 1872-1888.

[16] Coleman, J., E. Katz, H. Menzel 1957, "The Diffusion of an Innovation among Physicians," *Sociometry* 20(4), 253-270.

[17] Constant, D., L. Sproull, S. Kiesler 1996, "The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice," *Organization Science* 7, 119-135.

[18] Davis, J. 1970, "Clustering and Hierarchy in Interpersonal Relations: Testing Two Graph Theoretical Models on 742 Sociomatrices," *American Sociological Review* 35, 843-851.

[19] Duan, W., B. Gu, A. Whinston 2009, "Informational Cascades and Software Adoption on the Internet: An Empirical Investigation," *MIS Quarterly* 33(1), 23-48.

[20] Fischer, C. 1992, *America Calling: A Social History of the Telephone to 1940*, University of California Press.

[21] Freeman, L. 1977, "A Set of Measure of Centrality Based on Betweenness," *Sociometry* 40(1), 35-41.

[22] Friedkin, N. 1980, "A Test of Structural Features of Granovetter's Strength of Weak Ties Theory," *Social Networks* 2, 411-422.

[23] Friedkin, N. 1982, “Information Flow Through Strong and Weak Ties in Intraorganizational Social Networks,” *Social Networks* 3, 273-285.

[24] Goldenberg, J., S. Han, D. Lehmann 2009, “The Role of Hubs in the Adoption Process”, *Journal of Marketing* 73(2), 1-13.

[25] Granovetter, M. 1973, “The Strength of Weak Ties,” *The American Journal of Sociology* 78, 1360-1380.

[26] Granovetter, M. 1978, “Threshold Models of Collective Behavior”, *The American Journal of Sociology* 83(6), 1420-1443.

[27] Granovetter, M. 1983, “The Strength of Weak Ties: A Network Theory Revisited,” *Sociological Theory* 1, 201-233.

[28] Green, H. 2008, “Google: Harnessing the Power of Cliques,” *BusinessWeek*, October 6, 50.

[29] Hansen, M. 1999, “The Search-transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits,” *Administrative Science Quarterly* 44 82-111.

[30] Heider, F. 1958, *The Psychology of Interpersonal Relations*, John Wiley & Sons.

[31] Hill, S., F. Provost, C. Volinsky 2006, “Network-based Marketing: Identifying Likely Adopters via Consumer Networks,” *Statistical Science*, 21(2), 256-276.

[32] Homans, G. 1958, "Social Behavior as Exchange," *The American Journal of Sociology* 63, 597-606.

[33] Jackson, M. 2008, *Social and Economic Networks*, Princeton University Press.

[34] Katona, Z., P. Zubscek, M. Sarvary 2011, "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research* 48(3), 425-443.

[35] Katz, L. 1953, "A New Status Index Derived from Sociometric Analysis," *Psychometrika* 18, 39-43.

[36] Kempe, D., J. Kleinberg, É. Tardos 2003, "Maximizing the Spread of Influence in a Social Network", *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146.

[37] Kleinberg, J. 1999, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM* 46(5), 604-632.

[38] Koren, Y., R. Bell, C. Volinsky 2009, "Matrix Factorization Techniques for Recommendation Systems," *IEEE Computer* 42(8), 30-37.

[39] Kwak, H., C. Lee, H. Park, S. Moon 2010, "What is Twitter, a Social Network or a News Media," *Proceedings of the 19th International Conference Companion on World Wide Web*.

[40] Lee, D., H. Seung 1999, "Learning The Parts of Objects by Non-negative Matrix Factorization", *Nature* 401, 788.

[41] Lee, D., H. Seung 2001, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing* 13, 556-562.

[42] Levin, D., R. Cross 2004, "The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer," *Management Science*, 50 11 1477-1490.

[43] Leskovec, J., L. Adamic, B. Huberman 2007, "The Dynamics of Viral Marketing", *ACM Transactions on the Web*, 1(1).

[44] Liben-Nowell, D., J. Kleinberg 2007, "The Link-Prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology* 58(7), 1019-1031.

[45] Ma, H., H. Yang, M. Lyu, I. King 2008, "SoRec: Social Recommendation Using Probabilistic Matrix Factorization," *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 931-940.

[46] Mahajan, V., E. Muller, Y. Wind 2000, *New-Product Diffusion Models*, Springer.

[47] Manski, C. 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60(3), 531-542.

[48] Manski, C. 2000, "Economic Analysis of Social Interactions," *Journal of Economic Perspectives* 14(3), 115-136.

[49] Marlow, C., L. Byron, T. Lento, I. Rosenn 2009, "Maintained Relationships on Facebook," online at http://overstated.net/2009/03/09/maintained-relationships-on-facebook.

[50] Milgram, S. 1967, "The Small-World Problem," *Psychology Today* 2, 60-67.

[51] Morris, S. 2000, "Contagion", *Review of Economic Studies* 67, 57-78.

[52] Mundlak, Y. 1978, "On the Pooling of Time Series and Cross Section Data," *Econometrica* 46, 69-85.

[53] Nair, H., P. Manchanda, T. Bhatia 2010, "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research* 47(5), 883-895.

[54] Nelson, P. 1970, "Information and Consumer Behavior," *Journal of Political Economy* 78(2), 311-329.

[55] Newman, M. 2003, "The Structure and Function of Complex Networks," *SIAM Review* 45, 167-256.

[56] Olivera, F., P. Goodman, S. Tan 2008, "Contribution Behaviors in Distributed Environments," *MIS Quarterly* 32, 23-42.

[57] Onnela, J., J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.-L. Barabasi 2007, "Structure and Tie Strengths in Mobile Communication Networks," *Proceedings of the National Academy of Sciences USA*, 104 7332-7336.

[58] Rauch, J. 2010, "Does Network Theory Connect to the Rest of Us? A Review of Matthew O. Jackson's Social and Economic Networks," *Journal of Economic Literature* 48(4), 980-986.

[59] Richardson, M., P. Domingos 2002, "Mining Knowledge-Sharing Sites for Viral Marketing," *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 61-70.

[60] Rogers, E. 1995, *Diffusion of Innovations*, Edition 4, Free Press.

[61] Ryan, B., N. Gross 1943, "The Diffusion of Hybrid Seed Corn in Two Iowa Communities," *Rural Sociology* 8, 15-24.

[62] Socialflow 2011, "Breaking Bin Laden: Visualizing the Power of a Single Tweet," available online at http://blog.socialflow.com/.

[63] Van den Bulte, C., G. Lilien 1997, "Bias and Systematic Change in the Parameter Estimates of Macro-Level Diffusion Models," *Marketing Science*, 16(4), 338-353.

[64] Van den Bulte, C., G. Lilien 2001, "Medical Innovation Revisited: Social Contagion Versus Marketing Effect," *American Journal of Sociology*, 106(5), 1409-1435.

[65] Valente, T. 2005, "Models and Methods for Innovation Diffusion," *Models and Methods in Social Network Analysis*, Cambridge University Press.

[66] Wasko, M., S. Faraj 2005, "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly* 29, 35-57.

[67] Watts, D. 2002, "A Simple Model of Global Cascades on Random Networks", *Proceedings of the National Academy of Sciences USA* 99, 5766-5771.

[68] Watts, D., S. Strogatz 1998, "Collective Dynamics of 'small-world' networks", *Nature* 393, 440-442.

[69] Weng, J., E. Lim, J. Jiang, Q. He 2010, "TwitterRank: Finding Topic-sensitive Influential Twitterers," *Proc. 3rd ACM International Conference on Web Search and Data Mining*.

[70] Wooldridge, J. 2001, *Econometric Analysis of Cross Section and Panel Data* Edition 1, The MIT Press.

[71] Wu, S., J. Hofman, W. Mason, D. Watts 2011, "Who Says What to Whom on Twitter," *Proceedings of the 20th International Conference Companion on World Wide Web*.

[72] Young, P. 2009, “Innovation Diffusion in Heterogeneous Populations: Contagion, Social Influence, and Social Learning,” *The American Economic Review* 99, 1899-1924.

[73] Zhang, S., R. Wang, X. Zhang 2007, “Uncovering Fuzzy Community Structure in Complex Networks,” *Physical Review E* 76, 046103.