Information Sharing in Social Broadcast:

Evidences from Twitter

Zhan Shi   Huaxia Rui   Andrew Whinston

Abstract

On the rapidly growing social broadcast sites, information diffusion is driven by social interactions such as sharing. This paper provides an economic model to approach one micro-level question on this phenomenon: why network participants share/forward information to network neighbors and how their decisions relate to their social characteristics. In the empirical part, we run a series of econometric analyses using a data set collected from Twitter. Overall, the results are positive for our model.

1 Introduction

Recent years have seen an explosive growth of various online communities. Phenomenally successful social networks, like Twitter and Facebook, connecting and facilitating interactions among up to hundreds of millions of users, have had increasingly important roles in shaping the new social life landscape. Among the changes brought by these online services, one revolutionary shift has taken place in news (or information in general) consumption and production. Powered by what we call social broadcast tech-
nologies such as Twitter, Tumblr and Sina Weibo\(^1\), people to a large extent have gone from searching and consuming news through traditional media or news websites to having the news broadcast to them by their social network neighbors. On the production side, these social broadcast services also greatly eased the creation and exchange of *user-generated contents*, which, sometimes entitled “social media”, have fostered innovative ideas in areas such as collaboration, advertising, or even election campaign\(^2\).

One very important difference between the participation-centered social broadcast services and the publishing-centered traditional media is that on the social sites the informational content is typically spread through social interaction such as sharing or recommendation rather than using physical technologies like radio or TV. This implies that while competing for attentions the network participants rely on each other to spread information they create. Since the influence of a piece of informational content depends not only on its quality, but also on the size of reached audience, the social value of the information is thus co-created by its generator and any network participants who help the diffusion of the content. Given the growing importance of social broadcast in this “Attention Age”, understanding this diffusion process is of both intellectual and practical interests. Researchers in various academic disciplines have begun to study this phenomenon from their own unique angles (see Jackson and Rogers (2007) and Libai, Muller and Peres (2009) among others). Our paper, with an economic standpoint, takes a step to approach one very basic question on the micro level about the content diffusion process: why people, as rational agents, share information with their social neighbors and what roles do different social characteristics play in the decision making.

Broadly speaking, this paper contributes to a large body of economic, marketing and sociological

---


\(^2\)see “2010: The first social media election”, *Guardian*, May 3, 2010
literature that studies the diffusion of information, knowledge, or innovation among a heterogeneous population. However, in contrast with the situation of information sharing in a network which we want to model, in much of this body of research, it is usually explicitly or implicitly assumed that information or innovation is diffused through the learnings of later adopters, as opposed to the intentional dissemination by early adopters. Among these studies, Peyton Young (2009) took on one important aspect of the diffusion phenomenon: why long lags occur between an innovation’s creation and its general acceptance by a population. The models examined include the contagion model of marketing, the social influence model of sociology and social learning model of economics. The finding of the paper is that different mechanisms of individual adoption will drive quite different diffusion processes which lead to distinctive aggregate dynamics. As for the driving force of diffusion on the micro level, the contagion model and social influence model (as well as some other researches in word of mouth) both rely on the notion of exposure, an idea which could be traced back to models’ epidemiological origins. In other words, utility maximization was not the foundation of these two models. The social learning model was indeed founded upon rational decision-making framework, but as mentioned earlier the agent who makes a decision is the adopter, but not the sharer of innovation. In fact, the model assumes that innovation is diffused through the potential adopters’ observations and learnings, but not the intentional dissemination of early adopters.

Another key difference between Young (2009) and our model is that Young assumes the population to be infinite and encounters between individuals are purely random while we study the agents’ behavior in a finite and fixed social network. As a result the agents in our model are heterogeneous in the sense that they have different social characteristics determined by the existing structure of the neighborhood surrounding them. From this perspective of allowing social structure impact diffusion, our model is more
close to Jackson and Yariv (2007) and Golub and Jackson (2009). Jackson and Yariv (2007) models a potential innovation-adopter’s choice making decision, in which her payoff depends on the number of neighbors she has and her neighbors’ choices. Their very general framework is built on the simple idea of coordination game, and the interaction between agents is the model’s focal point. However in our paper, we instead model the decision of sharing information and do not introduce agents’ strategic interactions in order to maintain its econometric tractability.³

One difficulty of our current work is that unlike adopting an innovation, the return an agent enjoys from sharing general information to other people is unobvious. This could also be one of the reasons why general information sharing is not sufficiently modeled in existing economic literature. We solve this difficulty by thinking of the action of sharing as an investment in interpersonal relationships. The idea is that, though sharing appears to be an altruistic behavior, a person who disseminates new information to her social neighbors expects reciprocal services during their future coexistence, which justifies sharing as a rational behavior. One branch of previous sociological researches, which support this argument, study the strength of social ties (see Granovetter (1973) and Friedkin (1982)). Tie strength - a concept ranging from weak ties at one extreme to strong ties at the other - characterizes the closedness and interaction frequency of a social relationship. We borrow this concept from sociology and let it play a central role here. Its application in our model is twofold. First, we consider sharing behavior as an investment in interpersonal relationship, the value of which depends on the strength of social ties between the agent and her social neighbors. Second, one important hypothesis in the tie strength theory is that a pair of strongly tied persons share a larger overlap in their friendship circles than a pair of weakly tied persons do.

³We only observe one-shot sharing behavior. If we were to model the interaction between agents, we would like to observe one person sharing and whoever gets the shared information sharing back.
(see Granovetter (1973) for theoretical argument and Onnela et al (2007) for empirical results). In our model, we require agents to form beliefs in accordance with this principle when making their decisions.

Connecting theory and data is important in our research. Our model produces testable predictions which we examine in the empirical part using a data set collected from Twitter, the biggest and most vibrant social broadcast service in the world. In particular, we test the hypothesis mentioned in the previous paragraph, the Triadic Closure Property we will discuss later and the correlations between agents’ sharing decisions and their social characteristics predicted by our model. Additionally, we use maximum likelihood to estimate the scaled versions of the structural parameters. The data set produces estimates with signs which are consistent with our model. Overall, the empirical findings in our series of tests are positive for our model.

The remainder of this paper is organized as follows. In Section 2, we give a short introduction of Twitter and describe our data collecting strategy. We present a model of content sharing in social networks in Section 3, where we also discuss how an agent’s social network characteristics and interpersonal tie strength come into play in the decision-making process. In Section 4, we show our empirical findings and estimate some of the structural parameters of our model. We conclude the paper in Section 5 and point out future research directions.

2 Twitter and Data

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 festival at Austin, Texas, a showcase of Twitter impressed
the highly tech-savvy attendees and turned out to be an enormous marketing success. Since then, Twitter has entered a phase of rapid growth and gained popularity far beyond the technology industry insiders. As of April 2010, Twitter has more than 100 million registered users, who, in total, post an average of 55 million updates a day⁴. Now Twitter is one of the most vibrant micro-blogging and social networking services in the world.

Twitter can be considered as a decentralized content generation and consumption site where every user can post updates known as tweets⁵, which are text based messages of up to 140 characters and, by default, are available to be read by anyone. If we think of tweets as informational goods, then Twitter is a large and open marketplace where each user is both a producer and a consumer. As a consumer, a user (called Jim) may become a regular reader of another producer (called Jane), i.e. subscribing to Jane or, in Twitter’s terminology, following Jane. Meanwhile, as a producer, Jim updates tweets and may have his own subscribers (called followers). Therefore a user’s followers are those who subscribe to receive her tweets and a user’s followings are the users whose tweets she subscribes to receive. This following-follower relationship completely characterizes the interpersonal link on Twitter. It differs from the friendship on Facebook or some other social network site in two aspects: 1) the following-follower relationship on Twitter is relatively open in the sense that A following B does not require B’s consent and it does not map the real world friendship as that on Facebook does; 2) maybe more importantly the following-follower relationship is uni-directional (A following B does not imply B following A) while friendship is bi-directional ( A being a friend of B implies B being a friend of A). The uni-directional relationships collectively depict the unique social structure in the Twitter network. How this network was formed,

⁴For more statistics, see http://www.huffingtonpost.com/2010/04/14/twitter-user-statistics-r_n_537992.html
⁵Tweet can also be used as a verb, meaning to post. So “tweet a tweet” means “post an update”.
or how the heterogeneous informational goods producers and consumers are matched, and what factors determine its change are interesting questions on network dynamics. But it is out of scope of this paper. We take a snapshot of the existing Twitter social structure, consider it as it is and study users’ sharing behavior on top of it. The interplay of users' behavior and the underlying social structure will be devoted to future research.

Content sharing is an integral part of the Twitter experience. Besides composing and posting a tweet by themselves, Twitter users may also want to re-post other users’ (in most cases their followings’) tweets which they find of certain (informational, entertaining, etc) value, which, in Twitter’s terminology, is said to retweet. With the official retweet function, users only need to click one button to share. When many users retweet one single tweet, it can diffuse into the social network far beyond the tweeter’s direct followers and reach a much larger audience. In this perspective, retweet clearly mimics word of mouth by face-to-face communication in our physical world, except that retweet is a technology based word of mouth, so it is unique in the sense that the information stays exactly the same as when it is generated, no matter how long and how far the message has been passed. Hence, especially when we consider tweets as informational goods, retweet is then also an action of sharing with many people simultaneously. Through retweet, users who otherwise would not read the message could possibly consume (and derive utility from) the same information received by the retweeter. Our research centers around the question why the users, as rational decision makers living in a social network, intentionally disseminate useful information to their social neighbors. One weakness of the literature of word of mouth as well as the

---

6Retweet is also both a verb and a noun as tweet is. Twitter users call a user who retweets a retweeter.
7Posting others’ tweets simply by copying and pasting their tweets without mentioning the original author is not considered retweeting, rather, it is a highly criticized misbehavior in the Twitter community.
8The official retweet function is built in most mobile applications as well as Twitter’s own website.
literature of information sharing/diffusion is the lack of an individual decision model which potentially could be tested or even structurally estimated by un-experimental micro-level data, partly due to the impossibility of collecting a significant real world sample before the Internet age (or even before the popularization of the web 2.0 sites). The emergence of online social networks like Twitter, bringing in vast troves of machine-generated data about user relationships and activities, has made such research possible.

So we will use retweet in the Twitter world as the primary example of our model and use Twitter data to do empirical analysis. The mechanism of retweet is visually illustrated in Figure 1. Hereafter we will call the user who writes the original message the root¹⁹ and the root is denoted by R in the figure. The nodes represent different 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, their link is drawn in a solid line. Otherwise (only one of them follows the other), their link is a dashed line with an arrow pointing to the following. After R posts an update, without anyone retweeting it, only R’s followers A B C D and E will receive it. Now let us say, after reading the message, users A D and E

¹⁹The retweet data has tree structure and root is a borrowed terminology from the field of data structure.
retweet, thereby making $F G H$ and $K$, who are not direct followers of $R$, get the tweet and potentially become part of the audience.\textsuperscript{10} Then the new audience could also retweet (as $G$ and $H$ do shown in the figure), circulating the information further into the network. For the ease of expression, we will call users like $A D$ and $E$ the first layer retweeters, those like $G$ and $H$ the second layer retweeters, and so on. Provided that the retweet function is prevalently adopted across platforms and easy to use, anyone who gets the tweet can use it without significant technological burden, so in principle, the retweeting process could last\textsuperscript{1} for ever. In this paper, we approach the question why users like $A$ and $G$ share information with connected social neighbors (retweet in this Twitter example) while users like $B$ and $F$ don't, and what roles do different interpersonal links and people’s social network characteristics play in the decision making process.

Our data set is collected by using Twitter’s Application Programming Interface (API). Our basic strategy is picking up a number of tweets and, for each of them, looking at the retweeting behavior of a part of its readers. We choose our tweets from an official Twitter account called \textit{toptweets}\textsuperscript{11}. This account uses proprietary algorithm to choose and display the most popular tweets from the entire Twitter space\textsuperscript{12}. Clearly the updates displayed by this account are not a random and representative sample of all the tweets in the Twitter world, so whenever we use our sample to do any empirical analysis, we by default do not plan to interpret our results as evidences of the average behaviors on Twitter, but rather

\textsuperscript{10}If both $C$ and $D$ retweeted (using the official retweet function), $K$ gets only one copy of the message from whoever retweeted first.

\textsuperscript{11}http://twitter.com/toptweets/

\textsuperscript{12}Twitter’s Chief Scientist Abdur Chowdhury explained “Top Tweets is a new algorithm we developed that finds tweets that are catching the attention of other users”, “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”. 
only as the real world illustrations of our model.

The details about data collection are described below. From July 1st to July 30th 2010, at 0:05am of each day, we picked from toptweets the most recent post whose writer had fewer than 1000 followers¹³.

For each one of the tweets, our data set contains information of the user id of the root, the time when it’s posted as well as its content. For the root, we also know the number of her followings, the number of her followers, and the user ids of her followers (eg. A, B, C, D, E in Figure 1). The root’s followers are the set of people (denoted by A) we are primarily interested in, so we also collect the user ids of their followings and followers (their social neighbors). We kept monitoring the retweeting data for 96 hours and what we got are the user ids of the retweeters (eg. A, D, E, G, H in Figure 1, denoted by B) in the chronicle order of their retweets. Moreover, note that the intersect of A and B are the first layer retweeters. For each first layer retweeter, we collect the following ids and follower ids for each of her followers (eg. F, G, H, K in Figure 1). Therefore, we have the ids of second layer retweeters’ followings and followers.

(to be done. Do descriptive analysis of our data set) Figure 2 is a real world analog of Figure 1.

3 The Model

We model rational individuals’ information sharing/forwarding decisions in an environment of a finite and fixed social broadcast environment. In particular, an individual’s social characteristics, including the strength of the social ties via which she is bonded with her social neighbors, play important roles in her

¹³This is due to computational capacity constraint. Twitter API has a limit of 150 queries per minute per ip address
Figure 2: The Paths of Information Diffusion: An Example

decision making.

3.1 Payoffs

The set of agents we consider are the users who receive the information, whether from the generator of the information (the root) or some other sharer (e.g. a retweeter). The social structure of the network surrounding an agent $i$ is captured by the number of people whom $i$ subscribes to (denoted by $W_i$) and the number of people who subscribe to $i$ (denoted by $V_i$). We denote the agent from whom $i$ receives the information $r$. Therefore, $i$ is certainly a subscriber of $r$. But $r$ may not necessarily be a subscriber of $i$. 
We assume consuming a piece of information is a necessary first step before sharing it and receiving the message does not guarantee that an agent $i$ actually “sees” or “reads” it. We believe this is a plausible assumption given the limited time resource an agent is endowed with and the abundance of user generated contents available on various online communities that we study. Let $z_i$ be a binary variable which indicates whether agent $i$ consumes a message conditional on receiving it, with $z_i = 1$ meaning consuming it and 0 otherwise. The outcome of $z_i$ is determined partly by the frequency of $i$ receiving a piece of information from those $i$ subscribes to, and partly by the mount of effort $i$ puts in consuming the received messages. Intuitively, the more frequent new pieces of information $i$ receives, the less likely $i$ will “read” one particular message among them. Formally, we assume the frequency is an increasing and linear function of $W_i$, written $\tau W_i$, where $\tau$ is a positive parameter. The latent variable that determines the outcome of $z_i$ is $z^*_i = b\tau W_i + e_i$, where $e_i$ (effort) is an identically and independently distributed random variable. We normalize $b = -1$, so

$$z_i = 1 \text{ if and only if } z^*_i > 0 \iff -\tau W_i + e_i > 0$$  

(1)

After consuming the information, each agent has a choice of sharing it with her subscribers. We denote this binary action $y_i$, with $y_i = 1$ meaning sharing and $y_i = 0$ meaning not sharing. Agent $i$ considers sharing as an investment in relationships with social neighbors, including both the one between $i$ and $r$ and those between $i$ and $i$’s subscribers. So the utility $i$ gets from sharing comes from two parts. The first is the enhancement of the social link between $i$ and $r$, which we denote $u$. Alternatively, we can think of $u$ as the present value of the increased services $i$ can get from $r$ in the future because of the sharing. The second part comes from the relationship enhancement between $i$ and the subset of $i$’s subscribers, denoted by $pV_i$, who receive the information because of $i$’s sharing. Here note that $p$ is a
number between 0 and 1, since it might be the case that a portion of \( i \)'s subscribers (sized \((1 - p)V_i\)) have already received the information from earlier disseminator other than \( i \). For example, some of \( i \)'s subscribers may in the meantime subscribe to \( r \) and they can and do receive the message even without \( i \)'s sharing. Therefore, our assumption implies what \( i \) cares about is the size of increased audience of the shared content due to \( i \)'s action. We write \( i \)'s utility from the second source \( apV_i \), where \( a \) could be thought as the mean present value of the increased future services from the subset of \( i \)'s subscribers and is a positive constant. On the other hand, there is a cost for agent \( i \) to choose action 1, denoted by \( c_i \), which we assume to be identically and independently distributed across the population. So the net payoff from choosing sharing is \( u + pV_i - c_i \), where we normalize \( a = 1 \) without loss of generality. We also normalize the payoff from adopting action 0 to be 0. Therefore, \( i \) prefers 1 if and only if

\[
u + pV_i \geq c_i
\]  

\[(2)\]

3.2 Tie Strength and The Semi-rational Belief of \( p \)

The parameter \( p \) in equation (2) is critical in \( i \)'s decision making and we assume it is agent \( i \)'s belief of the proportion of her subscribers who have not been exposed to the information. Her subscribers who are not new to the information are those who have already been informed by earlier disseminators. In this model, the belief is assumed to be neither totally arbitrary nor completely rational (\( p \) is not equal to the true percentage value). Instead, we impose a condition requiring the agents to form beliefs semi-rationally, i.e. \( p \) is a function of the strength of social ties between \( i \) and earlier disseminators.

\[
\text{Strong Tie} \quad \text{Weak Tie} \quad \text{No Tie}
\]
Suppose \( r \) is the generator of the information to be diffused. \( i \), a subscriber of \( r \), forms the belief of her \( p \) according to the strength of social tie between \( r \) and \( i \). For simplicity, we define two levels of tie strength: strong and weak. If \( i \) and \( r \) mutually subscribe to each other, then their tie is strong. If \( r \) does not subscribe to \( i \), then their tie is weak. The classification and terminology we use should not be interpreted in absolute terms, but rather in relative terms. This ordinal definition is quite intuitive, but its legitimacy needs more explanation and support by evidences. We believe bi-directional tie is stronger than uni-directional tie because typically more interaction takes place between agents tied bi-directionally. In Granovetter (1973), the author emphasized the importance of reciprocity in personal ties: “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.” Moreover, Friedkin (1982) defined the strength of ties between college professors, in which asymmetrical contact is identified as weak tie and reciprocal tie as strong tie. So we believe our definition is consistent with the sociological tradition. Similar definition has also been used by researchers studying social networks (see Marlow et al (2009) analyzing the friendship on Facebook).

Our assumption implies \( i \) is informed about whether \( r \) subscribes to \( i \) or not. Let \( p_s \) and \( p_w \) be the beliefs when the link between \( i \) and \( r \) is strong and weak respectively. As explained earlier, by sociological theory, strongly tied agents should in general have larger overlap in their social neighbors, in this case, subscribers. Our model requires the agents to believe so. Therefore, everything else held equal, an agent whose link with \( r \) is strong will anticipate that a larger proportion of his subscribers are also subscribers of \( r \) than a peer whose link with \( r \) is weak, i.e. \( p_s < p_w \) (so smaller proportion are new to the information). Note that this prediction also provides us a way to verify our definition of tie strength using our Twitter data. Later in Section 4, we will do test and show the empirical results.
We note two more points about our model. When $r$ is not the generator of the information, but herself a sharer, we additionally require $i$ to form belief of $p$ according to her social distance from the generator. The farther $i$ is from the generator, the more earlier sharers there are, so the smaller $i$’s $p$ is. For example, in the retweet case, everything else held equal, a second layer retweeter’s $p_2$ ($G$ in Figure 1) is smaller than a first layer retweeter’s $p_1$ ($D$ in Figure 1). The second point is that the strength of interpersonal tie also plays a role is $u$ in equation (2). Recall that $u$ is the present value of increased future services of $r$ because of $i$’s investment in relationship. So we allow $u_s > u_w$.

### 3.3 The Probability of Sharing

Putting everything together, we still need to make one more assumption: the random variables $e_i$ and $c_i$ are independent. This implies that the effort an agent puts in consuming received information has no relationship with the cost of sharing to her subscribers. Let $p_{w-s} = p_w - p_s$, $u_{w-s} = u_w - u_s$ and $\delta_w$ be a binary variable which indicates an agent has a weak tie with $r$ ($\delta_w = 1$) or a strong tie ($\delta_w = 0$). Agent $i$ is a subscriber of the information generator $r$. Then from an econometrician’s point of view, the probability for $i$ to share can be written

$$P(y_i = 1) = P(-\tau W_i + e_i > 0)P(u + pV_i \geq c_i)$$

$$= P(-\tau W_i + e_i > 0)P(u_s + \delta_w u_{w-s} + (p_s + \delta_w p_{w-s})V_i - c_i \geq 0)$$

(3)

Several predictions on the relationships between an agent’s social characteristic variables and her sharing behavior can be drawn from equation (3). Everything else equal,

1. the probability of sharing is negatively correlated with $W_i$, since $\tau$ is a positive constant;
2. the probability of sharing is positively correlated with $V_i$, since $p_s + \delta_w p_{w-s}$ is always between 0 and 1;

3. the relationship between the probability of sharing and the strength of tie is a compromise between two effects:

$$u_{w-s} + p_{w-s} V_i$$

where $u_{w-s}$ is negative and $p_{w-s}$ is positive. So the sign of terms added up is unknown a priori.

Moreover, equation (3) will serve as the likelihood function of our maximum likelihood estimation in the part of empirical analysis.

4 Empirical Results

In this section, we use our Twitter data set to test various aspects of our model.

4.1 Strength of Tie in The Twitter World

Remember that the interpersonal link in the Twitter space is determined by the very following-follower relationship. The link between two users $A$ and $B$ can be 1) bi-directional if they follow each other, or 2) uni-directional if $A$ follows $B$ and $B$ doesn’t follow $A$ or vice versa, or 3) none if neither follows the other. As argued in the previous section, we define two strength levels of a social tie, if any: reciprocal relationship is strong and uni-directional relationship is weak.
Tie Strength and The Overlap of Friendship Circles

In Granovetter (1973), the author argued that the stronger the tie between two individuals, the greater the extent of overlap in their friendship circles. This hypothesis is a critical building stone of his famous “the strength of weak tie” theory and also of our model where agents make decisions partly based on this principle ($p_w > p_s$). So the first thing we do is to test whether this hypothesis is supported by the Twitter data set.

To do this exercise, we face two problems. The first is defining the analog of friendship circle on Twitter. We use a user’s social neighbors, namely her followings and her followers. Before any analysis, we do not know which group is closer, in the sociological sense, to that discussed in previous literature. So for completion, we will iterate using the two groups as the friendship circle in our test. The second problem is how to measure the overlap of two friendship circles. The difficult part is resulted from the fact that the sizes of following (and/or follower) sets differ wildly across users. In our data set, the user with the most followers (followings) has 5053697 followers (726266) and the one with fewest has none (0). So what we need is a quantity that can measure the similarity of two sets possibly with wildly different number of elements in a uniform and reasonable way. We use the same notations as in the model section and additionally let the size of the intersection of $R$ and $A$’s followers be $V_{RA}$. We define 

Follower Overlap Index (FwerOI) to be

$$\text{FwerOI} = \frac{V_{RA}}{\sqrt{V_R \cdot V_A}} \quad \text{FwerOI}_m = \frac{V_{RA}}{\min\{V_R, V_A\}}$$

Note that both FwerOI and FwerOI$_m$ are between 0 and 1. The difference between the two measures is evident when one user’s followers is a subset of the other’s. For example, say $R$ has 10 followers, out of which 5 are all of $A$’s followers. In this case, FwerOI = 0.71 and FwerOI$_m$ = 1. We similarly define
In this part we test whether there is a significant relationship between the strength of social tie linking two Twitter users and the overlap of their followings and followers. Specifically we run several linear probability models. The dependent variable is a dummy variable indicating strong tie \( \delta_s \), and the explanatory variable(s) is either one of these indexes or are all of them. Column 1 of Table 1 is the results of 4 separate regressions: regressing the binary outcome \( \delta_s \) on each of the four overlap indexes. Column 2 of the table is a single regression: regressing \( \delta_s \) on the four indexes in one equation. We can see that when we regress \( \delta_s \) on a single overlap index, the coefficients in all equations are significantly positive.\(^1\) This is a strong evidence that the likelihood of existing a bi-directional link between two

---

\(^{1}\)In all tables ** stands for significant under 5\% confidence level, * stands for significant under 10\% confidence level. \( t \)
people is positively correlated with the similarity of their social network neighbors. This result supports our definition that bi-directional ties are stronger than uni-directional ones in the Twitter world and our setting $p_w > p_s$ in the model section. The reason why we do column 2 is to identify which index of the four gives the most explanatory power of link type. We suggest $F\text{weO}I$ might be the best out of these four indexes in predicting tie strength.

**The Strong Triadic Closure Property**

Another important hypothesis in Granovetter (1973) is that if $A$ and $B$ are strongly tied, $A$ and $C$ are strongly tied, then there must be a (strong or weak) tie between $B$ and $C$. This hypothesis, which is called “The Strong Triadic Closure Property” by later sociological researchers, leads to the conclusion that a strong tie can not be a link that bridges two separate groups of people, and only weak ties may play that role. As a result, information transmitted through strong ties will be confined to the standing clique while weak ties enable information to spread outside of the clique and potentially reach more people.

We now test whether a weaker version of this hypothesis holds in the Twitter world. For each root, we divide his followers into two groups. The first group consists of the people who are strongly (bi-directionally) tied to the root, and the second group consists of the rest. If the hypothesis in Granovetter (1973) weakly holds, we would predict there should be more links among people in the first group than in the second. To do this exercise, we follow the literature (Newman (2003)) to calculate cluster coefficient (CC in Table 2) for each group, which is a metric used widely in topological analysis of a social network.
<table>
<thead>
<tr>
<th></th>
<th># of Followers</th>
<th>Group 1</th>
<th>Group 2</th>
<th>CC 1</th>
<th>CC 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>mean</td>
<td>2555</td>
<td>0.271</td>
<td>0.729</td>
<td>0.146</td>
<td>0.008</td>
</tr>
<tr>
<td>medium</td>
<td>1396</td>
<td>0.124</td>
<td>0.876</td>
<td>0.110</td>
<td>0.005</td>
</tr>
<tr>
<td>stdev</td>
<td>2788</td>
<td>0.291</td>
<td>0.291</td>
<td>0.151</td>
<td>0.009</td>
</tr>
<tr>
<td>max</td>
<td>10468</td>
<td>0.991</td>
<td>1.000</td>
<td>1.000</td>
<td>0.062</td>
</tr>
<tr>
<td>min</td>
<td>85</td>
<td>0.000</td>
<td>0.009</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2: Cluster Coefficients of Community Formed by Strongly Tied Followers and Community Formed by Weakly Tied Followers

It measures the density of interpersonal links within a social community. Formally, it is defined to be the ratio

$$CC = \frac{\text{Number of Existing Links}}{\text{Number of Possible Links}}$$

Table 2 shows the statistics of the roots’ numbers of followers, the fractions of the two groups and the computed cluster coefficients. Group 1 are people with strong tie, and the mean cluster coefficient is 0.146. Group 2 are people with weak tie, and the mean cluster coefficient is 0.008. So the links within Group 1 are 18 times denser than those within Group 2. Hence the statistical evidence from our data set supports that the Triadic Closure Property weakly holds in the Twitter world. This again means our ordinal definition of tie strength is consistent with theoretical prediction.
4.2 Regression Analysis

In this section we test the predictions of the model, as stated in Section 3.3. Basically we run a series of binary choice regressions where the dependent variable is a Twitter user’s sharing decision and the independent variables are the user’s social characteristics and the strength of social tie.

Technically, we apply econometric methods which deal with clustered sample. Our data set is clustered by tweets, which we denote by \( t \in \{1, \ldots, T\} \). For each tweet, we know who is the root \( t_t \), and all the root’s followers, which we denote by \( i \in \{1, \ldots, I_t\} \), where \( I_t \) is \( r_t \)’s total number of followers. \(^{15}\) \( I_t \) could be different across \( t \), so it’s an unbalanced cluster sample. For each \( r_t \) and \( i \in \{1, \ldots, I_t\} \), we know their followings and followers. Therefore, we can determine the strength of tie between the \( r_t \) and each \( i \). We use the same notations as in the model section, and the retweeting outcome is a binary variable \( y_{ti} \) which is 1 for retweeting and 0 otherwise. We use unobserved effects probit model, and the main assumption of this model is

\[
P(y_{ti} = 1 | x_t, c_t) = P(y_{ti} = 1 | x_{ti}, c_t) = \Phi(x_{ti} \beta + c_t) \quad i = 1, \ldots, I_t
\]  \hspace{1cm} (4)

where \( c_t \) is the unobserved effect, which we will talk about later, and \( x_t \) contains \( x_{ti} \) for all \( i \). \( x_{ti} \) is a vector of covariates that affects the retweeting behavior, including the dummy \( \delta_w \). Notice \( \delta_s \) is not included so that the group of strongly-tied users is the reference group. The coefficient of \( \delta_w \) should therefore be interpreted as the relative effect compared with \( \delta_s \).

Besides the dummy variable for tie strength, we also add in our regression equation two control

\(^{15}\)We preclude the root’s followers who set up privacy protection from our analysis. The reason is that we cannot observe followings and followers for these users. We believe dropping these users won’t introduce sample selection problem because for people who set up privacy protection and those who don’t, the proportions of retweeters are almost the same.
variables, $V$, the number of people who follow $i$ and $W$, the number of people whom $i$ follows. These two variables describe the structure of the social neighborhood surrounding a Twitter user. The vector of covariates $x_{ti}$ can be written

$$x_{ti} = (1, \delta, W_{ti}, V_{ti})$$

We summarizes these four variables in Table 3. Note that in computing these statistics, we group the users by tweets and calculate the average values for each of these variables. So the mean, max and min in Table 3 are across tweets, but not across users. For example, the max in “follower” column is 18509. This does not mean the user with the most followers have 18509 followers, but for the group who have the largest average number of followers the value is 18509. Also we can see that the distribution of tie types varies a lot across tweets. On average, weak tie users are more than strong tie users.

$c_t$ contains the unobserved effects that impact every $i$’s $\{1, \ldots, I_t\}$ retweeting behavior, for example, most obviously, the update’s “quality”. An update could be really funny, or it might tells some inside story about a hot topic, etc. In these cases, we would expect more people to retweet, all other variables held equal. $c_t$ may also include the timing effect of the tweet, since we would expect a message

<table>
<thead>
<tr>
<th></th>
<th>Weak Tie</th>
<th>Following</th>
<th>Follower</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>74.8</td>
<td>1490</td>
<td>2538</td>
</tr>
<tr>
<td>max</td>
<td>100</td>
<td>9800</td>
<td>18509</td>
</tr>
<tr>
<td>min</td>
<td>0.2</td>
<td>173</td>
<td>201</td>
</tr>
</tbody>
</table>

Table 3: Descriptive Statistics of Cluster Averaged Covariates
Figure 3: Number of Followers (left $y$ axis) and Number of Retweeters (right $y$ axis) across Tweets posted in daytime to be retweeted more than another which is posted in night. Other effects such as the writer’s reputation or whether the tweet is featured by Twitter might also be considered as parts of $c_t$. The existence of $c_t$ is supported by Fig 2, in which we plot, for each tweet, the number of followers (red bar with left $y$ axis) and the number of retweeters (blue bar with right $y$ axis). As we can see from the figure, it is clear that the percentage of retweeters varies wildly across tweets. Hence including $c_t$ is necessary.

We can now rewrite (4) as

$$P(y_{ti} = 1|x_t, c_t) = \Phi(\alpha + c_t + \gamma \delta_w + \beta_1 W + \beta_2 V) \quad i = 1, \ldots, I_t$$

(5)

We estimate four different econometric models: 1) random effects probit model; 2) pooled probit; 3) Chamberlain’s random effects probit model and 4) a pooled version of Chamberlain’s random effects
<table>
<thead>
<tr>
<th></th>
<th>RE Probit</th>
<th>Pooled Probit</th>
<th>Chamberlain RE</th>
<th>Chamberlain Pooled</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_w$</td>
<td>0.091**</td>
<td>0.023</td>
<td>0.098**</td>
<td>0.104**</td>
</tr>
<tr>
<td></td>
<td>(3.01)</td>
<td>(0.26)</td>
<td>(3.20)</td>
<td>(3.01)</td>
</tr>
<tr>
<td>$V$</td>
<td>6.7e-7**</td>
<td>5.5e-7**</td>
<td>6.7e-7**</td>
<td>6.1e-7**</td>
</tr>
<tr>
<td></td>
<td>(4.45)</td>
<td>(-3.73)</td>
<td>(4.49)</td>
<td>(4.61)</td>
</tr>
<tr>
<td>$W$</td>
<td>-7.1e-5**</td>
<td>-7.4e-5**</td>
<td>-7.1e-5**</td>
<td>-6.4e-5**</td>
</tr>
<tr>
<td></td>
<td>(-8.94)</td>
<td>(-3.60)</td>
<td>(-8.91)</td>
<td>(-3.61)</td>
</tr>
<tr>
<td>Observation</td>
<td></td>
<td></td>
<td>158614 in 78 groups</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Results of Binary Choice Models

probit model. The results are given in Table 4.

In order to perform the random effects probit analysis, we need to make two more assumptions besides (5):

$$y_{t1}, \ldots, y_{tI_t} \text{ are independent conditional on } (x_t, c_t)$$

(6)

$$c_t | x_t \sim \text{Normal}(0, \sigma_c^2)$$

(7)

Equations (5) (6) and (7) together ensure that random effects probit regression will give us the consistent estimates. The result is given in the first column of Table 4. $\delta_w$ is estimated to be positive and it’s significant under 1% confidence level, which means followers who are weakly tied to the root are more likely to retweet than strongly tied followers, everything else held equal. We also get significant coefficients for the control variables $V$ and $W$. Users with more followings and fewer followers are less likely to retweet,
which supports our model’s predictions.

The second column of Table 4 is the result of a pooled probit regression. This model is more plausible when we need to relax assumption (6), i.e. when we think people’s retweeting behavior is correlated with each other. If this is the case, equation (6) fails and random effects probit model will yield inconsistent estimates. One useful observation, however, is that, under (5) and (7), we have

$$P(y_{ti} = 1|x_t) = P(y_{ti} = 1|x_{ti}) = \Phi(x_{ti}\beta_c), \text{ where } \beta_c = \beta/(1 + \sigma_c^2)^{1/2}$$

Hence a pooled probit will consistently estimate a scaled version of the coefficients. Since we are primarily interested in the signs and significances of independent variables, but not in absolute values, this pooled regression is then an exercise that can test the robustness of our results. As shown in the table, though the coefficient of $\delta_w$ becomes insignificant, the signs and significances of $V$ and $W$ do not change.

Another thing we can do to relax the assumptions in random effects probit model is to explicitly allow unobservables to be correlated with explanatory variables. Here we appeal to Mundlak (1987) version of Chamberlain (1980)’s assumption:

$$c_t|x_t \sim \text{Normal}(\phi + \bar{x}_t\xi, \sigma_a^2)$$

where $\bar{x}_t$ is the average of $x_{ti}$, $i = 1, \ldots, I_t$ and $\sigma_a^2$ is the variance of $a_t$ in the equation $c_t = \phi + \bar{x}_t\xi + a_t$. We believe it is plausible to allow for correlation between the unobserved effect and some explanatory variables. The reason is that the mean of $\delta_s$ across a root’s followers is the proportion of them followed back by the root, which might be correlated with the root’s reputation that is in turn a part of $c_t$. Therefore, ex ante we can not rule out the possibility of correlation between $c_t$ and $x_{ti}$, and the Chamberlain model should be preferred. For the Chamberlain model, we still can estimate both the
random effects version and the pooled version, the results of which are given in column 3 and column 4 of Table 4 respectively. Note that for column 4, we are not estimating the coefficients themselves, but the scaled versions similar to those in column 2. Again, it is ok since we are only interested in signs and significances. We get the same signs and significances for the coefficients of $V$ and $W$, so we believe that it is a pretty robust result that users with more followers and less followings are more likely to retweet. We get positive estimates for $\delta_w$ in both columns and they are statistically significant.

### 4.3 MLE Analysis

In this section, we use maximum likelihood estimation techniques to estimate (scaled versions of) the structural parameters in equation (3) and the purpose of this exercise is to test whether the scaled parameters all have the predicted signs. Technically, first of all we need to specify the distributions of the two random variables $e_i$ and $c_i$. Let

$$e_i \sim N(\bar{e}, \sigma_e^2) \quad c_i \sim N(\bar{c}, \sigma_c^2)$$

where $\bar{e}$ and $\bar{c}$ are the mean effort and mean cost in the population respectively. We further use $\phi(\cdot)$ to denote the cumulative distribution function of standard normal. Then we can rewrite equation (3) as

$$P(y_i = 1) = \phi(\frac{\bar{e} - \tau W_i}{\sigma_e})\phi(-\frac{\bar{c}}{\sigma_c} + \frac{u_s}{\sigma_c} + \delta_w \frac{u_{w-s}}{\sigma_c} + (\frac{p_s}{\sigma_c} + \delta_w \frac{p_{w-s}}{\sigma_c})V_i)$$

$$= \phi(\bar{e} - \tau W_i)\phi(-\bar{c} + u_s + u_{w-s}\delta_w + p_s V_i + p_{w-s}\delta_w V_i)$$

where in the second line we still use the same notations for the structural parameters but they should be interpreted as the scaled versions of the original ones. In our MLE analysis, equation (11) serves as the likelihood function. Remember that $y_i$, $W_i$, $V_i$ and $\delta_w$ are data and $\bar{e}$, $\bar{c}$, $u_s$, $u_{w-s}$, $p_s$ and $p_{w-s}$ are
Table 5: Result of MLE Analysis

<table>
<thead>
<tr>
<th></th>
<th>$\tilde{e}$</th>
<th>$-\tau$</th>
<th>$-\tilde{e} + u_s$</th>
<th>$p_s$</th>
<th>$p_{w-s}$</th>
<th>$u_{w-s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>0.09</td>
<td>$-5.32e - 6^{**}$</td>
<td>$-1.06^{**}$</td>
<td>2.33e - 6</td>
<td>0.58e - 6</td>
<td>$-0.10^{**}$</td>
</tr>
<tr>
<td>se</td>
<td>0.29</td>
<td>$2.31e - 6$</td>
<td>0.16</td>
<td>1.93e - 6</td>
<td>3.42e - 6</td>
<td>0.06</td>
</tr>
<tr>
<td>obs</td>
<td>8193</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

scaled structural parameters. Since the covariates in the two $\phi(\cdot)$ functions are different, all the scaled structural parameters are identified except that $\tilde{e}$ and $u_s$ cannot be separated. The estimation results are reported in Table 5.

(The result is preliminary.) The theory predicts $-\tau < 0$, $p_s > 0$, $p_{w-s} > 0$ and $u_{w-s} < 0$. Our MLE estimates are all consistent with the prediction and the $-\tau$ and $p_{w-s}$ coefficients are both significant under 5% confidence level.

5 Conclusion

We believe our paper is among the first few to use both economic theory and empirical methods to analyze the spread of informational content in a social broadcast setting. We provide a utility-maximization based model to explain network participants’ information-sharing decisions. Moreover, the model incorporates the sociological concept of interpersonal tie strength mathematically, which we think few existing economic models have done. Then we take our model under examination by a real world non-
experimental data set. Our series of econometric analyses are overall supportive of our model.

There are also several other interesting things which are not studied in this research. The social structure information contained in our data set is essentially a snapshot of the constantly changing network environment. So in order to maintain econometric tractability, we model the agents’ decision-making in a single period setting where the social network structure is given and fixed. However, the readers would argue that agents do have dynamic incentives when sharing. For example, an agent may believe that by sharing useful information to others the agent is building a self-image as an informed person, so that more people would like to become her subscribers. Considering incentives like this would inevitably lead us to network dynamics, the research of which is infeasible with our current data set. We devote the study of the interplay between network structure and agents’ behaviors to future research.

Secondly, the observed variations we exploit in the regressions and MLE analysis are different agents’ different sharing decisions facing the same content but not a single agent’s different decisions facing different contents. To do the latter we need to redesign the data collecting strategy. Specifically, we would have to follow a single agent, and track all the contents she receives over a time span and her sharing actions during that period. Currently this is technically infeasible due to the limitation of Twitter API.

One factor that affects a network participant’s sharing behavior, the readers might argue, is the content’s “quality” or the participant’s preference toward that content. Whether a content is a piece of news, an innovative idea or a joke should have different impacts on receivers’ forwarding behaviors. In this paper, we do not explicitly model this effect, but instead leave it in the random error. Doing sentiment analysis on the contents might be an interesting future research.
References


