Article

Communication Research
2019, Vol. 46(2) 250–272

DOI: 10.1177/0093650216682900
journals.sagepub.com/home/crx
SAGE

Network Redundancy and Information Diffusion: The Impacts of Information Redundancy, Similarity, and Tie Strength

Hai Liang[1] and King-wa Fu[2]

Abstract

It remains controversial whether community structures in social networks are beneficial or not for information diffusion. This study examined the relationships among four core concepts in social network analysis—network redundancy, information redundancy, ego-alter similarity, and tie strength—and their impacts on information diffusion. By using more than 6,500 representative ego networks containing nearly 1 million following relationships from Twitter, the current study found that (1) network redundancy is positively associated with the probability of being retweeted even when competing variables are controlled for; (2) network redundancy is positively associated with information redundancy, which in turn decreases the probability of being retweeted; and (3) the inclusion of both ego-alter similarity and tie strength can attenuate the impact of network redundancy on the probability of being retweeted.

Keywords

information diffusion, network redundancy, information redundancy, social network, Twitter

Information diffusion in social networks has long been an interest in communication research (e.g., Greenberg, 1964; Katz & Lazarsfeld, 1955; Rogers, 1995). Although scholars generally believed that social networks are vital for the spread of contagious

[1]The Chinese University of Hong Kong, Shatin, Hong Kong
[2]The University of Hong Kong, Pokfulam, Hong Kong

Corresponding Author:
King-wa Fu, Journalism and Media Studies Centre, The University of Hong Kong, Room 206, Eliot Hall, Pokfulam, Hong Kong.
Email: kwfu@hku.hk

ideas and behaviors, it is far from conclusive how the community structures in social networks affect information flow. Some scholars argue that tightly connected networks with many redundant contacts (i.e., network redundancy) are less efficient for information diffusion (Burt, 1992, 2005; Granovetter, 1973, 1983), whereas other scholars argue that redundant contacts in social networks provide multiple exposures, thus facilitating information diffusion (e.g., Bakshy, Rosenn, Marlow, & Adamic, 2012; Centola, 2010; Hodas & Lerman, 2014).

This controversy relies on various theoretical assumptions that require further examination. First, previous studies assume that network redundancy naturally implies information redundancy (Burt, 1992), which in turn inhibits information flow in social networks. However, there is no empirical evidence directly testing the mediating role of information redundancy. Second, competing factors, such as tie strength and ego-alter similarity, could serve as alternative explanations for the impact of network redundancy on information diffusion (see Harrigan, Achananuparp, & Lim, 2012).

First, strong ties tend to exist in network triads and are less appropriate for access to new information (Granovetter, 1973). In this sense, network structures with redundant contacts are expected to be negatively associated with information diffusion due to the confounding effect of tie strength. Second, according to the homophily principle, users in social networks tend to bond more with those who are similar to them. And the existence of such similarity can facilitate information diffusion (Brown & Reingen, 1987).

Therefore, the present study aims to test the competing roles of information redundancy, ego-alter similarity, and tie strength between network redundancy and information diffusion. In doing so, we collected thousands of representative ego networks from a popular social media platform—Twitter. We measured information redundancy explicitly by using text mining techniques for each individual with regard to their ego. And then we examined the relationships between network redundancy, information redundancy, information similarity, and tie strength at the dyadic level. Finally, by using multilevel mediation analysis, we examined the relative importance of information redundancy, similarity, and tie strength as potential mechanisms relating network redundancy to information diffusion.

Information Diffusion in Online Social Networks

Social scientists have long recognized the importance of social networks in the spread of information (e.g., Burt, 1992; Granovetter, 1973; Rogers, 1995). The emergence of social media has twofold implications for information diffusion studies. First, social media have made social networks ubiquitous, which is significant in that those networks play a major role in the diffusion of information by increasing the spread of novel information and diverse viewpoints (Guille, Hacid, Favre, & Zighed, 2013). It is estimated that majorities of Twitter and Facebook users consider that each platform serves as a source for news (Barthel, Shearer, Gottfried, & Mitchell, 2015). Second, the wide availability of data from online social media platforms enables researchers to investigate diffusion patterns in more empirical-driven and fine-grained ways (Golder

& Macy, 2014). Those data offer a rich source of evidence for studying the dynamics of individual and group behavior, the structure of networks, and global patterns of flow of information through them (Lerman & Ghosh, 2010).

On social media, the diffusion process can be represented by individual *A* sharing a message *B* posted by an individual *C* (i.e., $A \xleftarrow{B} C$), where *A* and *C* are usually connected via the "following" function on social media. The following relationship can be directed (e.g., Twitter) or undirected (e.g., Facebook). Information dissemination on social media platforms is achieved by the function of "retweet" or "share." If *A* retweets *C*'s messages, it indicates that information flows from *C* to *A*. Researchers show different interests in this process. For example, studies focused on content, that is, the role of *B*, found that the presence of hashtags or URLs in the messages influences the probability of being retweeted on Twitter (Liang & Fu, 2015); studies focused on the user being retweeted, that is, the role of *C*, found that opinion leaders with many followers and other individual characteristics are more likely to be retweeted (Bakshy et al., 2012; Suh, Hong, Pirolli, & Chi, 2010); and studies have also focused on the formation of information cascades at the network level (e.g., Goel, Anderson, Hofman, & Watts, 2015).

The current study focuses on the impact of network structures on information diffusion, which is assessed by the relationship between the dyadic or triadic relationships between *A* and *C* in *A*'s ego network (i.e., community structures) and whether *A* retweeted *C*'s messages (information diffusion). An ego network is defined as a part of a social network formed from a focal node, termed ego, and the other persons whom the ego has followed, termed alters. Figure 1 illustrates this relationship using Twitter as an example. Users (egos) spread (the solid lines in Figure 1) any messages posted by the users they have followed.

Based on this framework, researchers have found several important predictors for information diffusion between egos and alters, such as the similarity of demographic characteristics between egos and alters (e.g., Brown & Reingen, 1987), and various measures of the centrality of alters (e.g., Cha, Haddadi, Benevenuto, & Gummadi, 2010; Suh et al., 2010), as well as tie strength between egos and alters (e.g., Brown & Reingen, 1987; Friedkin, 1982). These studies are based on the argument that individuals may be more likely to exhibit behavior similar to that of their friends because of peer influence (Bakshy et al., 2012). However, most previous studies simply inferred network structures from information flow without knowing the underlying social networks. As a result, it remains unknown how network structures affect the dynamics of information diffusion (Lerman & Ghosh, 2010; Ma, Lee, & Goh, 2014).

Network Redundancy and Information Diffusion

Many studies have discussed the negative role of community structures in the spread of ideas (see Harrigan et al., 2012). Among the various characteristics of social networks, at the heart of these studies is the impact of network redundancy on information flow. Network redundancy is a measure to quantify the level of redundant contacts

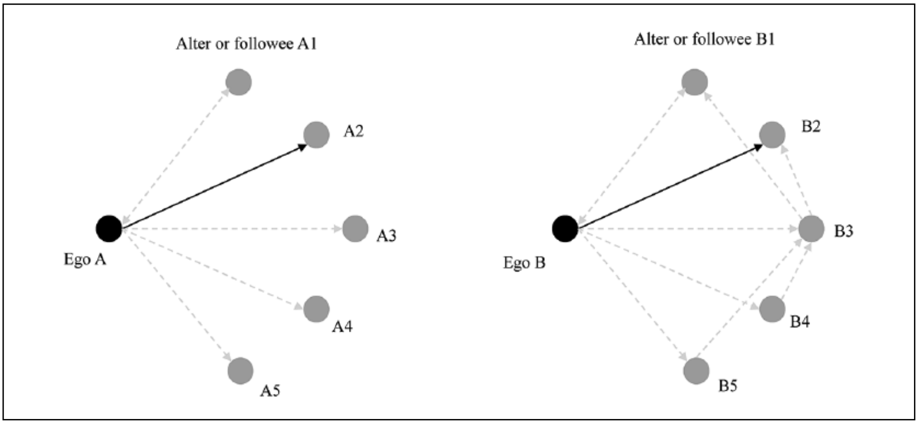

Figure 1. An illustration of network structure and information diffusion in Twitter ego networks.

Note. Egos and alters are Twitter users. The dashed lines indicate the egos' all following relationships (e.g., *A* follows *A3*), while the solid lines indicate that the egos have retweeted the corresponding alters' messages (e.g., *B* retweeted *B2* and the information flows from *B2* to *B*). The following relationship could be reciprocal (e.g., *A-A1, B-B1*) or nonreciprocal (e.g., *A-A3*). *B1* and *B2* are the shared followees between Ego *B* and Alter *B3*. *B4* and *B5* are the mediated followees between Ego *B* and Alter *B3*.

in ego networks. It refers to the extent to which an alter is structurally redundant with ego's other alters. According to Burt (1992), a structurally redundant network is a network whose neighbors are themselves tightly connected. That pattern predicts that individuals in low redundancy networks have advantages compared with those with higher redundancy.

In undirected networks, network redundancy is a prerequisite for measuring Burt's effective size and network constraints (see Burt, 2005). Alter *j*'s network redundancy with ego *i* is given by $R_{ij} = M_{jq} / N$, where N is the total number of alters for ego i, and M_{jq} is the number of ego's alters that are also connected with j ($j \neq q$). In Figure 1, regardless of the tie direction, network redundancy for Alter *A2* is 0 because *A2* is not connected with any other alters. Network redundancy for Alter *B3* is 4/5, because *B3* is connected with four alters. Originally, network redundancy is measured at the network level, which is the sum of the network redundancy values of all alters in an ego network. At the network level, network redundancy is proportional to network density measured by the number of ties among alters. In Figure 1, the redundancy of Network B is higher than Network A. Network B is tightly connected and Network A is a sparse network. The present study will measure network redundancy for each alter with regard to its ego to investigate its impact on the probability information flow at the dyadic level rather than at the network level.

Social networks on many social media platforms, like Twitter, are directed. It is necessary to extend the measure of network redundancy for directed networks. As presented in Figure 1, *B1* and *B2* are two shared followees between Ego *B* and Alter

B3. That suggests that *B* and *B3* shared common information sources in the ego network, and thus, they may receive similar messages. Therefore, *B3* might be redundant for Ego *B*. In addition, *B4* and *B5* are two intermediaries between *B* and *B3*. Messages posted by *B3* could be retweeted by *B4* and *B5* and thus later were received by Ego *B*. In this sense, *B3* is redundant for Ego *B*.

Network redundancy might have negative impacts on information diffusion (see Harrigan et al., 2012). It is argued that messages passed within a tightly connected community will tend to be redundant, and therefore lack novelty. Such lack of novelty lowers the incentive of senders to spread messages as well as the interest of recipients in receiving such messages (Burt, 1992, 2005; Granovetter, 1973, 1983). Nevertheless, the empirical finding by Harrigan et al. (2012) suggests that densely connected community structure is positively associated with diffusion. Under a different theoretical framework, Haythornthwaite (1996) argued that information in sparse networks can flow through only one route, whereas information in tightly connected networks can flow from and to a number of different actors. Information is expected to flow more freely among members of a closed network than a sparse network. In addition, Harrigan et al. (2012) summarized two more explanations for the positive effect of community structures. The similarity explanation states that individuals within a community tend to be similar, thus increasing the relevance of each other's messages. The social bonding explanation considers information sharing as a form of social bonding and further hypothesizes that individuals sharing messages within communities tend to get much higher social bonding rewards from such messaging behavior.

Despite the theoretical controversy, empirical studies, in general, support the idea that network redundancy is beneficial for the diffusion of information or behaviors. Jenssen and Greve (2002) found that redundancy is positively related to access to information resources for business start-ups. Reagans and McEvily (2003) found that social cohesion will be positively associated with ease of knowledge transfer. They argued that two individuals are more willing to share knowledge with each other when they are surrounded by strong third-party connections. In an online experiment, Centola (2010) demonstrated that health behavior in networks with a high level of clustering created by redundant ties spreads farther and faster than in randomly created networks. Harrigan et al. (2012) found that community structures (mutual ties and triads) are positively correlated with retweeting on Twitter. They further concluded that structural redundancy increases message contagion. Therefore, we hypothesize that alters with higher network redundancy are more likely to be retweeted by their egos.

Hypothesis 1 (H1): Network redundancy is positively associated with the probability of alters' tweets being retweeted by their egos.

Competing Explanations

There are several competing explanations for the relationship between network redundancy and information diffusion. In Burt's theory, information redundancy is naturally implied in the measure of network redundancy. It assumes that if an ego's alters are in contact with each other (network redundancy), these alters may possess the same information because they talk to each other, thus creating information redundancy (Burt, 1992, 2005). Information redundancy refers to the similar messages an ego received from its alters in social networks. Such redundancy reduces the novelty of messages and further reduces the likelihood of sending and receiving such messages (see Harrigan et al., 2012).

Two gaps exist in this theoretical justification. First, there is no clear evidence suggesting that network redundancy naturally implies information redundancy. The relationship is largely assumed by scholars (e.g., Burt, 1992, 2005). Network redundancy is usually measured at the network level to predict informational variables. For example, people with less redundant contacts in their ego networks are more likely to have ideas evaluated as valuable (Burt, 2004). The theoretical assumption in these empirical models is that redundant information comes from the redundant contacts in the ego networks, which is at the dyadic level (see Burt, 1992). Second, there are contradictory findings on the impact of information redundancy on information diffusion. Wu and Huberman (2007) found that content novelty, which was considered to be the opposite of information redundancy, attracts user attention on digg.com. Yang and Leskovec (2010) observed that novelty exerts a strong force on the adoption of short textual phrases in online news media. Cha et al. (2010) found that the most retweeted users were content aggregation services and hypothesized that this was precisely because of lack of redundancy of their posts.

On the contrary, the complex contagion hypothesis posits that repeated exposures to an idea are particularly crucial for adopting novel ideas (Centola, 2010; Centola & Macy, 2007). Bakshy et al. (2012) confirmed that the ego's number of sharing friends was positively associated with the ego's probability of sharing the message in an online experiment conducted on Facebook. Later studies on Twitter also found that successive exposures indeed increase the probability that the user will begin mentioning specific hashtags (Romero, Meeder, & Kleinberg, 2011) or URLs (Hodas & Lerman, 2014).

In order to examine whether information redundancy is a mechanism between network redundancy and information diffusion, we need to formally test whether network redundancy is associated with information redundancy and whether information redundancy is associated with information diffusion and in what direction. Thus, we ask the following:

> **Research Question 1 (RQ1):** Is the association between network redundancy and the probability of being retweeted mediated by information redundancy in ego networks?

Tie strength is another factor that can confound the relationship between network redundancy and information flow. According to Granovetter (1973), the strength of a tie is a combination of amount of time, emotional intensity, intimacy, and reciprocity. The earlier research by Granovetter (1973, 1983) mixed the impact of tie strength with tightly connected community structures on information diffusion. Granovetter proposed that weak ties rather than strong ties are more appropriate for access to new information. Nevertheless, subsequent empirical studies have reaffirmed the strength of strong ties for information diffusion (see Haythornthwaite, 1996). Friedkin (1982) found that strong ties are more important than weak ties in promoting information flow about work activities within organizations, whereas weak ties are more important for information flow outside a group. Strong ties can lead to more effective communication; have greater motivation, time, and effort to be of assistance; and also facilitate the formation of trust (see Reagans & McEvily, 2003). Similarly, Harrigan et al. (2012) argued that sharing messages on social media is a way of signaling respect, solidarity, common values, and shared social identity. They conceptualized information sharing as a form of social bonding that is thus related to strong ties. Through an online experiment on Facebook, Bakshy et al. (2012) found that tie strength is positively associated with an ego's probability of sharing a link given that their friends have shared the link previously. Weak ties expose friends to information they would not have otherwise shared.

Another point is that Granovetter's original claim suggests that tie strength affects information diffusion through information redundancy. Strong ties tend to exist in triads (closed structures), and triads tend to act like echo chambers, thereby reproducing redundant messages. Meanwhile, weak ties are argued to have access to more diverse information because they are expected to have fewer mutual contacts (Granovetter, 1973, 1983). Burt (1992) made a conceptual separation between tie strength and community structures. It is argued that it is network redundancy, and not tie strength, that yields information redundancy, which further inhibits information diffusion. This indicates that tie strength is irrelevant to information redundancy. Therefore, it is unlikely that the negative impact of tie strength on information flow is due to the increase of information redundancy. However, the relationship between tie strength and information redundancy has never been empirically tested.

Besides, network redundancy might imply strong ties in ego networks. As argued by Granovetter (1973, 1983), strong ties usually exist in triads. In tightly connected ego networks, the common friendships shared by ego and alters could either induce strong ties between individuals or reflect strong ties between individuals. Using a mobile communication data set, Onnela et al. (2007) confirmed this hypothesis. The majority of the strong ties are found within the clusters, indicating that users spend most of their time talking to member of their close friends. In contrast, most links connecting different communities are weaker than the links within the communities. However, the co-occurrence of network redundancy and strong ties does not imply any causal relationships between the variables. If the strong ties are induced by densely connected structures, tie strength could serve as a mediator between network redundancy and information diffusion. If network redundancy merely reflects the strength of social ties, tie strength is a confounder. Although it is not easy to disentangle the direction of causality, we are more interested in how tie strength can influence the relationship between network redundancy and information diffusion. Statistically, mediation and confounding are identical and can be distinguished only on conceptual grounds (MacKinnon, Krull, & Lockwood, 2000). Therefore, we hypothesize the following:

> **Hypothesis 2 (H2):** The positive association between network redundancy and the probability of being retweeted is weaker when tie strength is controlled for.

In addition to information redundancy and tie strength, homophily could be another mechanism that bridges network redundancy and information diffusion. Similarity is closely correlated with network redundancy. According to the homophily principle, people are more inclined to connect with similar others in terms of certain attributes, such as age, sex, education, and social status (Rogers, 1995). As a consequence of interaction in these homogeneous constructs, individuals tend to become interpersonally tied to each other. Eventually, social networks usually contain subgroups where individuals are more densely connected with each other (De Choudhury, Sundaram, John, Seligmann, & Kelliher, 2010; Feld, 1981).

Ego-alter similarity can facilitate information diffusion in social networks. There are several ways that similarity can influence information diffusion. First, Granovetter (1973) suggested that tie strength is positively correlated with similarity. Strong ties usually imply that the egos and alters are similar in terms of different attributes. Therefore, tie strength is a mediator bridging similarity and information diffusion. Second, similar individuals are more likely to interact with each other. So homophilous ties may have a greater chance of being activated for information flow (see Brown & Reingen, 1987). Third, similar sources of information may be perceived as more credible than different ones (Rogers, 1995), and thus, similar ties are more likely to be activated for information flow (Brown & Reingen, 1987). Even though there are many attributes (e.g., sex and age) that could be used to measure similarity, they are usually invisible on social media platforms. We could measure the semantic similarity based on the stories the egos and alters tweeted previously. This kind of similarity could reflect the common interest between a pair of users (Ma et al., 2014), which is likely to facilitate the information flow between them.

As we discussed above, the ego-alter similarity might be positively associated with network redundancy and information diffusion. However, the ego-alter similarity could serve as both mediator and confounder between network redundancy and information flow. As a confounder, similarity causes network redundancy through the homophily mechanism and facilitates information flow between similar users. As a mediator, similarity is caused by network redundancy through social influence (e.g., Lewis, Gonzalez, & Kaufman, 2012), and then it facilitates information flow. No matter what, we will expect that the inclusion of ego-alter similarity will partially explain the positive relationship between network redundancy and information flow. Therefore,

> **Hypothesis 3 (H3):** The positive association between network redundancy and the probability of being retweeted is weaker when ego-alter similarity is controlled for.

Method

Data Collection

Our data set was collected by using Twitter's REST APIs (Application Program Interfaces). To overcome the representativeness problem, this study sampled Twitter accounts randomly from the population. First, we employed a method reported in Liang and Fu (2015) to generate random Twitter user IDs. The Twitter ID is a unique value that every account on Twitter has. A list of random Twitter IDs represents a random sample of Twitter users. Using their method, we obtained 34,006 valid user accounts.

Second, we obtained the egos' user profiles, their followees' IDs, and up to 3,200 tweets and retweets (i.e., timeline) for each ego user. Users are not equally active in terms of tweeting. The collected tweets might represent the last month, year, or 5 years of someone's timeline. To control this, we selected the tweets posted in the past 2 weeks. We updated our data in March 2015. Due to the privacy settings on Twitter, we could only get tweets from public accounts (31,883). We further excluded the inactive users who did not post any tweets and did not change their followees in the past 3 months (7,609). As we need to measure informational variables based on the egos' timelines, we removed the egos who did not post any tweets in the past 2 weeks (6,551).

Third, we constructed ego networks in which nodes are users and ties are the following relationships between egos and followees (1,839,660). We further collected the followees' profiles and tweets (in the past 2 weeks), and their followees' followees. We excluded the followees whose tweets and following relationships are kept in private. The final data set for our analyses includes 6,551 ego networks containing 962,859 ego-followee edges.

Measures

Network redundancy of a followee with ego's other followees was measured by two indicators: the proportion of shared followees between the ego and the followee to the total number of the ego's followees, and the proportion of mediated followees between the ego and the followee to the total number of the ego's followees (see Figure 1). The two indicators were normalized by dividing the total number of followees of the ego. As these two indicators are highly correlated (r = .78, p < .01, N = 962,860) and are conceptually relevant, we took the average as a measure of network redundancy on Twitter. Theoretically, the minimal value is 0 (nonredundant) and the maximum value is 1 (purely redundant). As a result, the mean of network redundancy is 0.075 (SD = 0.100, median = 0.037) in our data set.

Information redundancy of a followee/alter with its ego was measured by text mining techniques. First, we created a term-document matrix for each ego and its followees (i.e., 6,115 term-document matrices in total). In each term-document matrix, rows are the users (including an ego and its alters) and columns are the unique words in the users' tweets (all available tweets). For non-English tweets, we translated the text into English automatically using Google Translate API for further analysis according to Lucas et al. (2015). Following the general text mining procedure, we also stemmed the words (e.g., likes - like) and removed the stop words (e.g., a, an). For those tweets including URLs, we expanded the URLs and included them as terms in the matrix. Second, the words used by user u were encoded into a feature vector of term frequency–inverse document frequency (tf-idf) $\varphi(u)$. The ith element $\varphi_i(u)$ represents the frequency of the word indexed by i in all the words used by u, scaled by the inverse document frequency (see Salton, Wong, & Yang, 1975). The tf-idf value is directly proportional to the number of times a word appears in the document but is adjusted to be inversely proportional to the word frequency in the corpus (Jain, Agarwal, & Pruthi, 2015).

Information redundancy was measured based on the tf-idf values. Words are not equally important in terms of their uniqueness. The tf-idf is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Robertson, 2004). It quantifies the word importance relative to all words used by alters in an ego network. If a followee has many words with high tf-idf values, it means that the user included many unique words and showed less redundancy. Therefore, we calculated the sum of all tf-idf values for each followee as an indicator of information uniqueness (IU). And then we subtract the value using the maximum value to measure information redundancy (IR = max(IU) − IU). We further normalized the raw score to range from 0 to 1 by using the formula: (IR − min(IR)) / (max(IR) − min(IR)). As a result, the mean of information redundancy is 0.645 (SD = 0.186, Mdn = 0.699).

Ego-alter similarity was measured at the dyadic level by the semantic similarity between the tweets posted by the egos and those posted by their alters. As a result, the score quantifying the similarity of information posted by two users u_1 and u_2 is given by a cosine similarity measure: $\phi(u_1), \phi(u_2) / (\| \phi(u_1) \| \cdot \| \phi(u_2) \|)$. Theoretically, ego-alter similarity ranges from 0 (completely dissimilar) to 1 (actually the same). The mean of the similarity is 0.015 (Mdn = 0, SD = 0.040). If two words are referring to similar ideas (e.g., #love and #like), users are inclined to use them together, and thus, the semantic similarity score between them will be high. Due to the highly skewed distribution, we used the square root in formal analyses.

Tie strength was measured by three indicators: whether the followee has been mentioned by the ego, whether the followee has been replied to by the ego, and reciprocity between the ego and the followee. The first two indicators are measures of communication frequency. If a followee of an ego is also a follower of the ego, we said the tie is reciprocal. Among the 962,859 ties, 33% are reciprocal, 4% of followees have been mentioned, and 9% of followees have been replied to. Finally, we operationalized strong ties as those edges that are reciprocal and for which the followees have been mentioned or replied to at least once (9.2%). Overall, only 4.4% were considered as strong ties in our data.

Being retweeted was measured at the edge level to indicate whether a followee has been retweeted by its ego at least once in the ego's posts during our observation. For any retweets, Twitter official API only returns the users who originally posted the tweets rather than the users from whom the egos directly retweeted. For example, Ego *A* retweeted a tweet originally posted by User *B* from *A*'s followee *C*. The Twitter API returns that *A* retweet *B* instead of *C*. To solve this problem, we examined the timelines of all *A*'s followees to see how many users actually retweeted the same tweet. If more than one followee have retweeted the same tweet, we considered the most recent one as the user that the ego retweeted from. Among the 962,859 ego-followee edges, only 1.3% of the followees have been retweeted by their egos in the past 2 weeks.

We included two types of control variables. The ego-specific predictors are the characteristics of ego users, whereas alter-specific predictors are the characteristics of the followees. Ego-specific variables include the number of followers (*M* = 2,828, *Mdn* = 221, *SD* = 7,969), the number of followees (*M* = 2,332, *Mdn* = 619, *SD* = 5,051), the number of tweets (*M* = 7,358, *Mdn* = 686, *SD* = 41,771), and years since registration (*M* = 3.22, *Mdn* = 3, *SD* = 1.6), all of which are directly provided by Twitter's profile API. Alter-specific variables include the number of followers (*M* = 1,156,000, *Mdn* = 26,500, *SD* = 4,873,998), the number of followees (*M* = 18,820, *Mdn* = 609, *SD* = 82,667), the number of tweets (*M* = 16,740, *Mdn* = 5,820, *SD* = 39,510), years since registration (*M* = 4.08, *Mdn* = 4, *SD* = 1.79), and tweeting frequency (i.e., the number of tweets posted in the last 2 weeks, *M* = 137, *Mdn* = 35, *SD* = 323).

Data Analysis

We used multilevel generalized linear models (Snijders & Bosker, 2012) to test our hypotheses. The multilevel framework has been successfully employed to model egocentric networks (e.g., Golder & Yardi, 2010). The strength of multilevel modeling lies in its capability to take into account internal homogeneity within groups. In our study, the unit of analysis is the tie between egos and alters. Each retweeting relationship nested under the same ego user could be influenced by the unique characteristics of that particular ego. All tie measures and alter-specific measures are Level 1 variables. All ego-level predictors are Level 2 variables.

To formally test the confounding or mediation effects of information redundancy, similarity, and tie strength on the relationship between network redundancy and retweeting, we employed the multilevel mediation analysis developed by Imai, Keele, and Tingley (2010) and Tingley, Yamamoto, Hirose, Keele, and Imai (2014; given that the two types of effects are statistically identical). We considered network redundancy as the treatment variable (T) and let $M_i(t)$ denote the potential value of mediator (information redundancy, similarity, or tie strength) for interest for unit i under treatment status $T_i = t, t \in [0,1]$. The total unit treatment effect is given by $\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$. The causal mediation effects are represented by $\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0))$, and the direct effects of the treatment are $\zeta_i(t) = Y_i(1, M_i(t)) - Y_i(1, M_i(t))$ (see Imai et al., 2010). According to this operationalization, $\delta_i(t)$ measures the difference of the probabilities of being retweeted given different mediator values for a single unit i. We are mainly interested in whether the average of causal mediation effects, that is, $\bar{\delta}(t) = \text{mean}(\delta_i(t))$, is statistically significant.

Results

Bivariate Analysis

We examined the relationships among network redundancy, information redundancy, ego-alter similarity, tie strength, and the probability of being retweeted. As expected, network redundancy and information redundancy are positively correlated (Spearman's rho = 0.22, $p < .01$). Network redundancy is also related to strong ties. In our data, the average of network redundancy for strong ties is 0.11, while the value is significantly smaller for weak ties (0.07, $F = 8{,}525$, $p < .01$). However, the bivariate analysis suggests that the correlation between network redundancy and similarity is not significant (Spearman's rho = 0.01, $p = .07$).

Furthermore, both network redundancy and ego-alter similarity are positively associated with being retweeted. On average, the users being retweeted have 0.11 network redundancy and 0.09 similarity, compared with 0.07 network redundancy ($F = 1{,}858$, $p < .01$) and 0.01 similarity ($F = 52{,}913$, $p < .01$) for those users not being retweeted. Yet, the users being retweeted have lower values in information redundancy (0.71 vs. 0.75, $F = 466$, $p < .01$). In addition, users connected through strong ties are more likely to be retweeted than users connected through weak ties. The odds ratio of being retweeted for strong ties is 0.07 (i.e., the number of strong ties being retweeted divided by the number of strong ties not being retweeted), whereas the ratio for weak ties is 0.01 ($\chi^2 = 8{,}373$, $df = 1$, $p < .01$).

Multilevel Analysis

We conducted a series of multilevel models (see Table 1) to examine the role of network redundancy in information diffusion. The dependent variable is whether a followee has been retweeted at least once or not. Due to the multilevel structure of our data, we first calculate the intraclass correlation coefficient (ICC) based on a null model including only the intercept. The ICC value is fairly high (55%), indicating that 55% of the variance could be explained at the ego (second) level. In other words, it suggests that whether a user will be retweeted by its followers mainly depends on the characteristics of their followers. More important, it indicates the necessity of using multilevel modeling.

In the full model (Model IV), network redundancy is positively correlated with the probability of being retweeted, which indicates that followees who are structurally redundant are more likely to be retweeted by their ego users ($B = 1.72$, $SE = 0.11$, $p < .01$). That means the odds of being retweeted for users who are purely network redundant (i.e., the value is 1) are about 5 times higher than the odds for those users with zero network redundancy, because $\exp(1.72) = 5.58$. Therefore, H1 is confirmed.

Information redundancy is negatively associated with the probability of being retweeted ($B = -1.51$, $SE = 0.09$, $p < .01$). Holding all other variables at a fixed value, we will see a 22% decrease in the odds of being retweeted for a 0.1 increase in information redundancy. Meanwhile, the odds of being retweeted by their ego users for the followees connected via strong ties are more than 2 times higher than the odds for the

Table 1. Multilevel Logistic Regression Models Predicting Being Retweeted.

	Model I Estimate (SE)	Model II	Model III	Model IV
Edge/alter-specific factors (first level)				
Network redundancy	2.33**	2.13**	1.60**	1.72**
	(0.11)	(0.11)	(0.11)	(0.11)
Information redundancy	−2.59**	−1.47**		−1.51**
	(0.08)	(0.10)		(0.09)
Strong vs. weak tie	0.95**		0.90**	0.91**
	(0.03)		(0.03)	(0.03)
Ego-alter similarity		7.46**	7.66**	7.42**
		(0.10)	(0.09)	(0.10)
No. of followers[a]	−0.11**	−0.10**	−0.07**	−0.06**
	(0.00)	(0.00)	(0.00)	(0.00)
No. of followees[a]	−0.06**	−0.06**	−0.06**	−0.06**
	(0.01)	(0.01)	(0.00)	(0.01)
No. of statuses[a]	−0.11**	−0.09**	−0.11**	−0.11**
	(0.01)	(0.01)	(0.01)	(0.01)
Years since registration	−0.07**	−0.06**	−0.06**	−0.06**
	(0.01)	(0.01)	(0.01)	(0.01)
Tweeting frequency[a]	0.56**	0.42**	0.43**	0.42**
	(0.01)	(0.01)	(0.01)	(0.01)
Ego-specific factors (second level)				
No. of followers[a]	0.01	−0.01	−0.01	−0.01
	(0.02)	(0.02)	(0.02)	(0.02)
No. of followees[a]	−0.85**	−0.83**	−0.75**	−0.80**
	(0.03)	(0.03)	(0.03)	(0.03)
No. of statuses[a]	0.30**	0.26**	0.23**	0.23**
	(0.01)	(0.01)	(0.01)	(0.01)
Years since registration	−0.10**	−0.09**	−0.10**	−0.10**
	(0.01)	(0.02)	(0.02)	(0.02)
Intercept	0.24	−0.91**	−2.39**	−1.02**
	(0.13)	(0.15)	(0.12)	(0.15)
Model summary				
Variance of intercepts (*SD*)	1.21	1.37	1.38	1.38
	(1.10)	(1.17)	(1.18)	(1.18)
Log-likelihood	−46,827	−43,814	−43,566	−43,436
AIC	93,682	87,656	87,160	86,903
Conditional R^2	47.0%	50.5%	49.5%	50.3%
No. of ties	962,859			
No. of ego users	6,551			

Note. AIC = Akaike information criterion.
[a]Variables were log-transformed for multilevel analyses. The square root of similarity was used.
$^{*}p < .05$. $^{**}p < .01$.

followees connected via weak ties ($B = 0.91$, $SE = 0.03$, $p < .01$). In addition, ego-alter similarity is positively associated with the odds of being retweeted ($B = 7.42$, $SE = 0.10$, $p < .01$).

Furthermore, Models I to IV consistently suggest that network redundancy is positively associated with the probability of being retweeted. Comparing Model I with Model IV, we found that the inclusion of ego-alter similarity attenuates the impact of network redundancy on the probability of being retweeted (2.33 vs. 1.72, $Z = 3.90$, $p < .01$). This is consistent with H3. Comparing Model II with Model IV, we found that the inclusion of tie strength attenuates the impact of network redundancy (2.13 vs. 1.72, $Z = 2.55$, $p < .05$). This is consistent with H2. Comparing Model III and Model IV, we found that the exclusion of information redundancy actually suppresses the impact of network redundancy on the probability of being retweeted. However, the difference is not statistically significant (1.60 vs. 1.72, $Z = -0.77$, $p > .05$).

Mediation Analysis

According to Table 1, we found that the impacts of network redundancy on information diffusion are attenuated by including similarity and tie strength, whereas the impact is amplified by the inclusion of information redundancy. It appears that the ego-alter similarity has a larger effect than tie strength, because the coefficient of network redundancy decreases. Table 2 further shows that network redundancy is positively associated with information redundancy when all other variable values are fixed (see Model V: $B = 0.10$, $SE = 0.00$, $p < .01$). Model VI suggests that structurally redundant users are more likely to be the users who connected via strong ties ($B = 5.43$, $SE = 0.08$, $p < .01$). Model VII suggests that network redundancy is also positively associated with information similarity ($B = 0.07$, $SE = 0.00$, $p < .01$). Findings are summarized in Figure 2. They suggest that network redundancy could exert its influence on information diffusion through information redundancy, similarity, and tie strength.

To formally test the mediation effects of information redundancy, tie strength, and similarity, we employed multilevel mediation analysis. Before we present the formal results, we should note that the explanatory factors could be both mediators and confounders. Concerning RQ1, we tested the mediating role of information redundancy. The average direct effects of network redundancy $(\bar{\zeta}(t))$ on information diffusion is 0.030, and the 95% confidence interval is [0.025, 0.036]. The average causal mediation effects $(\bar{\delta}(t))$ via information redundancy is −0.003, and the 95% confidence interval is [−0.003, −0.002]. The proportion of mediated effects via information redundancy is −10.4%, and the confidence interval is [−12.2%, −8.6%]. The findings suggest that the mediation effect of information redundancy is statistically significant and the direction of the indirect effect is negative.

Concerning H2, we tested the mediating role of tie strength. The average direct effect of network redundancy is 0.036, and the 95% confidence interval is [0.030, 0.043]. The average of mediation effects is 0.011, and the 95% confidence interval is [0.010, 0.012]. The proportion of mediated effect via tie strength is 23.3%, and the

Table 2. Multilevel Regression Models Predicting Information Redundancy, Similarity, and Strong Ties.

	Model V Information redundancy Estimate (SE)	Model VI Strong vs. weak ties	Model VII Ego-alter similarity
Network redundancy	9.86×10^{-02}**	5.43**	7.13×10^{-02}**
	(2.31×10^{-03})	(0.08)	(1.12×10^{-03})
Information redundancy		0.54**	-1.41×10^{-01}**
		(0.05)	(4.66×10^{-04})
Strong vs. weak tie	7.56×10^{-03}**		1.22×10^{-02}**
	(8.64×10^{-04})		(4.13×10^{-04})
Ego-alter similarity	-6.13×10^{-01}**	1.26**	
	(2.03×10^{-03})	(0.07)	
No. of followers[a]	-6.33×10^{-03}**	−0.80**	-4.18×10^{-03}**
	(6.91×10^{-05})	(0.01)	(3.30×10^{-05})
No. of followees[a]	-1.31×10^{-03}**	0.38**	-6.67×10^{-04}**
	(8.15×10^{-05})	(0.01)	(3.90×10^{-05})
No. of statuses[a]	3.96×10^{-03}**	0.36**	1.97×10^{-03}**
	(1.38×10^{-04})	(0.01)	(6.61×10^{-05})
Years since registration	-5.08×10^{-03}**	−0.03**	-1.51×10^{-03}**
	(1.26×10^{-04})	(0.01)	(6.02×10^{-05}**)
Tweeting frequency[a]	4.82×10^{-04}**	0.04**	1.03×10^{-02}**
	(1.18×10^{-04})	(0.00)	(5.54×10^{-05}**)
No. of followers[a]	3.357×10^{-03}**	0.20**	2.09×10^{-03}**
	(9.06×10^{-04})	(0.03)	(7.44×10^{-04})
No. of followees[a]	-2.90×10^{-02}**	−0.81**	-5.50×10^{-03}**
	(9.96×10^{-04})	(0.03)	(7.99×10^{-04})
No. of statuses[a]	6.05×10^{-03}**	0.54**	5.38×10^{-03}**
	(5.26×10^{-04})	(0.02)	(4.33×10^{-04})
Years since registration	-5.15×10^{-04}	0.21**	-1.27×10^{-04}
	(5.92×10^{-04})	(0.02)	(4.88×10^{-04})
Intercept	8.49×10^{-01}**	−4.58**	1.59×10^{-01}**
	(4.00×10^{-03})	(0.13)	(3.08×10^{-03})
Variance of intercepts (*SD*)	0.004	1.66**	0.003
	(0.066)	(1.29)	(0.057)
Log-likelihood	432,050	−98,752	1,141,075
Conditional R^2	30.7%	74.8%	47.9%
No. of ties		962,859	
No. of ego users		6,551	

[a]Variables were log-transformed for multilevel analyses. The square root of similarity was used.
*$p < .05$. **$p < .01$.

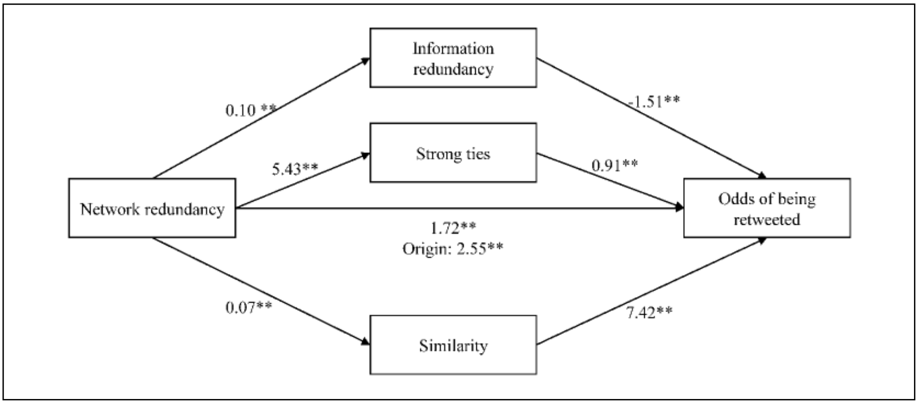

Figure 2. The explanatory roles of information redundancy, tie strength, and similarity.
Note. This figure is for illustrative purpose only. It is not the exact model we fitted in Tables 1 and 2. All control variables are not presented in this figure. According to Table 2, information redundancy, tie strength, and similarity are interrelated. "Origin" here indicates the coefficient of network redundancy including only control variables (SE = 0.11).
$^{*}p < .05$. $^{**}p < .01$.

confidence interval is [21.3%, 26.0%]. The findings confirmed that the indirect effect of tie strength is significantly positive. Therefore, H2 is confirmed. In addition, the mediation effect of tie strength is stronger than the effect of information redundancy (23.3% > 10.4%).

Concerning H3, we tested the mediating role of ego-alter similarity. The average direct effects of network redundancy is 0.035, and the 95% confidence interval is [0.028, 0.042]. The average of mediation effects is 0.012, and the 95% confidence interval is [0.011, 0.013]. The proportion of mediated effect via ego-alter similarity is 25.0%, and the confidence interval is [23.3%, 27.6%]. The findings confirmed that the indirect effect of ego-alter similarity is significantly positive. Therefore, H3 is confirmed. In addition, the mediation effect of ego-alter similarity is stronger than the effect of information redundancy (25.0% > 10.4%).

Discussion

Network Structure and Information Diffusion

Although there are many approaches to testing the impact of network structures on information diffusion, this study focused on one of the core concepts: network redundancy. There are many debates and contradictory findings concerning the topic. Traditional studies were based on small data sets, and social networks were inferred by self-reports on relationships that are less accurate (Ma et al., 2014). Thanks to the popularity of social media platforms, we can reexamine this question in a more objective and reliable way using large-scale online social network data.

First, the current study found that network redundancy is positively correlated with the probability of being retweeted in ego networks. This indicates that alters sharing many followees and intermediaries are more likely to be retweeted by their egos. This finding is largely consistent with previous empirical studies (e.g., Harrigan et al., 2012; Jenssen & Greve, 2002), whereas it is clearly different from the expectations in the classical theories (Burt, 1992; Granovetter, 1973). This discrepancy is related to the following controversies.

Second, Burt's hypothesis is based on the mediating role of information redundancy between network redundancy and information diffusion. It is argued that network redundancy is positively associated with redundant information, which in turn inhibits the flow of novel information (Burt, 1992, 2005). In our study, we confirmed that structurally redundant alters contain more redundant information with regard to their egos. Furthermore, we found that information redundancy is negatively correlated with the probability of being retweeted. Although the inclusion of information redundancy does not change the impact of network redundancy on information diffusion significantly, a formal mediation test suggests that information redundancy does have a negative mediation effect. This finding contradicts the well-known phenomenon of complex contagion in online social networks (Bakshy et al., 2012; Centola, 2010), whereas it echoes the claim that people are likely to retweet posts with novel information in online social networks (Cha et al., 2010; Wu & Huberman, 2007; Yang & Leskovec, 2010). Given the positive direct effect and negative indirect effect of network redundancy on information flow, it clearly suggests that information redundancy is not the only mechanism connecting network redundancy and information flow.

Third, Granovetter's hypothesis is based on the confounding role of tie strength (Granovetter, 1973, 1983). According to Granovetter, tie strength and network structure are two intertwined concepts. This predicts that network redundancy is positively associated with tie strength, which is confirmed by subsequent studies (e.g., Onnela et al., 2007). Our study further confirmed this relationship. However, unlike the original "strength of weak ties" hypothesis, strong ties are more important in promoting information flow in our study. Nevertheless, this finding is consistent with previous diffusion studies on Twitter (Harrigan et al., 2012) and Facebook (Bakshy et al., 2012).

The impact of ego-alter similarity on the relationship between network redundancy and information flow is similar to the impact of tie strength. We demonstrated that the semantic similarity between the tweets posted by egos and alters is positively associated with both densely connected communities and the probability of being retweeted by egos. The positive correlation between network redundancy and information flow could be partially explained by the ego-alter similarity. However, the causal directions are unidentifiable in the current study. We do not know whether the ego-alter similarity (also tie strength) is a mediator or just a confounder. Nevertheless, the mediation effect via similarity is the largest one in the present study.

In summary, network redundancy plays a vital role in information diffusion within ego networks both directly and indirectly. Our study suggests that well-connected communities with less redundant information, homogeneous, and strong ties are found to be more conductive to information flow in Twitter ego networks. Network redundancy shows an independent impact on information diffusion between egos and alters when all competing variables are controlled for. The mediation effect of information redundancy is negative, and the effect size is much smaller than that of tie strength or ego-alter similarity. This further weakens the argument in favor of using information redundancy to explain the relationship between network redundancy and information diffusion in social networks. Instead, tie strength and ego-alter similarity appear to be the more plausible explanations.

Diffusion Within/Across Communities

A major concern of the present study is whether community structures, that is, tightly connected networks, facilitate or inhibit information diffusion in social networks. The present study found that homogeneous community structures with redundancy and strong ties are better for information flow. Yet, we cannot conclude that this is universal. Compared with existing literature, we are inclined to argue that whether tightly connected structures are good or bad for information diffusion is conditional on whether within or across community diffusion is of research interest.

In the current study, we simplified the diffusion process as a chain of $A \xleftarrow{B} C$. Under this framework, the present study investigated the probability of an alter being retweeted by their ego within ego networks. This design could overestimate the importance of community structure in information diffusion. As argued by Friedkin (1982), weak ties are less efficient to promote information within communities than strong ties. Given this, it is highly possible that strong ties are more important in our study because ego networks could be considered as communities. We should note that even though individuals are more likely to be influenced by their strong ties, weak ties are more important in terms of exposing individuals to information they might otherwise not have been exposed to (Bakshy et al., 2012).

In addition, our dyadic analysis is also different from the macro-level analysis. The strength of weak ties in promoting boundary-spanning information flows lies not in their individual efficiency but in their numbers (Friedkin, 1982), because most contagion occurs along weak ties (Bakshy et al., 2012). This is also true in our data. There are 10,001 (79%) retweeting edges that occur along weak ties, whereas only 2,664 (21%) retweeting edges occur along strong ties. This is empirically valid but lacks theoretical implications, because weak ties always outnumber strong ties in social networks.

Second, the retweeting behavior on Twitter needs time and effort. Therefore, users expect social rewards from this behavior. As Harrigan et al. (2012) discussed, reposting a friend's post is one of the compliments in Facebook etiquette, and thus, information diffusion is a form of social bonding reward. Our results are consistent with this argument. We found that tightly connected structures are associated with stronger ties and both of them are positively associated with the probability of being retweeted on Twitter. This social bonding effect might be only correct for information spreading within communities. In tightly connected communities, strong ties dominated and community members are under social pressures to retweet others' posts. In sparse networks, individuals might be less susceptible to the social bonding effect and strong ties might play a less important role than weak ties.

Limitations and Future Research

The current study is subject to several biases that need to be addressed in future studies. First, the mediation analysis suggests plausible causal paths among the variables as we presented in Figure 2. Although all paths are grounded in existing theories, we must note that the figure does not naturally indicate the true casual relationships. For example, individuals may build their personal networks strategically based on information redundancy. In order to reduce information overload, social media users can easily remove the redundant contacts from their networks. Therefore, network redundancy and information redundancy might present a mutual causal relationship. Moreover, network structures and information diffusion could also be mutually determined. For example, individuals are inclined to follow new users who they have retweeted on Twitter (Antoniades & Dovrolis, 2015). We are uncertain about to what extent mutual causality can bias our findings. Future studies are highly encouraged to test the causal relationships among network structures, tie strength, and information flow, which are beyond the scope of the present study. We suggest two workable ways to achieve this goal. The researcher can either use a panel design, run alternative structural equation models to simulate competing explanations, and compare each model's goodness of fit, or run simulations with different diffusion mechanisms, and then compare the results with the observed data.

The second bias may come from the unit of analysis. The present study focused on the dyadic relationships between egos and alters. Although dyadic analysis has its advantages to understand more micro-level behaviors, it could not substitute for network-level analysis. As we mentioned, strong ties are individually more influential. However, most viral messages may be bridged by weak ties. Besides, successful information cascades in social networks may have their global properties that we cannot know from ego network analysis (e.g., Goel et al., 2015).

Third, we need to differentiate retweeting and being retweeted. In the literature on information diffusion, scholars have overemphasized the probability of messages being retweeted by their followers in social networks. Few studies have examined individuals' inclination to retweet others' information. Watts and Dodds (2007) pointed out that large cascades of influence are driven by a critical mass of easily influenced individuals. Similarly, Romero, Galuba, Asur, and Huberman (2011) argued that social media users with more passive followers are more influential. The current study suggests that users with more followers retweeted more frequently, whereas users following more users are less likely to retweet others' messages. A possible explanation is that a large number of followers may encourage a user to inform their audience by retweeting messages. Large numbers of followees may discourage retweeting because of information overload. Future studies should investigate the profiles and incentives of users who are more likely to retweet others' messages.

Fourth, measurement errors may be caused by the current operationalization of the key variables. The present study employed the text mining techniques to measure information redundancy and ego-alter similarity unobtrusively. The potential problem of this kind of operationalization is that it is difficult to evaluate the validity of these measures. The measure of information redundancy could be interpreted in different ways. For example, it might indicate that ego users actually received similar words belonging to a same thematic community. However, it does not necessarily mean that the information in the tweets is redundant for the ego users. It depends on how the users process the information. A possible solution for this limitation is to incorporate a survey into the research design. Besides, future studies are recommended to measure ego-alter similarity by using additional variables, such as user demographics, to increase study validity.

Finally, our study might be subject to platform selection bias. Twitter is a social media platform focusing on information sharing. However, there are platforms emphasizing social networking functions, like Facebook. It remains largely unknown whether the characteristics and purposes of social media platforms can influence the patterns of information diffusion. More important, most theories we mentioned above were originally proposed for explaining offline phenomena. Future studies are encouraged to compare the differences of information diffusion patterns between online and offline social networks. In this way, we can see whether the discrepancy between classical theories and current findings is caused by the mediated contexts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Small Project Funding from the University of Hong Kong (201409176011) and the Public Policy Research Funding Scheme from the Central Policy Unit of the Hong Kong Special Administrative Region Government (2013.A8.009.14A).

References

Antoniades, D., & Dovrolis, C. (2015). Co-evolutionary dynamics in social networks: A case study of Twitter. *Computational Social Networks*, *2*(1), 1. doi:10.1186/s40649-015-0023-6

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). *The role of social networks in information diffusion*. Paper presented at the Proceedings of the 21st International Conference on World Wide Web, WWW 2012, Lyon, France.

Barthel, M., Shearer, E., Gottfried, J., & Mitchell, A. (2015). *The evolving role of news on Twitter and Facebook*. Retrieved from http://www.journalism.org/2015/07/14/the-evolving-role-of-news-on-twitter-and-facebook/

Brown, J. J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, *14*, 350-362.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, *110*, 349-399. doi:10.1086/421787

Burt, R. S. (2005). *Brokerage and closure: An introduction to social capital*. Oxford, UK: Oxford University Press.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*, 1194-1197.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, *113*, 702-734.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010, May). *Measuring user influence in Twitter: The million follower fallacy*. Paper presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC.

De Choudhury, M., Sundaram, H., John, A., Seligmann, D. D., & Kelliher, A. (2010). *Birds of a feather: Does user homophily impact information diffusion in social media?* (arXiv preprint arXiv:1006.1702). Retrived from https://arxiv.org/abs/1006.1702

Feld, S. L. (1981). The focused organization of social ties. *American Journal of Sociology*, *86*, 1015-1035. doi:10.1086/227352

Friedkin, N. E. (1982). Information flow through strong and weak ties in intra-organizational social networks. *Social Networks*, *3*, 273-285. doi:10.1016/0378-8733(82)90003-X

Goel, S., Anderson, A., Hofman, J., & Watts, D. (2015). The structural virality of online diffusion. *Management Science*, *62*, 180-196. doi:10.1287/mnsc.2015.2158

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*, 129-152.

Golder, S. A., & Yardi, S. (2010, August). *Structural predictors of tie formation in Twitter: Transitivity and mutuality*. Paper presented at the IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, MN.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *78*, 1360-1380. doi:10.1086/225469

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, *1*, 201-233.

Greenberg, B. S. (1964). Diffusion of news of the Kennedy assassination. *Public Opinion Quarterly*, *28*, 225-232. doi:10.1086/267239

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, *42*(2), 17-28.

Harrigan, N., Achananuparp, P., & Lim, E. P. (2012). Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. *Social Networks*, *34*, 470-480. doi:10.1016/j.socnet.2012.02.005

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research*, *18*, 323-342.

Hodas, N. O., & Lerman, K. (2014). The simple rules of social contagion. *Scientific Reports, 4*, Article 4343.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309-334.

Jain, N., Agarwal, P., & Pruthi, J. (2015). HashJacker—Detection and analysis of hashtag hijacking on Twitter. *International Journal of Computer Applications*, *114*(19), 17-20.

Jenssen, J. I., & Greve, A. (2002). Does the degree of redundancy in social networks influence the success of business start-ups? *International Journal of Entrepreneurial Behavior & Research*, *8*, 254-267.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence, the part played by people in the flow of mass communications*. Glencoe, IL: Free Press.

Lerman, K., & Ghosh, R. (2010, May). *Information contagion: An empirical study of the spread of news on Digg and Twitter social networks*. Paper presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC.

Lewis, K., Gonzalez, M., & Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 68-72. doi:10.1073/pnas.1109739109

Liang, H., & Fu, K. W. (2015). Testing propositions derived from Twitter studies: Generalization and replication in computational social science. *PLoS ONE*, *10*(8). doi:10.1371/journal.pone.0134270

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*, 254-277. doi:10.1093/pan/mpu019

Ma, L., Lee, C. S., & Goh, D. H. (2014). Understanding news sharing in social media from the diffusion of innovations perspective. *Online Information Review*, *38*(5), 598-615. doi:10.1108/OIR-10-2013-0239

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*, 173-181. doi:10.1023/a:1026595011371

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., . . . Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*, 7332-7336.

Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly*, *48*, 240-267.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*, 503-520. doi:10.1108/00220410560582

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York, NY: The Free Press.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, September). *Influence and passivity in social media*. Paper presented at the European Conference, ECML PKDD 2011, Athens, Greece.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011, March). *Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter*. Paper presented at the Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613-620. doi:10.1145/361219.361220

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: SAGE.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). *Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network*. Paper presented at the IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, MN.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1-38. doi:10.18637/jss.v059.i05

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*, 441-458. doi:10.1086/518527

Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 17599-17601. doi:10.1073/pnas.0704916104

Yang, J., & Leskovec, J. (2010, December). *Modeling information diffusion in implicit networks*. Paper presented at the IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia.

Author Biographies

Hai Liang is assistant professor in the School of Journalism and Communication at the Chinese University of Hong Kong. His research interests include political communication, dynamic communication process, social media analytics, and computational social science.

King-wa Fu is associate professor at the Journalism and Media Studies Centre at the University of Hong Kong. His research interests cover political participation and media use, computational media studies, health and the media, and younger generation's Internet use.