Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Empirical Analysis of the Relation between Community Structure and Cascading Retweet Diffusion

Sho Tsugawa
Faculty of Engineering, Information and Systems, University of Tsukuba
Tsukuba, Ibaraki 305–8573, Japan
s-tugawa@cs.tsukuba.ac.jp

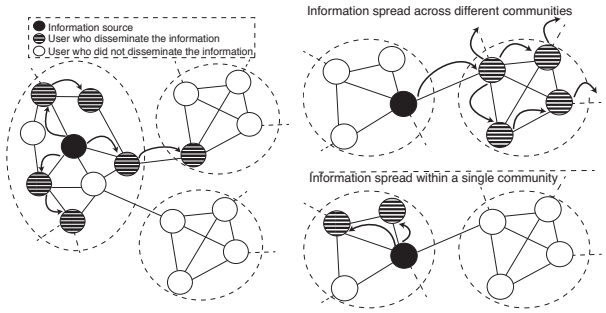
Abstract

Social networks have community structure, in which the network is composed of highly clustered subnetworks (communities) with sparse links between them. Such community structure is expected to affect information diffusion among individuals. This paper empirically investigates how the community structure of a social network among Twitter users affects cascading diffusion of retweets among them. The results show that the frequency of retweets between users who are in the same community is approximately two times that between users who are in different communities. In contrast, the results also show that tweets disseminated via inter-community retweets have future popularity about 1.5-fold that of tweets disseminated via intra-community retweets. By using this fact, we construct classifiers to predict the future popularity of tweets from community-based features as well as features related to influence of users and tweet contents. Our experimental results show that contrary to our expectations, community-based features have little contributions for predicting the future popularity of tweets. This paper discusses the implications of the counterintuitive result.

Introduction

Studying the dynamics of information diffusion by the cascades of *reposting* of messages on social media is important for several application domains, such as political campaigns (Stieglitz and Dang-Xuan 2013), viral marketing (Kempe, Kleinberg, and Tardos 2003), information diffusion in crisis events (Olteanu, Vieweg, and Castillo 2015), and preventing the spread of rumors (Friggeri et al. 2014). On the popular social media platform, Twitter, users can disseminate tweets posted by other users to their followers via a functionality called *retweeting*. The retweeted tweets can be further disseminated by followers. Such cascades of retweets cause some tweets to be disseminated to many users. Information widely disseminated in a social network has the potential to affect public opinion, brand awareness, and product market share (Bakshy et al. 2011). However, it has been shown that only a small fraction of information reaches many people (Dow, Adamic, and Friggeri 2013). Thus, many studies have investigated factors affecting information diffusion, particularly focusing on tweet diffusion on the Twitter social media platform (Naveed et al. 2011; Suh et al. 2010; Tsugawa and Ohsaki 2017; Ferrara and Yang 2015; Stieglitz and Dang-Xuan 2013; Recuero, Araujo, and Zago 2011). These studies have found major factors affecting tweet diffusion that include the influence of the tweet publisher (Suh et al. 2010; Martin et al. 2016), URLs (Suh et al. 2010), hashtags (Suh et al. 2010; Naveed et al. 2011), and emotional intensity (Stieglitz and Dang-Xuan 2013; Ferrara and Yang 2015; Tsugawa and Ohsaki 2017) in tweets.

Existing studies suggest that having a community structure affects information diffusion in a social network (Onnela et al. 2007; Weng, Menczer, and Ahn 2013; Nematzadeh et al. 2014; Li, Lin, and Yeh 2015; Galstyan and Cohen 2007; De Meo et al. 2014). Many social networks have a community structure, in which the network is composed of highly clustered subnetworks (communities) with sparse links between them (Newman and Girvan 2004; Ferrara 2012). There is no universal definition of communities, and several definitions of communities can be found in the literature (Fortunato 2010). However, there is one widely accepted basic concept of a community: there must be more links within the community than links connecting to nodes outside the community (Fortunato 2010). This paper follows this basic concept, and communities in a social network are defined as highly clustered subnetworks in the network that have sparse links between them. In the context of information diffusion on social media, community structure is expected to have the following two effects. First, information diffusion tends to be trapped within a community; this will be called the *trapping effect* (Weng, Menczer, and Ahn 2013; Nematzadeh et al. 2014; Onnela et al. 2007; De Meo et al. 2014) (Fig. 1(a)). Information diffusion across different communities is suggested to be a rare event, and most information is diffused within a community (Weng, Menczer, and Ahn 2013; 2013; Nematzadeh et al. 2014; Onnela et al. 2007; De Meo et al. 2014). The other effect is that if information is spread across different communities, the information will be widely spread (Weng, Menczer, and Ahn 2013; Granovetter 1973; De Meo et al. 2014; Bakshy et al. 2012) (Fig. 1(b)). This is suggested by the well known theory of Granovetter called *the strength of weak ties* (Granovetter 1973). The strength of weak ties suggests that information disseminated through weak inter-



(a) Trapping effect (b) The strength of weak ties

Figure 1: Effects of community structure on information diffusion: (a) Most information is spread within a community and inter-community diffusion is a rare event. (b) If information is spread across different communities, the information will be widely spread.

community ties is expected to be popular in the future (Granovetter 1973).

However, empirical studies that analyze the effects of community structure of a social network on information diffusion on social media are still limited. One major approach for studying the effects of community structure on information diffusion is using stochastic information diffusion models (Nematzadeh et al. 2014; Galstyan and Cohen 2007; De Meo et al. 2014; Onnela et al. 2007). Existing studies have investigated the effects of community structure on information diffusion under several models, such as independent cascade and linear threshold models (De Meo et al. 2014; Nematzadeh et al. 2014). In contrast, how community structure affects real information diffusion has not been fully explored. In particular, how community structure affects the cascading diffusion of retweets has not yet been investigated. The work of (Weng, Menczer, and Ahn 2013) is the most related to this paper. They extensively investigated the causes and effects of trapping on information diffusion. They also showed that the future popularity of a hashtag on Twitter can be predicted using community structure. As we will discuss in more detail in the next section, this paper extends the work of (Weng, Menczer, and Ahn 2013) in several directions.

This paper aims to provide empirical evidence for the effects of community structure on information diffusion. In particular, we focus on the cascading diffusion of retweets on Twitter, and address the following three research questions.

(RQ1) How does the community structure trap tweet diffusion?

(RQ2) How does inter-community diffusion of a tweet affect its future popularity?

(RQ3) How is the size of influence of the inter-community diffusion on future popularity of tweets compared with other factors?

Related Work and Contributions of This Paper

The roles of the strength of social ties in human interactions have been actively studied. In the seminal work by Granovetter (1973), the strength of a social tie is defined as "a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and reciprocal services which characterize the tie". Based on this definition, Granovetter's theory of *the strength of weak ties* (Granovetter 1973) suggests that weak ties play important roles in several contexts such as the diffusion of ideas and job seeking. In the context of social media, information about social tie strength between users is expected to be useful for various applications such as privacy control and viral marketing (Gilbert and Karahalios 2009). However, the strength of social ties, defined in (Granovetter 1973), between social media users cannot be directly observed in most social media. Therefore, Gilbert and Karahalios (2009) examined the problem of predicting the social tie strength between social media users from observable log data of user activities on social media. Another study investigated the factors affecting the changes of social tie strength on Facebook (Burke and Kraut 2014). Bakshy et al. (2012) measured the strength of social ties between Facebook users from online and offline interactions, and showed that weak ties play an important role for large-scale diffusion of news on Facebook. Although this finding is closely related to our work, they focused on social tie strength defined as the frequency of communication between users, which is different to the community structure focus of this study.

Social tie strength and community structure of social networks are closely related to each other. It has been shown that people in the same community communicate with each other more frequently than with people in different communities (Palla, Barabási, and Vicsek 2007; Tsugawa and Ohsaki 2015). This characteristic is observed in several modes of communication, such as mobile phone (Palla, Barabási, and Vicsek 2007; Tsugawa and Ohsaki 2015), email (Tsugawa and Ohsaki 2015), and messaging on social networking services (Tsugawa and Ohsaki 2015). These studies suggest that social ties within a community (i.e., intra-community links) are so-called strong ties that carry many communications whereas social ties across different communities (i.e., inter-community links) are weak ties that carry few communications.

Using the observations that community structure and social tie strength are closely related to each other, De Meo et al. (2014) introduced alternative definitions of strong and weak ties. They proposed classifying links connecting nodes belonging to different communities as weak ties, and those connecting nodes in the same community as strong ties (De Meo et al. 2014). Using this definition, they investigated the roles of strong and weak ties. They showed through simulations of information diffusion on the Facebook social network that diffusion through inter-community links, which are considered as weak ties, plays an important role for wide information diffusion (De Meo et al. 2014). Our study follows this definition of the strong and weak ties, and investigates how inter-community diffusion of a tweet affects its future popularity.

The effects of community structure on diffusion phenomena have been investigated also in other studies. Major approach is using model-based simulation experiments. Onnela et al. (2007) showed through simulation experiments on a phone call social network that strong ties trap information within a community. Galstyan and Cohen (2007) and Nematzadeh et al. (2014) showed through simulations on synthetic networks that the modularity (Newman and Girvan 2004) of a network, a measure of the strength of community structure, affects information diffusion on the network. There also exist empirical results supporting the strength of weak ties. Centola (2010) investigated the spread of health behavior under controlled experiments in artificial online social networks. It was found that the behavior spread farther and faster across networks with a community structure than across random networks. Li et al. (2015) showed that nodes that act as bridges between communities play an important role for large-scale information diffusion on Twitter. While Li et al. (2015) focused on the roles of nodes, we focus on the roles of ties and communities.

Our study follows above mentioned studies, and empirically investigates the effects of community structure on information diffusion, particularly focuses on the cascading retweet diffusion on Twitter. To the best of our knowledge, the effects of inter-community diffusion of a tweet on its cascading retweet diffusion have not been empirically investigated, but Weng et al. (2013) conducted a closely related study. They empirically investigated the effects of community structure on hashtag diffusion on Twitter and found that most of the hashtags spread within a highly clustered community (*trapping effects*) whereas a small fraction of viral ones spread across many communities. Moreover, they showed that viral hashtags can be predicted using spreading patterns in an early stage of diffusion. Namely, they show that hashtags used in many communities in an early stage of their diffusion will be viral (*the strength of weak ties*). This study follows the approach in (Weng, Menczer, and Ahn 2013), and extends it in the following directions. First, we examine the generalizability of the existing findings in different contexts, which should be important for empirical studies. In particular, we examine the applicability of the findings to different types of diffusion and different types of users from Weng et al. (2013). Retweets and hashtags have different roles on Twitter. Yang et al. (2012) discuss that the primary role of a hashtag is a bookmark of contents, and it is also used for expressing a membership of a particular group. In contrast, retweets are primarily used for spreading other users' tweets, and also often used for conversations (Boyd, Golder, and Lotan 2010). Moreover, while Weng et al. (2013) studied English-speaking Twitter users, we study both of English-speaking and Japanese-speaking users. English-speaking users and Japanese-speaking users may have different behavioral characteristics (Hong, Convertino, and Chi 2011). Using such different data, we examine how the findings in (Weng, Menczer, and Ahn 2013) can be applicable to different types of diffusion among different types of users. Second, this paper quantifies the effects of inter-community diffusion and intra-community diffusion of a tweet on its future popularity, which was not directly addressed by Weng et al. (2013). Quantifying the influence of inter-community diffusion is important for understanding the dynamics of information diffusion, and constructing a predictive model of information diffusion. We compare the effects of inter-community diffusion of a tweet on future popularity with those of other factors affecting popularity of tweets such as URL and hashtag inclusions in tweets, and influence of the tweet publisher.

Contributions Our main contributions and findings are summarized as follows. (1) We provide empirical evidence for the effects of community structure on tweet diffusion. Most tweets are disseminated within a community, whereas tweets spread across different communities will be popular in the future. (2) We examine the effects of inter-community diffusion of a tweet on its future popularity comparing other factors affecting tweet popularity. Our results suggest that inter-community diffusion of a tweet has a statistically significant effect on its future popularity, yet, contrary to our expectations, the effect is quite weak. We show that community-based features have little contributions for predicting future popularity of tweets, and discuss the implications of this counterintuitive result. (3) We examine the effects of community structure on retweet diffusion for different types of users (i.e., Japanese-speaking and English-speaking users). We show that our main findings can be applicable to both types of users.

Dataset and Preliminaries

Dataset

For analyzing cascading retweet diffusion, we collected Japanese retweets. We started data collection from July, 2013, and have been collecting Japanese retweets everyday using the Twitter application programming interface (API)[1]. The collection process successfully gave a large-number of retweets, and our retweet dataset contains more than 50M retweets per a week. This dataset has been also used in our other projects (Tsugawa and Kimura 2018; Tsugawa and Kito 2017).

We also obtained the social network of a sample of Twitter users. Due to the restrictions of the Twitter API, it was impossible to obtain the social network of all Twitter users who involved in our retweet dataset. We therefore determined target Twitter users whose social relationships would be examined. For the purposes of the analyses, we extracted active users who frequently post retweets. The procedures are as follows. We first extracted Japanese retweets posted during December 11-17, 2013 from the retweet dataset explained above, which gave 52,129,804 retweets of 14,220,864 original tweets. We next counted the number of retweets of the original tweets posted during the period, and extracted original tweets having between 10 and 100 retweets. We then extracted users who retweeted 10 or more of these tweets, which gave 356,453 users. The lower threshold of retweet cascade size (10) is intended to exclude users who post

[1]We used the Search API in the Twitter REST API v1.1, and collected Japanese tweets using the query q=RT, lang=ja.

retweets only for conversational purposes, and the upper threshold (100) is intended to exclude users who only post retweets to very popular tweets. We next obtained the social network of the 356,453 users by finding the followers and followees of the 356,453 users. The data collection using the Twitter API started January 1, 2014, and it finished January 11, 2014. We then constructed a network representing the follower and followee relationships, and extracted the largest weakly connected component of the network for the following analyses. The number of nodes belonging to the largest component is 351,870. We denote the network as $G = (V, E)$, where V is the set of nodes representing the 351,870 users, and E is the set of links representing their "following" relationships. A link $(u, v) \in E$ represents the relationship that user u follows user v.

We next detected communities in the obtained network G. There are various definitions of communities and various algorithms for detecting them (Fortunato 2010). Among the various algorithms, we primarily used the Louvain algorithm (Blondel et al. 2008) since it can be applied to large-scale networks and has been widely used (e.g., (De Meo et al. 2014)). The Louvain algorithm detects communities by maximizing the modularity (Newman and Girvan 2004), which is a measure for quantifying the quality of community detection based on both of the intra-community link density and inter-community link sparsity. In the community detection, we ignored link direction following (Weng, Menczer, and Ahn 2013). Note that, for each node u, the Louvain algorithm gives a single community $c(u)$ to which node u belongs. However, we checked the robustness of the results obtained by comparing the results obtained with an overlapping-community detection algorithm, which will be shown in later section.

Finally, we extracted retweets to be analyzed from the retweet dataset. Since the social network of the target users was obtained early January, 2014, we decided to use retweets posted in the period from January 1–31, 2014. In our dataset, the number of original tweets posted by the target users during the period was 13,996,348, and the number of retweets to the original tweets were 63,449,098. Note that the 63,449,098 retweets include retweets posted by users not in the target user set V. Among the 63,449,098 retweets, 29,270,742 retweets were posted by the target users. In what follows, we mainly use the 29,270,742 retweets posted by the target users since the communities of users posting these retweets are known. All retweet data are used for obtaining the number of retweets of the original tweets in order to accurately know the popularity of the original tweets. In the following section, these tweets and retweets, together with the obtained network and communities, are used for the analyses. Several statistics of the dataset are shown in Tab. 1.

We also use a dataset of English retweets for validating the results obtained from the Japanese dataset. The explanation about the English dataset will be given in the later section.

Table 1: Statistics of the Japanese dataset

Num. of nodes	351,870
Num. of links	29,535,522
Num. of detected communities	27
Num. of original tweets	13,996,348
Num. of all retweets	63,449,098
Num. of retweets posted by the target users	29,270,742

Table 2: Symbols used in this paper

Symbol	Definition
G	The network
V	Set of nodes in G
E	Set of links in G
E_{intra}	Set of intra-community links
E_{inter}	Set of inter-community links
u	User, i.e., node in V
c	Community
t	Original tweet
$r_k(t)$	The k-th retweet of original tweet t
$u(t)$	User that posts original tweet t
$u(r_k(t))$	User that posts retweet $r_t(k)$
$c(u)$	Community to which user u belongs

Preliminaries

Here, we define the terminology and symbols used in the following analyses. An original tweet is denoted by t, and its k-th retweet is denoted by $r_t(k)$. Note that for determining k, we use all the 63,449,098 retweets in the dataset. The users $u(t)$ and $u(r_t(k))$ are those who posted the original tweet t and retweet $r_t(k)$, respectively. Community $c(u)$ is the community to which user u belongs. Retweet $r_t(k)$ is called an *intra-community retweet* if user $u(t)$ and user $u(r_t(k))$ belong to the same community (i.e., $c(u(t)) = c(u(r_t(k)))$), otherwise $r_t(k)$ is called an *inter-community retweet*. E_{intra} and E_{inter} are the sets of intra-community links and inter-community links, respectively. Specifically, $E_{\mathrm{intra}} = \{(u, v) | (u, v) \in E \wedge c(u) = c(v)\}$, and $E_{\mathrm{inter}} = \{(u, v) | (u, v) \in E \wedge c(u) \neq c(v)\}$. These symbols are summarized in Tab. 2 for convenience.

Analysis

(RQ1): How does the community structure trap tweet diffusion?

Following (Weng, Menczer, and Ahn 2013), we define the link weight $w(u, v)$ on link (u, v) as the number of retweets by user u to original tweets by user v for investigating the trapping effect. We then compare the average link weights of intra-community and inter-community links, defined as

$$\langle w_{\mathrm{intra}}(u) \rangle = \frac{1}{k_{\mathrm{intra}}(u)} \sum_{(u,v) \in E_{\mathrm{intra}}} w(u, v), \quad (1)$$

$$\langle w_{\mathrm{inter}}(u) \rangle = \frac{1}{k_{\mathrm{inter}}(u)} \sum_{(u,v) \in E_{\mathrm{inter}}} w(u, v), \quad (2)$$

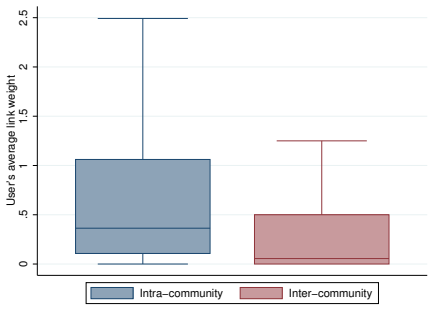
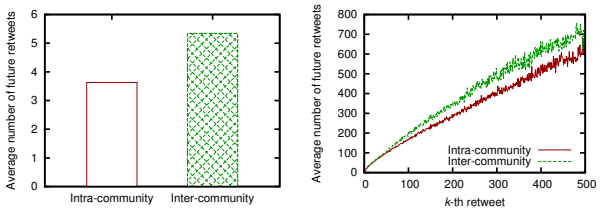


Figure 2: Box plots for comparing average link weights of intra-community and inter-community links: Users more frequently retweet the tweets posted by members of their own community.



(a) 1st retweet (b) k-th retweet

Figure 3: Comparison of average numbers of future retweets for intra-community and inter-community diffusions. Future popularity is higher when the tweets are disseminated via inter-community retweet than when the tweets are disseminated via intra-community retweet.

where $k_{\text{intra}}(u)$ and $k_{\text{inter}}(u)$ are the numbers of intra-community and inter-community links originating from user u, respectively. Normalizing the frequency of inter-community retweets and intra-community retweets by the number of inter-community links and intra-community links, we aim to eliminate the effects of the difference in the exposure rate between tweets from the same community and tweets from different communities. If users' retweeting is independent of the community of origin of a tweet, there would be no difference between $\langle w_{\text{intra}}(u)\rangle$ and $\langle w_{\text{inter}}(u)\rangle$. Our expectation is that users prefer to retweet tweets from their own community, and as a result, $\langle w_{\text{intra}}(u)\rangle$ will be higher than $\langle w_{\text{inter}}(u)\rangle$ due to the effects of homophily (McPherson, Smith-Lovin, and Cook 2001). Figure 2 shows box plots of $\langle w_{\text{intra}}(u)\rangle$ and $\langle w_{\text{inter}}(u)\rangle$ for the 309,470 users who posted retweets (users who did not post any retweets are excluded).

Figure 2 shows that average link weights on intra-community links are generally higher than average link weights on inter-community links. The median of $\langle w_{\text{intra}}(u)\rangle$ and $\langle w_{\text{inter}}(u)\rangle$ is 0.36 and 0.056, respectively. The mean of $\langle w_{\text{intra}}(u)\rangle$ and $\langle w_{\text{inter}}(u)\rangle$ is 1.32 and 0.71, respectively. A Mann–Whitney U test shows that there is a significant difference between $\langle w_{\text{intra}}(u)\rangle$ and $\langle w_{\text{inter}}(u)\rangle$ ($n = 309,470$, $U = 66,756,864,850$, $p < 0.01$). This result suggests that a intra-community link carries more retweets than a inter-community link. This result is consistent with the results in (Weng, Menczer, and Ahn 2013), which suggests that the relation between community structure and social tie strength in terms of retweet frequency is a universal characteristic of social media.

Summary of findings Community structure traps the cascading diffusion of retweets: intra-community link carries more retweets than an inter-community link.

(RQ2) How does inter-community diffusion of a tweet affect its future popularity?

Results in the previous section show that inter-community links carry much fewer retweets than intra-community links, which implies that inter-community links are weak ties in terms of retweet frequency. Granovetter's theory of the strength of weak ties suggests that weak ties convey information that have potential to be disseminated to many people (Granovetter 1973). Therefore, a tweet spread across different communities should be more popular in the future than a tweet spread within a community.

We compare the future popularity of tweets spread across different communities and that of tweets spread within a community. For each retweet $r_k(t)$ posted by the target users, we investigate the number of future retweets after $r_k(t)$ is posted, which is defined as $N(t) - k$, where $N(t)$ is the total number of retweets to the original tweet t. We then calculate the average of future retweets when $r_k(t)$ is an intra-community retweet and when $r_k(t)$ is an inter-community retweet. Figure 3(a) shows a comparison of future popularity of tweets after their first retweet ($k = 1$) for intra-community and inter-community retweets. We calculated the average and standard error of the number of future retweets. Since the standard errors are sufficiently small (0.04 for inter-community retweet and 0.01 for intra-community retweet), error bars are not shown in Fig. 3(a). The future popularity of tweets disseminated via inter-community retweet is approximately 1.5-fold the future popularity of tweets disseminated via intra-community retweet. The Welch's t-test shows that the difference is statistically significant ($p < 0.01$). We also investigated the future popularity of tweets disseminated via intra-community retweet and those disseminated via inter-community retweet for $k \geq 1$ (see Fig. 3(b)). This result also suggests that inter-community diffusion is correlated with the increases in future popularity of tweets.

These results support the idea that inter-community diffusion of a tweet is significantly correlated with increase in its future popularity. On average, a tweet spread across different communities has a future popularity 1.5-fold that of a tweet spread within the same community.

Summary of findings Inter-community diffusion of a tweet is significantly correlated with increase in its future popularity. A tweet spread across different communities has

a future popularity approximately 1.5-fold that of a tweet spread within the same community.

(RQ3): How is the size of influence of the inter-community diffusion on future popularity of tweets compared with other factors?

The previous subsection shows that inter-community diffusion affects future popularity. However, several factors have been reported as affecting tweet popularity, such as the presence of a URL and the number of followers of the tweet's publisher (Suh et al. 2010). We, therefore, examine the significance of the effects of inter-community diffusion on future popularity by taking into account of the effects of other factors. For instance, influential nodes may have many inter-community links, and therefore, tweets posted by influential nodes are expected to receive many inter-community retweets. Moreover, tweets posted by influential nodes are expected to be popular. Therefore, there is a concern that the tweets spread via inter-community retweet have higher future popularity only because such tweets are posted by highly influential nodes. This subsection aims to address such concerns regarding other factors that may related to community structure and tweet popularity. Moreover, we investigate the size of the influence of inter-community diffusion on the future popularity of tweets compared with that of other factors.

Regression analysis For examining the significance of the effects of inter-community diffusion on future popularity by taking into account of the effects of other factors, a regression model is constructed. For each retweet $r_t(k)$ posted by the target users, the dependent variable is the number of retweets of tweet t after $r_t(k)$ is posted (*RT-num*). The candidates of the independent variables are summarized in Tab. 3. The main concern here is the influence of the variable *inter-com*, which is a categorical variable representing whether $r_t(k)$ is inter-community retweet or not. Other variables include factors that are shown to affect popularity of tweets (Suh et al. 2010; Naveed et al. 2011; Bakshy et al. 2012; Martin et al. 2016) such as the number of followers of $u(t)$ and the presence of a URL in tweet t. Variable *influence* is defined as the number of retweets to the tweets posted by user $u(t)$ in the previous month (i.e., during December, 2013). This variable is shown to strongly affect the popularity of tweets (Martin et al. 2016; Bakshy et al. 2012). Since analyzing all tweets and retweets in our dataset is computationally expensive, we randomly selected two million original tweets from all 13,996,348 original tweets in the dataset and we used their all retweets posted by the target users for the regression analysis. Since the popularity of tweets has a long-tailed distribution, we log-transformed the dependent variable, and the Ordinary Least Squares (OLS) linear regression model is used. We checked the normality assumption with the Kolmogorov-Smirnov test. The results of the test applied to the dependent variable and residuals suggest the null hypothesis of normal distribution cannot be rejected. The results of regression analysis are shown in Tab. 4. Note that to avoid multicollinearity, deg_{orig}, deg_{rt}, and *intra-com-rate* are not included in the model. deg_{orig}, deg_{rt}, and *intra-com-rate* have strong correlation (i.e., correlation coefficient more than 0.7) with *influence*, *avedeg*, and *inter-com*, respectively, and therefore the constructed model was not interpretable when including these variables.

Table 4 shows that considering other factors, inter-community diffusion of a tweet has a significant influence on its future popularity. The size of the influence of each variable is estimated from e^{β}, where β is the regression coefficient in the model. From the value of e^{β} for *inter-com*, we can see that inter-community diffusion of a tweet has the effect of increasing its future popularity by a factor of approximately 1.3, which is consistent with the results in the previous subsection. From this result, we can confirm that eliminating the effects of other factors such as node influence and degree of users who retweet the tweets, inter-community diffusion has a significant positive effect on future popularity. Note that the regression coefficients of different variables cannot be directly compared with each other since some of them are quantitative and some of them are binary variables. We should also note that the value of R^2 is not high in our model, but it is comparable to those in existing studies (e.g., (Martin et al. 2016)). As discussed in (Martin et al. 2016), the cascade size prediction is a difficult task, which results in relatively low R^2 value.

Prediction To further investigate the effects of inter-community diffusion on the future popularity of tweets, we conducted an experiment to predict future popularity of tweets using features obtained from community structure. The task is predicting whether the number of retweets of tweet t will be over θ or not at the time when retweet $r_t(k)$ is posted ($k < \theta$). θ and k will be called as the virality threshold, and the training period, respectively. The aim here is to find viral tweets at an early stage of their diffusion. As the virality threshold θ, we used 1%-tile, and 0.5%-tile values of the cascade size of the tweets in the dataset. As the training period k, we used $k = \lfloor 0.05\theta \rfloor, \lfloor 0.1\theta \rfloor, \lfloor 0.2\theta \rfloor, \lfloor 0.4\theta \rfloor$.

To construct classifiers predicting the popularity of tweets, we used random forests (Breiman 2001). The training data are the retweets used in the regression analysis. As the features of the classifiers, we used the variables shown in Tab. 3. For each of k, we constructed models using several combinations of features: a model only using features related to tweet contents (contents), that using features related to tweet contents and community (w/ community), that using features related to tweet contents and node degree (w/ degree), that using features related to tweet contents and node influence (w/ influence), that using all features (full), and that without using features related to community (w/o community). The number of decision trees was 500, and each decision tree was trained with randomly selected $\lfloor \sqrt{f} \rfloor$ features, where f is the number of features used in the model.

We investigated the prediction accuracy of the constructed models. One million original tweets randomly extracted from the dataset and all their retweets posted by the target users were used as the test data. Note that the test data did not include any tweets in training data. Using the test data, we calculated the precision, recall, and F_1-measure (Rijsber-

Table 3: Candidates of independent variables used in the regression analysis

Variables related to community	
inter-com	Categorical variable indicating whether $r_t(k)$ is inter-community retweet or not
intra-com-rate	Fraction of intra-community retweets among the first k retweets
$comsize_{orig}$	Size of community $c(u(t))$
$comsize_{rt}$	Size of community $c(r_t(k))$
Variables related to node degree	
deg_{orig}	Number of followers of $u(t)$
deg_{rt}	Number of followers of $u(r_t(k))$
avedeg	Average of number of followers of users in $U_{k,t}$
Variable related to node influence	
influence	Average number of retweets of $u(t)$'s tweets in the previous month (i.e., Dec., 2013)
Variables related to tweet contents	
k	Retweet count k
URL	Categorical variable indicating whether tweet t contains a URL or not
hash	Categorical variable indicating whether tweet t contains any hashtag or not
word	Number of words in tweet t

Table 4: Results of regression analysis for investigating the effect of inter-community retweets on the future popularity of a tweet. Inter-community diffusion of a tweet significantly affects its future popularity. (**: $p < 0.01$)

Dependent variable: log(*RTnum*)		
Independent variables	Coeff. β	e^{β}
*k***	1.488e-03	1.002
*inter-com***	2.870e-01	1.332
$comsize_{orig}$**	-1.408e-06	1.000
$comsize_{rt}$**	4.818e-06	1.000
log(*avedeg*)**	8.858e-02	1.093
log(*influence*)**	5.674e-01	1.764
*URL***	8.852e-01	2.423
*hash***	3.485e-01	1.417
*word***	1.496e-02	1.015
Num. of observations		3,585,062
R^2		0.3676

gen 1979) of the constructed models as the measures of prediction accuracy (Tab. 5). Prediction accuracies of random guess (dented as random) are also included in the table. The random guess selects n_p random tweets as viral, where n_p is the actual number of tweets exceeding the virality threshold (Weng, Menczer, and Ahn 2013).

Contrary to our expectations, Tab. 5 indicates that the community-based features have little contributions to improve the accuracy for predicting the future popularity of tweets. The differences of accuracies between the full model and the w/o community model are small. Moreover, comparing the accuracies of the w/ community model and those of the w/ degree and the w/ influence models, it is also suggested that the contribution of community related features is smaller than features related to node degree and influence. From these observations, we conclude that community-related features have little contributions for improving the prediction accuracy, but other features such as node influence and node degree have more predictive power. Moreover, although the task here is difficult (see accuracies of random), it also should be noted that the prediction accuracies of constructed model is not very high in this experiment. To achieve higher prediction accuracy, the use of additional features and other prediction models should be considered.

Summary of findings Although the effect of inter-community diffusion of a tweet is suggested to be statistically significant, the effect is quite weak. Other features such as node influence and node degree have larger influence on future popularity than inter-community diffusion.

Robustness of results

Validation with overlapping community detection

Methodology To check the robustness of the results and findings in the previous section we used a different community detection algorithm. The Louvain algorithm used in the previous section is a disjoint community detection algorithm that assumes each node belongs to a single community. However, overlapping community detection algorithms that assume each node belongs to multiple communities also exist (Fortunato 2010). In this section we show that similar results can be obtained when using overlapping community detection.

We used the popular overlapping community detection algorithm, called the link clustering algorithm, proposed in (Ahn, Bagrow, and Lehmann 2010) with a similarity threshold value of $t = 0.2$. Table 6 shows several statistics of the obtained communities. Some nodes do not belong to any communities. We excluded such nodes from the following analyses. We also excluded two-nodes communities.

Table 5: Comparison of prediction accuracies for each model when predicting the future popularity of tweets. Each model predicts whether a tweet will be retweeted θ times or more when its k-th retweet is posted.

Predicting top 1% tweets ($\theta = 58$)

	Precision				Recall				F_1-measure			
training period	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ
random	0.05	0.10	0.19	0.37	0.05	0.10	0.19	0.37	0.05	0.10	0.19	0.37
contents	0.00	0.00	0.78	0.60	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.38
w/ community	0.00	0.67	0.65	0.63	0.00	0.00	0.01	0.05	0.00	0.00	0.03	0.09
w/ degree	0.68	0.71	0.71	0.66	0.05	0.07	0.10	0.20	0.10	0.13	0.18	0.31
w/ influence	0.63	0.67	0.69	0.56	0.07	0.08	0.11	0.28	0.13	0.15	0.19	0.38
w/o community	0.64	0.67	0.68	0.65	0.11	0.11	0.15	0.27	0.18	0.19	0.24	0.38
full	0.62	0.66	0.67	0.61	0.11	0.12	0.17	0.33	0.19	0.20	0.27	0.43

Predicting top 0.5% tweets ($\theta = 98$)

	Precision				Recall				F_1-measure			
training period	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ
random	0.04	0.08	0.16	0.33	0.04	0.08	0.16	0.33	0.04	0.08	0.16	0.33
contents	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.05
w/ community	0.00	0.50	0.70	0.60	0.00	0.00	0.01	0.04	0.00	0.01	0.02	0.08
w/ degree	0.68	0.72	0.67	0.62	0.04	0.05	0.08	0.15	0.08	0.10	0.14	0.25
w/ influence	0.66	0.64	0.69	0.65	0.04	0.06	0.09	0.18	0.08	0.11	0.16	0.28
w/o community	0.64	0.63	0.68	0.63	0.09	0.11	0.14	0.23	0.16	0.19	0.23	0.34
full	0.64	0.64	0.66	0.61	0.09	0.11	0.13	0.26	0.16	0.18	0.22	0.37

Table 6: Statistics of overlapping communities obtained with link clustering ($t = 0.2$)

Num. of communities	588,929
Ave. size of communities	6.57
Ave. num. of communities to which users belong	13.98
Num. of users belonging to at least one community	276,769

We defined the overlap of communities for users u and v as

$$O(u,v) = \frac{|C(u) \cap C(v)|}{|C(u) \cup C(v)|}, \tag{3}$$

where $C(u)$ is the set of communities to which node u belongs. We use this measure in the following analyses. Extending the concept of weak and strong ties in disjoint communities (De Meo et al. 2014) to overlapping communities, we consider links with low community overlap as weak ties and links with high community overlap as strong ties.

Results We first check how overlapping community structure affects the retweet frequency between users. Figure 4 shows the relation between the community overlap between two nodes and average number of retweets occurring between those nodes. As the community overlap between two nodes increases the number of retweets occurring between those nodes also increases. Thus, a high community overlap between two nodes indicates a strong relationship between them. This suggests that strong ties with high community overlap convey more information than weak ties with low community overlap.

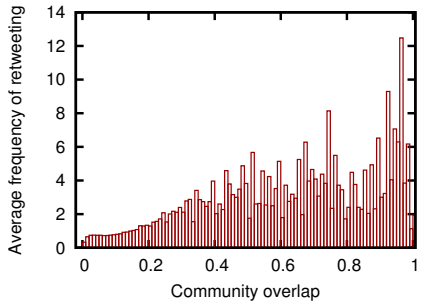


Figure 4: Relation between community overlap between two nodes and the average number of retweets occurring between those nodes: As the community overlap between two nodes increases the number of retweets occurring between those nodes also increases.

We next investigated how tweet diffusion through a low-community-overlap link, which we regard as a weak tie, affects its future popularity. A regression model was constructed with the same dependent variable used for the analysis in the previous section. For the independent variables, we used community overlap $O(u,v)$ (*overlap*) between two nodes instead of *inter-com*, while $comsize_{orig}$ and $comsize_{rt}$ were replaced with the average sizes of the communities to which user $u(t)$ (*ave-comsize*$_{orig}$) and user $u(r_t(k))$ belong (*ave-comsize*$_{rt}$). Linear regression was used. Table 7 shows the results. It can be seen that *overlap* significantly affects the popularity of tweets. This suggests that a tweet spread through a low-community-overlap link will be spread widely.

Table 7: Results of a regression analysis for investigating the effect of community overlap on future popularity of a tweet. Diffusion of a tweet between low-community-overlap users significantly affects its future popularity. (**: $p < 0.01$)

Dependent variable: log(*RTnum*)		
Independent variables	Coeff. β	e^{β}
*k***	1.514e-03	1.002
*overlap***	-2.307e+00	0.100
ave-comsize$_{orig}$**	-4.985e-04	1.000
ave-comsize$_{rt}$**	-1.907e-04	1.000
log(*avedeg*)**	8.998e-02	1.094
log(*influence*)**	5.850e-01	1.795
*URL***	8.170e-01	2.264
*hash***	2.956e-01	1.344
*word***	1.378e-02	1.014
Num. of observations		2,955,521
R^2		0.3719

We also revisited the experiment of future popularity prediction (Tab. 8). The settings are same with those in the previous section, but community related features are replaced with overlapping-community-based ones. We used *overlap*, *ave-comsize*$_{orig}$, and *ave-comsize*$_{rt}$ instead of *inter-com*, *comsize*$_{orig}$, and *comsize*$_{rt}$, respectively. We also used $\frac{\sum_{u \in U_{k,t}} O(u,u(t))}{|U_{k,t}|}$ instead of *intra-com-rate*. Due to space limitations, we only show the results for $\theta = 98$. Table 8 indicates that overlapping-community-based features have little influence on future popularity prediction, which is consistent with the results in the previous section.

Overall, the results discussed in this section suggest that findings in the previous section can be replicated with overlapping communities. Since similar results are obtained from such different definitions of community, we expect that the effects of community structure on information diffusion shown in this paper will observed for several definitions of community.

Validation with different user set

Methodology For further checking the robustness of the results and findings, we used a dataset of a different user set. We collected English tweets, their retweets, and the social network of a sample of users who involved in those tweets and retweets. We collected English retweets during November 19–25, 2018, which gave 208,673,006 retweets of 25,813,058 original tweets. From the original tweets, we randomly extracted 10,000 tweets. We collected followers and followees of the users who posted the 10,000 original tweets, and users who posted retweets to the 10,000 original tweets in early December 2018. We then constructed a social network of these users, and extracted the largest weakly connected component of the network. The number of nodes belonging to the largest component is 110,566, and these users are the target users used in the following analyses. The communities of the target users were determined using the Louvain algorithm (Blondel et al. 2008). From the collected tweets, we extracted original tweets posted by the target users. Several statistics of the English dataset are shown in Tab. 9.

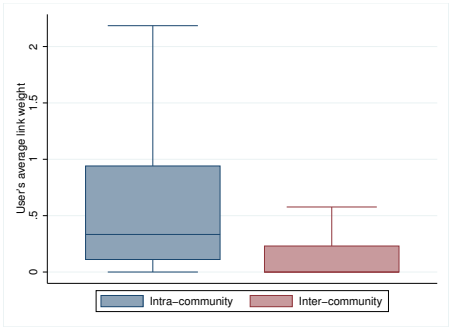


Figure 5: Box plots for comparing average link weights of intra-community and inter-community links of users in the English dataset: This confirms the results in Fig. 2.

Results We first check the effects of community structure on the retweet frequency between users. Figure 5 shows box plots of $\langle w_{\mathrm{intra}}(u) \rangle$ and $\langle w_{\mathrm{inter}}(u) \rangle$ for the users in the English dataset. This result confirms that an intra-community link carries more retweets than an inter-community link.

We next check the effects of inter-community diffusion of a tweet on its future popularity. A regression model was constructed with the same dependent and independent variables used for the analysis in the previous section (Tab. 10). The variable *influence* was calculated using retweets posted during the period from October 20, 2018 to November 18, 2018. This result shows that the variable *inter-com* has a statistically significant positive effect on future popularity. Although the regression coefficient of each variable and R^2 value of the model are different from the result of the Japanese dataset, this result supports the finding that inter-community diffusion of a tweet is significantly correlated with increase in its future popularity.

We finally revisited the experiment of future popularity prediction using the English dataset. We used randomly selected 80% original tweets and their retweets in the dataset as the training data, and the rest of the data were used as the test data. Table 11 shows the prediction accuracies for each model. These results show that also for the English dataset, community-related features have little contributions for improving the prediction accuracy. Other features such as influence and degree have larger impact on prediction accuracy than community-related features. These are consistent with the results for the Japanese dataset, suggesting that our results are robust for different types of users.

Discussion

Implications

Our key finding in this paper is that the community-related features only have very weak influence on future popularity of tweets, which is contrary to our expectations and the results in the existing studies (Weng, Menczer, and Ahn 2013;

Table 8: Comparison of prediction accuracies for each model when predicting the future popularity of tweets using features related to overlapping communities. Each model predicts whether a tweet will be retweeted θ times or more when its k-th retweet is posted.

Predicting top 0.5% tweets ($\theta = 98$)

	Precision				Recall				F_1-measure			
training period	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ
random	0.04	0.08	0.16	0.33	0.04	0.08	0.16	0.33	0.04	0.08	0.16	0.33
contents	0.00	0.00	0.00	0.72	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04
w/ community	0.62	0.65	0.68	0.58	0.01	0.03	0.03	0.10	0.03	0.05	0.06	0.17
w/ degree	0.67	0.68	0.65	0.61	0.04	0.06	0.08	0.15	0.08	0.11	0.14	0.25
w/ influence	0.65	0.63	0.70	0.66	0.04	0.06	0.09	0.16	0.08	0.12	0.16	0.26
w/o community	0.65	0.62	0.68	0.62	0.09	0.11	0.14	0.22	0.15	0.19	0.23	0.33
full	0.60	0.62	0.66	0.62	0.09	0.11	0.13	0.25	0.15	0.19	0.22	0.35

Table 9: Statistics of the English dataset

Num. of nodes	110,566
Num. of links	3,129,091
Num. of detected communities	39
Num. of original tweets	1,014,480
Num. of all retweets	14,143,862
Num. of retweets posted by the target users	1,991,484

Nematzadeh et al. 2014; Li, Lin, and Yeh 2015; Galstyan and Cohen 2007; De Meo et al. 2014). This finding is supported by the results from different types of definitions of communities (i.e., overlapping community and disjoint community) and from different types of users (i.e., Japanese-speaking and English-speaking users). This suggests that for the cascade size prediction problem (Cheng et al. 2014; Martin et al. 2016), community-related features are not effective, and other features such as node influence and degree should be used.

Our findings are expected to be useful for realistic simulations of information diffusion on social networks. Information diffusion models have been used, for instance, for finding influential nodes in social networks (Lü et al. 2016). The cascade sizes of information diffusion triggered by identified nodes are considered as representing their influence, and the effectiveness of the algorithms for finding influential nodes is also based on cascade size. Typically, in the simulation of information diffusion, probabilities of information diffusion between nodes are determined without considering the community structure, and a fixed information diffusion probability is assumed for every pair of nodes (Lü et al. 2016). In contrast, our results show that the probabilities of inter-community diffusion and intra-community diffusion are different. If information diffusion probabilities between nodes are determined based on this finding (i.e., the probability of intra-community diffusion is higher than that of inter-community diffusion), more realistic information diffusion can be generated in a simulation, which may change the understanding of the effectiveness of algorithms for finding influential nodes.

Table 10: Results of regression analysis for investigating the effect of inter-community retweets on the future popularity of a tweet in the English dataset. This confirms the result in Tab. 4. (**: $p < 0.01$)

Dependent variable: $\log$(*RTnum*)		
Independent variables	Coeff. β	e^{β}
*k***	2.531e-04	1.000
*inter-com***	3.206e-01	1.378
$comsize_{orig}$**	-3.539e-06	1.000
$comsize_{rt}$**	2.359e-05	1.000
$\log$(*avedeg*)**	5.276e-02	1.054
$\log$(*influence*)**	9.041e-01	2.470
*URL***	1.439e-01	1.155
*hash***	-2.321e-01	0.793
*word***	3.342e-02	1.034
Num. of observations		1,991,484
R^2		0.5496

Limitations

Although this paper investigates the effects of community structure on information diffusion, the opposite effects are still not understood. Information spreading has been shown to affect network structure (Weng et al. 2013; Hutto, Yardi, and Gilbert 2013), which may change the community structure. To study the mutual interactions between community structure and information diffusion is an interesting direction of future research.

Further investigation on the differences between the results of the Japanese dataset and those of English dataset will be necessary. Although we obtained similar results from the both datasets, there are some differences in the results. For instance, R^2 in the English dataset is larger than that in the Japanese dataset. The cause of such differences is still unclear in this study. Moreover, we should note that there are 5-years difference between the data collection periods of the two datasets. The difference of the data collection periods might affect our results.

Table 11: Comparison of prediction accuracies for each model when predicting the future popularity of tweets in the English dataset. Each model predicts whether a tweet will be retweeted θ times or more when its k-th retweet is posted.

Predicting top 0.5% tweets ($\theta = 336$)

	Precision				Recall				F_1-measure			
training period	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ	0.05θ	0.1θ	0.2θ	0.4θ
random	0.04	0.08	0.20	0.35	0.04	0.08	0.20	0.35	0.04	0.08	0.20	0.35
contents	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
w/ community	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.39
w/ degree	0.00	0.40	0.58	0.49	0.00	0.02	0.08	0.36	0.00	0.03	0.14	0.41
w/ influence	0.44	0.41	0.55	0.55	0.06	0.11	0.18	0.47	0.11	0.18	0.27	0.51
w/o community	0.65	0.53	0.59	0.59	0.11	0.15	0.20	0.45	0.19	0.24	0.30	0.51
full	0.64	0.48	0.58	0.58	0.14	0.19	0.19	0.52	0.23	0.27	0.29	0.55

For clarifying the reason of the difference in the effects of community structure on retweet diffusion and hashtag diffusion, more efforts are still needed. For the hashtag diffusion, it is shown that community-related features have large influence on future hashtag popularity (Weng, Menczer, and Ahn 2013). In contrast, our results only show the weak effects of community-related features on the cascade sizes of retweets. To investigate the cause of such difference is an important future work.

Moreover, how the characteristics of *each* community affect information diffusion is also still an open issue. Although this paper shows the size of community has little effect on information diffusion, the density of the community or the strength of community structure may affect both intra and inter community diffusion. Measures for quantifying the characteristics of an individual community have been proposed (Leskovec, Lang, and Mahoney 2010) and these should be useful for future analyses of the relation between community characteristics and information diffusion within the community. Moreover, using other types of community detection such as hierarchical community detection and community detection considering link directions (Fortunato 2010) is also interesting.

Finally, there still exist other factors that may related to community structure and information diffusion. For instance, a node with high betweenness centrality (Freeman 1979) tends to have high influence that can spared tweets to many other users (Lü et al. 2016). Moreover, such node is expected to posts many inter-community retweets since a high-betweenness node tends to be a bridge of multiple communities (Newman and Girvan 2004). To distinguish the effects of community structure on retweet diffusion and the effects of node role on the retweet diffusion, there remains a room for improvement in the design of analyses.

Conclusion

In this paper we have investigated how the community structure of a social network of Twitter users affects the cascading diffusion of retweets. The results have shown that the frequency of intra-community retweets by a user is approximately double the frequency of inter-community retweets. This suggests that cascading diffusions of retweets are typically trapped within a community, and inter-community diffusion is a rare event. In contrast, the results have also shown that tweets disseminated via inter-community retweets have higher future popularity approximately 1.5-fold that of tweets disseminated via intra-community retweets. By using this fact, we constructed classifiers to predict the future popularity of tweets from community-based features as well as other features affecting future popularity of tweets. Our results have shown that, contrary to our expectations and results in the existing studies (Weng, Menczer, and Ahn 2013; Nematzadeh et al. 2014; Li, Lin, and Yeh 2015; Galstyan and Cohen 2007; De Meo et al. 2014), community-based features have little contributions for predicting the future popularity of tweets. Moreover, these findings are obtained from both of the English and Japanese datasets and are robust against changes in the definition of community. Overall, this paper provides empirical evidence for effects of community structure on information diffusion that have long been believed but rarely empirically validated.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 17H01733.

References

Ahn, Y.-Y.; Bagrow, J. P.; and Lehmann, S. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466(7307):761–764.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proc. WSDM'11*, 65–74.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Proc. WWW'12*, 519–528.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Proc. HICSS'10*, 1–10.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Burke, M., and Kraut, R. E. 2014. Growing closer on Facebook: changes in tie strength through social network site use. In *Proc. CHI'14*, 4187–4196.

Centola, D. 2010. The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proc. WWW'14*, 925–936.

De Meo, P.; Ferrara, E.; Fiumara, G.; and Provetti, A. 2014. On Facebook, most ties are weak. *Communications of the ACM* 57(11):78–84.

Dow, P. A.; Adamic, L. A.; and Friggeri, A. 2013. The anatomy of large Facebook cascades. In *Proc. ICWSM'13*, 145–154.

Ferrara, E., and Yang, Z. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1:e26.

Ferrara, E. 2012. A large-scale community structure analysis in Facebook. *EPJ Data Science* 1(1).

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486(3):75–174.

Freeman, L. 1979. Centrality in social networks conceptual clarification. *Social Networks* 1(3):215–239.

Friggeri, A.; Adamic, L. A.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *Proc. ICWSM'14*, 101–110.

Galstyan, A., and Cohen, P. 2007. Cascading dynamics in modular networks. *Phys. Rev. E* 75(3):036109.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proc. CHI'09*, 211–220.

Granovetter, M. S. 1973. The strength of weak ties. *American Journal of Sociology* 78(6):1360–1380.

Hong, L.; Convertino, G.; and Chi, E. H. 2011. Language matters in Twitter: A large scale study. In *Proc. ICWSM'11*, 518–521.

Hutto, C. J.; Yardi, S.; and Gilbert, E. 2013. A longitudinal study of follow predictors on Twitter. In *Proc. CHI'13*, 821–830.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *Proc. KDD'03*, 137–146.

Leskovec, J.; Lang, K. J.; and Mahoney, M. 2010. Empirical comparison of algorithms for network community detection. In *Proc. WWW'10*, 631–640.

Li, C.-T.; Lin, Y.-J.; and Yeh, M.-Y. 2015. The roles of network communities in social information diffusion. In *Proc. IEEE Big Data'15*, 391–400.

Lü, L.; Chen, D.; Ren, X.-L.; Zhang, Q.-M.; Zhang, Y.-C.; and Zhou, T. 2016. Vital nodes identification in complex networks. *Physics Reports* 650:1–63.

Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A.; and Watts, D. J. 2016. Exploring limits to prediction in complex social systems. In *Proc. WWW'16*, 683–694.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Naveed, N.; Gottron, T.; Kunegis, J.; and Alhadi, A. C. 2011. Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proc. WebSci'11*, 1–7.

Nematzadeh, A.; Ferrara, E.; Flammini, A.; and Ahn, Y.-Y. 2014. Optimal network modularity for information diffusion. *Phys. Rev. Letters* 113(8):088701.

Newman, M. E. J., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69(2):026113.

Olteanu, A.; Vieweg, S.; and Castillo, C. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proc. CSCW'15*, 994–1009.

Onnela, J.-P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaski, K.; Kertész, J.; and Barabási, A.-L. 2007. Structure and tie strengths in mobile communication networks. *PNAS* 104(18):7332–7336.

Palla, G.; Barabási, A.-L.; and Vicsek, T. 2007. Quantifying social group evolution. *Nature* 446(7136):664–667.

Recuero, R.; Araujo, R.; and Zago, G. 2011. How does social capital affect retweets? In *Proc. ICWSM'11*, 305–312.

Rijsbergen, C. J. V. 1979. *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 2nd edition.

Stieglitz, S., and Dang-Xuan, L. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* 29(4):217–247.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proc. SocialCom'10*, 177–184.

Tsugawa, S., and Kimura, K. 2018. Identifying influencers from sampled social networks. *Physica A: Statistical Mechanics and its Applications* 507:294–303.

Tsugawa, S., and Kito, K. 2017. Retweets as a predictor of relationships among users on social media. *PloS One* 12(1):e0170279.

Tsugawa, S., and Ohsaki, H. 2015. Community structure and interaction locality in social networks. *Journal of Information Processing* 23(4):402–410.

Tsugawa, S., and Ohsaki, H. 2017. On the relation between message sentiment and its virality on social media. *Social Network Analysis and Mining* 7(1):19.

Weng, L.; Ratkiewicz, J.; Perra, N.; Gonçalves, B.; Castillo, C.; Bonchi, F.; Schifanella, R.; Menczer, F.; and Flammini, A. 2013. The role of information diffusion in the evolution of social networks. In *Proc. KDD'13*, 356–364.

Weng, L.; Menczer, F.; and Ahn, Y.-Y. 2013. Virality prediction and community structure in social networks. *Scientific Reports* 3:2522.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We know what@ you# tag: does the dual role affect hashtag adoption? In *Proc. WWW'12*, 261–270.