Topicality and Impact in Social Media: Diverse Messages, Focused Messengers

Lilian Weng*, Filippo Menczer

Center for Complex Networks and Systems Research School of Informatics and Computing, Indiana University Bloomington, Bloomington, USA

* weng@umail.iu.edu

Abstract

We have a limited understanding of the factors that make people influential and topics popular in social media. Are users who comment on a variety of matters more likely to achieve high influence than those who stay focused? Do general subjects tend to be more popular than specific ones? Questions like these demand a way to detect the topics hidden behind messages associated with an individual or a keyword, and a gauge of similarity among these topics. Here we develop such an approach to identify clusters of similar hashtags in Twitter by detecting communities in the hashtag co-occurrence network. Then the topical diversity of a user's interests is quantified by the entropy of her hashtags across different topic clusters. A similar measure is applied to hashtags, based on co-occurring tags. We find that high topical diversity of early adopters or co-occurring tags implies high future popularity of hashtags. In contrast, low diversity helps an individual accumulate social influence. In short, diverse messages and focused messengers are more likely to gain impact.

Introduction

Online social media provide platforms on the Internet in which people can easily and cheaply exchange messages. A great body of information is generated and tracked digitally, creating unprecedented opportunities for studying user generated content and information diffusion processes [1, 2]. Despite a growing body of research around content and content creators, our understanding of the factors that make messages popular and messengers influential is still quite limited.

Identifying which strategies of sharing and spreading content online are more likely to obtain popularity or social impact is a Holy Grail in social media research. Here, in particular, we focus on the role of topical diversity of messages and messengers: Are users who comment on a variety of matters more or less likely to achieve high influence compared to those who delve into one focused field? Do general Twitter hashtags, such as #lol, tend to be more or less popular than novel ones, such as #instantlyinlove? We need to identify topical interests and furthermore to measure topical diversity of messages and messengers. By quantifying topical diversity, we can acquire the capability to distinguish messengers with focused expertise from those with broad interests, and to tell broadly popular messages apart from novel ones. This

problem involves several prerequisite tasks to be solved. First, we should be able to extract topics out of online content and identify user interests, while the connections between content and users and social links among users should be considered. Most existing methods for detecting topics and user interests delve into the content of the messages without leveraging knowledge of the social network topology. Second, topics should be characterized in such a way that one can naturally measure their diversity. The diversity of user interests has not yet been thoroughly investigated in the literature. Eventually, we expect to apply the extracted topics and topical diversity to some practical applications like recommendation and virality/traffic prediction.

In this paper, we study the topical diversity of messages and messengers. We develop a method to detect topics using the social network structure, but not the content, and then propose a way to distinguish messengers with diverse interests from those with focused attention, as well as messages on general matters from those on particular domains. Thus we are able to tell which categories of users and messages have better chances to gain influence and impact.

We use data from Twitter, one of the most popular social media platforms where Internet *memes* are generated and propagated. A user *follows* others to subscribe to the information they share. This generates a directed social network structure along which messages spread. Each Twitter user can post short messages called *tweets*, which may contain explicit topical tags, words, or phrases following a hash symbol ('#'), named *hashtags*. By using hashtags, people explicitly declare their interests in corresponding discussions and help others with similar preferences find appealing content. In this study we focus on hashtags as operational proxies for memes.

Messages in social media involve a variety of *topics*. Content, messages, or ideas are deemed semantically similar if they discuss, comment, or debate about the same topic; conversely, we can detect a topic by clustering a group of related messages observed. Several studies examined the recognition of topics in the online scenario and social media [3–7]. Leskovec *et al.* grouped short, distinctive phrases by single-rooted directed acyclic graphs used as signatures for different topics [3]. Features extracted from content, metadata, network, and their combinations were leveraged to detect events in social streams [7, 8]. Another approach is based on the discovery of dense clusters in the inferred graph of correlated keywords, extracted from messages in a given time frame [6, 9]. Here we adopt a similar strategy to identify clusters of similar hashtags by detecting communities in the network topology [10, 11] on account of topic locality. Topic locality in the Web describes the phenomenon of Web pages tending to link with related content [12, 13]. The effect of topic locality is utilized in focused Web crawlers [14], collaborative filtering [15, 16], interest discovery in social tagging [17, 18], and many other applications [9, 19–22]. In our scenario, topic locality refers to the assumption that semantically similar hashtags are more likely to be mentioned in the same messages and therefore to be close to each other in a hashtag co-occurrence network.

Meanwhile, we see a growing literature on discovering user interests and topics [21–27]. A common approach is to use a vector representation generated from all the posts by a user to represent her interest. Then whether a user would be interested in a newly incoming message is determined by the similarity between feature vectors of user interests and the message [26, 28]. LDA has also been applied to extract user interests from user generated content [21]. Java *et al.* [23] looked into communities of users in the reciprocal Twitter follower network and summarized user intent into several categories (daily chatter, conversations, information sharing, and news updates); a user could talk about various topics with friends in different communities. Michelson and Macskassy discovered entities mentioned in tweets according to predefined folksonomy-based categories to allocate topics so as to build an entity-based topic profile [25]. The diversity of user interests has not yet been thoroughly investigated. An exception is the work of An *et al.*, who explored which news sources Twitter users are following and correlated

the observation with the diversity of their political opinions [29]. In this paper we propose a simple but powerful method to detect topics and infer user interests, as well as definitions of topical diversity of users and content.

Hashtag popularity has been examined from various perspectives, including their innate attractiveness [30, 31], the network diffusion processes [32–38], user behavior [28, 39, 40], and the role of influentials along with their adoption patterns [41, 42]. Romero *et al.* predicted popularity of a tag based on the social connections of its early adopters, but did not consider topicality and connections among tags [43]. Many methods for quantifying social impact and identifying influential users have been proposed. User influence can be quantified in terms of high in-degree in the follower network [44–46], information forwarding activity [45, 47], seeding larger cascades [41, 42], or topical similarity [9, 21]. We believe that the proposed measurement of topical diversity would prompt new approaches to the prediction of future hashtag popularity. Several previous studies have supported our intuition. For example, network diversity was shown to be positively correlated with regional economic development [48, 49]; community diversity at the early stage tend to boost the chances of a meme going viral [37, 38].

The social network topology is determined by how people are connected. Each individual is represented as a node and each following relationship as an edge linking a pair of users (see the bottom layer in Fig. 1). Hashtags spread among people through these social connections and can be mapped into a semantic space, in which each node is a tag and similar ones are coupled forming topic clusters (see the top layer in Fig. 1). Note that even though social media users can obtain information about certain events from external sources instead of their social connections, how they elect to describe these events can be very different from person to person. The emergence of consensus on the hashtag used to describe an event depends on a diffusion process on the social network [50]. By examining which topics are attached to a user's
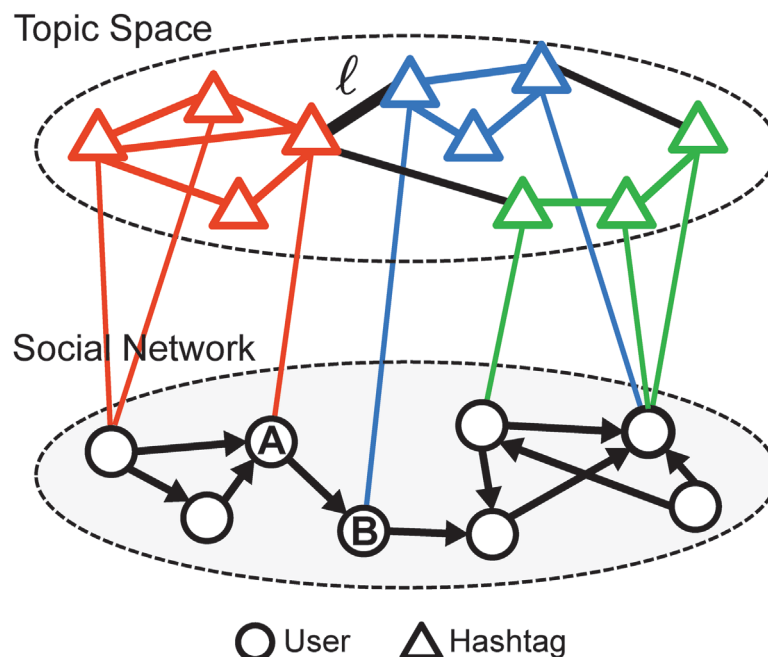

Fig 1. Representation of online conversation topics in social media by a multi-layer network. The social network connects people. In the topic network, nodes represent hashtags that are linked when they co-occur; clusters represent topics (shown in colors). A person and a hashtag are connected when the person uses the hashtag.

messages, we can infer her interests; by examining the topics of tags that co-occur with a given hashtag, we can learn what that hashtag is about. In reality we are able to observe the social network structure and information diffusion flows, but not topic formation in the semantic space. To the best of our knowledge, our understanding of the connection between these two layers of information diffusion is still in its infancy [43, 51].

Let us highlight the main contributions of this work:

- We develop a way to extract topics from online conversations. A network of hashtags is built by counting how many times a pair of hashtags appear together in a post. Communities (clusters of densely connected nodes) in such a network are found to well represent topics as sets of semantically related hashtags.

- Given a user, we gauge the diversity of his topical interests by examining to which topic clusters each of his hashtags belongs. We can thus distinguish users with diverse interests from those with focused attention. The topical diversity of a hashtag is measured similarly, by considering its co-occurring hashtags.

- We find that when a hashtag is adopted by people with diverse interests, or co-occurs with other tags on assorted themes, it is more likely for the tag to become popular. One interpretation is that diversity increases the probability of the hashtag being exposed to different audience groups. We show that the topical diversity of early adopters and co-occurring tags is a good predictor for the future popularity of hashtags.

- In contrast, high topical diversity is not a helpful factor in the growth of individual social impact. Focusing on one or a few topics may be a sign of expertise. Inactive users attract followers by mentioning a variety of topics, while active users tend to obtain many followers by maintaining focused topical interests. Focused topical preferences promote the content appeal of ordinary users and celebrities alike.

Methods

In this section, we introduce our dataset and define several key concepts to facilitate the subsequent discussion.

Data

We collected data from January to March 2013 using the Twitter streaming API (http://dev. twitter.com/docs/streaming-apis) without any customized query string to filter the tweets. We received a random sample of up to 10% of public messages, yielding a collection of about 1.2–1.3 billion tweets per month. We did not filter spammer accounts or bots from the dataset, as we expect the topic community structure to be robust with respect to a minority of irregular users. In the study, we set the first two months as the *observation period* and the last month as the *test period*; the former is used to build up the topic network and quantify user topical interests, and the latter for evaluating the results of prediction tasks. Table 1 shows several basic statistics about the dataset, which is publicly available. (http://thedata.harvard.edu/dvn/dv/topic_diversity) Note that about 13% of the tweets contain hashtags and about 43% of the users use hashtags.

Hashtags during March 2013 are used for prediction tasks. We are interested in newly emergent tags, so that we are able to identify the start of their lifetime and track their growth for at least three weeks. We select hashtags that do not appear during January and February 2013, but are used by at least three distinct users during March 2013. In addition, only tags with the first tweet observed during the first week of March are considered, so that we can track their

Table 1. Basic statistics of the dataset, which is split into two periods: observation and testing.
About 13% of the tweets contain hashtags.

	Jan-Feb 2013 (Observation)	Mar 2013 (Testing)
# Tweets	2,449,711,388	1,339,702,599
# Tweets with hashtags	316,668,998	173,823,786
# Hashtags	27,923,499	16,802,087
# Users	92,356,790	72,963,020

doi:10.1371/journal.pone.0118410.t001

usage during the whole month. Eventually, 509,868 hashtags (3.03% of all hashtags in March) were chosen as *emergent* hashtags.

Basic concepts

Let us first clarify several basic concepts as below.

User (*u*) *Users* are people using and creating content on the social media platform. They autonomously form the social network by connecting with one another, i.e. following on Twitter and friending on Facebook. Users are creators of online content, also referred to as *messengers*.

Hashtag (*h*) *Hashtag* are topical tags, words, or phrases following a hash symbol ('#') used by Twitter users to explicitly identify topics in tweets. A tweet may or may not contain hashtags. Here hashtags are also referred to as *messages*.

Hashtag adopter Users who post about a given hashtag are considered *adopters* of this hashtag.

Popularity Popular hashtags are expected to attract many adopters and top popular hashtags are deemed viral. Given a hashtag, we define the total number of adopters after a month as a measure of its *popularity* or *virality*.

Tweets A *tweet* is a post generated by a Twitter user.

Followers Twitter users can subscribe to others to receive their tweets; a subscriber is known as a *follower*.

Retweets Twitter users can forward interesting posts to their followers, seeding cascades of interesting content on the follower social network. The forwarded posts are *retweets*.

Content interestingness (*β*) A measure of how *interesting* is the content posted by a given user. Lerman [52] studied the interestingness of online content on Digg and defined it as "the probability it will get reposted when viewed."

Topic clusters

Hashtags are explicit topic identifiers on Twitter that are invented autonomously by millions of content generators. Since there is no predefined consensus on how to name a topic, multiple hashtags may be developed to represent the same event, theme, or object. For instance, #followback, #followfriday, #ff, #teamfollowback, and #tfb are all about asking others to follow someone back or suggesting people to follow; #tcot, #ttxcot, #twcot, and #ccot label politically conservative groups on Twitter. To reduce the duplication, we shift attention from single hashtags to more general categories—clusters of semantically similar hashtags—that we call *topic clusters*.

Table 2. Examples of topic clusters in the hashtag co-occurrence network.

Cluster	Example #hashtags
Technology	google, microsoft, supercomputers, ibm, wikipedia, startuptip, topworkplace
Politics	tcot, p2, top, usgovernment, dems, owe, politics, teaparty
Lifestyle	pizza, pepsi, health, vacation, caribbean, ford, honda, volkswagen, hm, timberland
Twitter-specific	followme, followfriday, friday, justretweet, instantfollower, rt
Mobile devices	apple, galaxy3, note2, iosapp, mp3player, releases

doi:10.1371/journal.pone.0118410.t002

Under the topic locality assumption that semantically similar hashtags are more likely to appear in the same tweets together, such topic clusters are expected to be densely connected. We detect these clusters by finding communities in the hashtag co-occurrence network. First we recover the network by only considering hashtags used by at least three distinct users and join occurrences observed in at least three messages. We do this to filter out noise from accidental co-occurrence and spam. The recovered hashtag co-occurrence network contains 974,529 nodes and 7,325,492 edges. Each node represents a hashtag and each edge is weighted by the number of messages containing both tags. We expect to discover topic clusters based on the assumption that related tags tend to appear together frequently and to form densely connected components. Several community detection methods are designed to identify such dense clusters. We selected the Louvain method [10] because of its efficiency. With the Louvain method, we obtain 37,067 communities (level 2 in the hierarchical community structure). As exemplified in Table 2, communities in the hashtag co-occurrence network capture coherent topics. At the macroscopic level we can still observe strong topic locality (see Fig. 2). We do not expect the use of alternative clustering algorithms to affect the results; indeed, the choice of community detection method was not a factor in our prior work where we measured meme diversity based on social communities [37, 38].

It is common for people to use different spellings for hashtags that refer to the same topic. These tag variations can be captured in the proper topical clusters if they frequently co-occur with each other, or with other related tags. One illustrative case if that of hashtag abbreviations. For instance, both #followfriday and its abbreviation #ff can be used to suggest people to follow
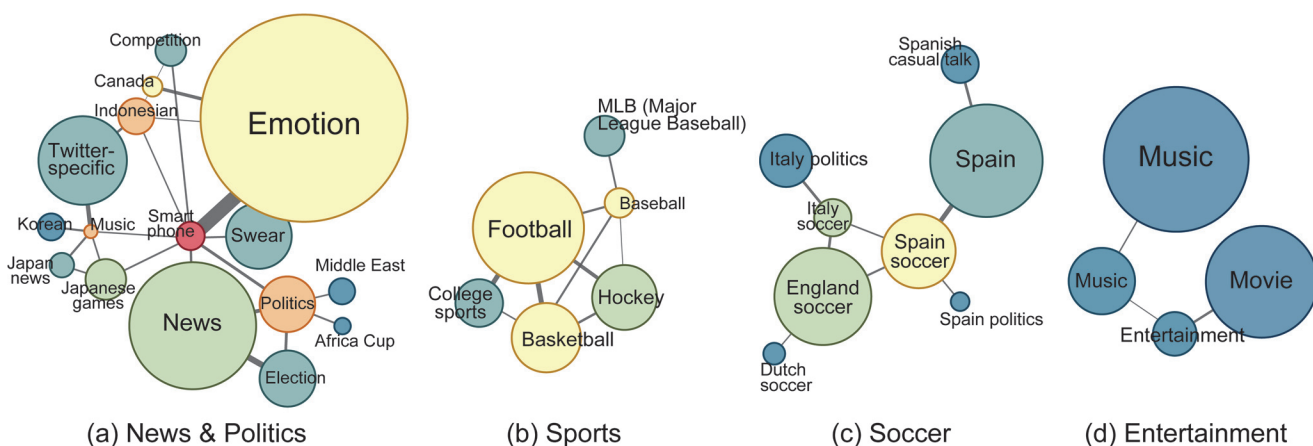

(a) News & Politics   (b) Sports   (c) Soccer   (d) Entertainment

Fig 2. Examples of connected topic clusters of related themes. Themes include: (a) news and politics, (b) sports, (c) soccer, and (d) music and entertainment. Each node represents a cluster of hashtags on the topic as labeled; the area is proportional to the number of hashtags that the topic cluster contains; the color is assigned according to the degree, so that high degree clusters are more red and low degree clusters more blue. All these examples are consistent with the existence of topic locality.

doi:10.1371/journal.pone.0118410.g002

on Friday; #occupywallstreet and #ows both refer to a protest movement against social and economic inequality. We observed high co-occurrence between #followfriday and #ff, as well as between #occupywallstreet and #ows. This implies a tendency toward adoption of a consensus representation for a specific topic. In other cases, tags may not be used together but they may share semantically related neighbors in the hashtag co-occurrence network. As an example, the related political hashtags #twcot (top women conservatives on Twitter) and #txcot (Texas conservatives on Twitter) do not co-occur in the same tweets, but both appear together with #tcot (top conservatives on Twitter) quite often. In both cases, our method is able to capture hashtags that are about the same topic in the same clusters, irrespective of spelling variations.

Diversity of user interests

Given a messenger u, we can track the sequence of hashtags (with repetition) that he used in the past, $h_1, h_2, \ldots, h_{n_u}$. Each hashtag h_i is attached to a topic $T(h_i)$, given by:

$$T(h_i) = \begin{cases} C(h_i) & \text{if } h_i \text{ exists in the hashtag co} - \text{occurrence network} \\ h_i & \text{otherwise} \end{cases} \quad (1)$$

where $C(h)$ is a community containing h in the hashtag co-occurrence network. The set of distinct topics associated with all of u's hashtags is denoted as $\mathbb{T}_u$, $T(h_i) \in \mathbb{T}_u$. The topical diversity of a user's interests can be estimated by computing the entropy of hashtags across topics:

$$H_1(u) = - \sum_{T_j \in \mathbb{T}_u} P(T_j) \log P(T_j) \quad (2)$$

where

$$P(T_j) = \frac{1}{n_u} |\{h_i : T(h_i) = T_j, 1 \leq i \leq n_u\}|. \quad (3)$$

Table 3 compares two people, each having used a set of 10 distinct hashtags 20 times. User A was interested in trendy Twitter-specific tags almost exclusively (low $H_1(A)$), while user B paid attention to a set of very diverse conversations about countries, movies, books, and horoscope (high $H_1(B)$). Note that the opposite (and wrong!) conclusion, $H_1(A) > H_1(B)$, would be drawn had we measured entropy based on hashtags rather than topic clusters.

Table 3. Comparison of two users with different diversity of topical interest.

User	C	#Hashtag (usage count)
A	20	nowplaying(1)
	96	rt(4), follow(3), tfb(2), ff(2), 500aday(2), teamfollow(2), teamfollowback(2), f4f(1), rt2gain(1)
		$n_A = 20$, $|\mathbb{T}_A| = 2$, $H_1(A) = 0.2864$
B	9	australia(1)
	20	cosmicconsciousness(1), thenotebook(1)
	33	thedescendants(1)
	57	friendswithbenefits(1)
	79	thepowerofnow(2)
	139	gemini(8), geminis(2)
	806	tdl(2)
	–	tipfortheday (1)
		$n_B = 20$, $|\mathbb{T}_B| = 8$, $H_1(B) = 2.3610$

doi:10.1371/journal.pone.0118410.t003

Diversity of content

Similarly, given a hashtag h, we recover the sequence of other hashtags (with repetition) that co-occurred with it, $h_1, h_2, \ldots, h_{m_h}$. Each co-occurring hashtag (*co-tag*) h_i is assigned to topic $T(h_i)$ based on the topic cluster to which it belongs (see Equation 1). Then the co-tag diversity of h, $H_2(h)$, is measured in the same way as the user diversity H_1 (see Equation 2).

Results

In this section we test two prediction tasks using the concept of topical diversity. The first task aims at predicting the popularity of hashtags, measured as the total number of adopters after a month, given the information about early adopters and co-occurring tags. The second is to predict the social influence of an individual, quantified by how many times he or she is retweeted by others, using diversity-based measures.

Predicting hashtag popularity

Do diversity measures help us detect hashtags that will go viral in the future? We first explore whether the topical diversity of a hashtag's adopters or its co-tags can imply its future popularity.

Prediction via user diversity. Hashtags in Twitter can be treated as channels connecting people with shared interests, because hashtags label and index messages enabling people to easily retrieve information and broadcast to certain groups. As illustrated in Fig. 3, users with focused interests are linked with few groups, while people who care about diverse issues are exposed to a larger number of interest groups through hashtag channels. We expect the latter category of users to play a critical bridging role, connecting many groups in the network. This would allow them to spread innovative information across groups, as suggested by the weak tie hypothesis [53], thus boosting the diffusion of hashtags [37, 38, 54]. In other words, we hypothesize that if a hashtag has early adopters with diverse topical interests, it is more likely to go viral.

Given a hashtag h, we track the users who adopt it within t hours after h is created and compute the average interest diversity among these early adopters as a simple predictor. Irrespective of how long we track, we observe a positive correlation between the average user diversity and the future popularity of the hashtags, measured as the total number of adopters after one month (see Fig. 4a).
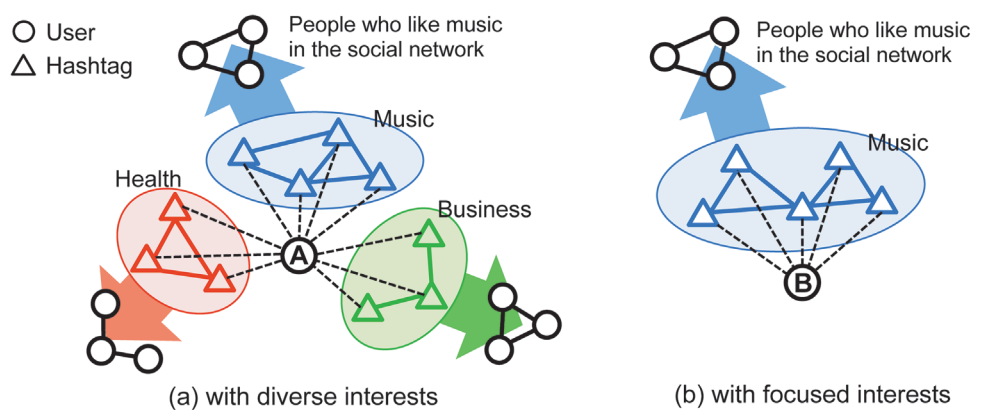

Fig 3. **User A has diverse topical interests and user B displays more focused interests.** Each connected group in the illustration corresponds to a social circle with common interests.

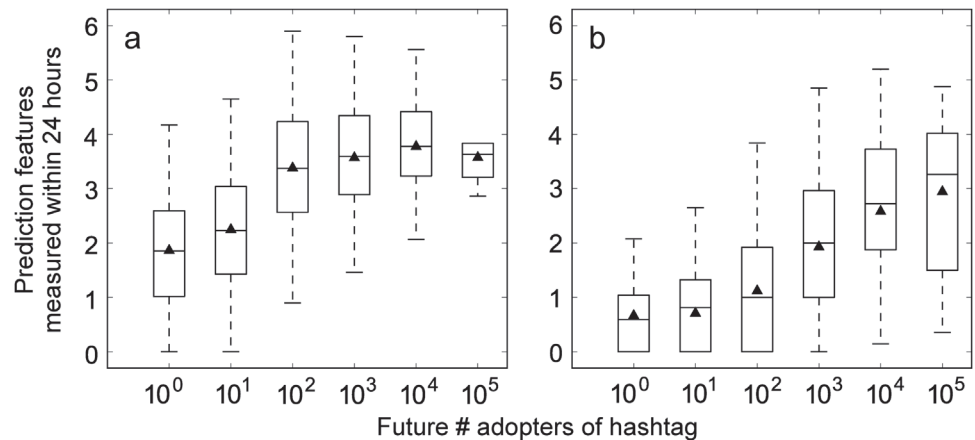doi:10.1371/journal.pone.0118410.g003

Fig 4. Correlation between topical entropy and hashtag future popularity. (a) Correlation between the average topic-based entropy H_1 of adopters in the first 24 hours and the total number of future hashtag adopters. (b) Entropy H_2 of hashtags co-occurring in the first 24 hours across topic clusters as a function of future popularity of emergent hashtags.

To better evaluate the predictive power of adopter diversity, let us run a simple prediction task based on information at the early stage to forecast which hashtags from the test period will be viral in the future. Different formulations of this task have been proposed in the literature [38, 55]. Here, a hashtag is deemed viral if the number of distinct adopters at the end of the test period is above a given threshold. Our evaluation algorithm has three steps:

i) For each feature, we compute its value for each newly emergent hashtag h in the test period based on the set of early adopters of h within t hours after the birth of h. A hashtag is born when the first tweet containing it appears. The feature is either a measure of user characteristics averaged among early adopters, or a linear combination of several such measures. We track adoption events for $t = 1, 6,$ and 24 hours since birth.

ii) Hashtags are ranked by the feature values in descending order.

iii) We set a percentile threshold for labeling popular hashtags. The most popular hashtags are deemed viral. Based on this ground truth, we can measure false positive and true positive rates and draw a receiver-operating-characteristic (ROC) plot. The area under the ROC curve (AUC) is our evaluation metric. The higher the AUC value, the better the feature as a predictor of future hashtag popularity.

We consider several user attributes of early adopters that have been shown in the literature to be strong predictors of virality [38, 44, 45, 47, 56]. These include the number of early adopters n, number of followers fol (potential audience), and number of tweets twt that a user has produced during the observation period (activity). We additionally consider the diversity of topical interests, H_1. The goal of our experiment is not to achieve the highest accuracy (a task for which different learning algorithms could be explored). Rather, we aim to compare the predictive powers of different features. Therefore we focus on the relative differences between AUC values generated by single or combined features rather than on the absolute AUC values. AUC values measured using different features are listed in Table 4. Among individual features, n is the most effective. When we combine it with other features, fol yields high AUC consistently, but the differences are very small. The performance of the diversity metric is competitive,

Table 4. AUC of prediction results using different adopter features within *t* early hours. Prediction features include the number of followers (*fol*), the number of tweets (*twt*), the diversity of topical interests of adopters (H_1), and the number of early adopters (*n*). The threshold is expressed as a top percentile of most popular hashtags that are deemed viral for evaluation purposes. Best results for each column are bolded.

Threshold	50%			10%			1%			0.1%		
t (hours)	1	6	24	1	6	24	1	6	24	1	6	24
twt	0.57	0.58	0.59	0.62	0.63	0.64	0.67	0.68	0.68	0.70	0.71	0.72
fol	0.57	0.58	0.59	0.62	0.64	0.65	0.67	0.69	0.70	0.74	0.75	0.76
H_1	0.55	0.55	0.55	0.58	0.58	0.58	0.60	0.60	0.61	0.63	0.63	0.64
n	0.64	**0.68**	0.71	0.75	0.79	0.84	0.82	0.86	**0.90**	0.86	**0.89**	0.91
n + twt[†]	0.64	**0.68**	0.71	0.75	**0.80**	0.84	**0.83**	**0.87**	**0.90**	**0.87**	**0.89**	**0.92**
n + fol[†]	**0.65**	**0.68**	**0.72**	**0.76**	**0.80**	**0.85**	**0.83**	**0.87**	**0.90**	**0.87**	**0.89**	**0.92**
n + H_1[†]	0.64	**0.68**	0.71	0.75	**0.80**	**0.85**	**0.83**	0.86	**0.90**	0.86	**0.89**	**0.92**

† A linear combination with coefficients determined by regression fitting using least squared error.

matching the top results in several experimental configurations. These results are not particularly sensitive to the popularity threshold or the duration of the early observation window.

Prediction via content diversity. In this section we examine whether the future popularity of a hashtag is affected by the topical diversity of its early co-occurring tags. How people apply hashtags to label their messages depicts their topical interests and determines the topology of the tag co-occurrence network. In Fig. 1, a link between the topic layer and the social layer of the network marks an association between a user and a hashtag. This tag may attract an audience in the social network. The co-occurrence of two tags extends the audience groups of both. For example, link ℓ in Fig. 1 exposes user A to the blue topic and user B to the red cluster. Therefore we expect a hashtag to be exposed to more potential adopters, making it more likely to go viral, if it often co-occurs with many other hashtags. To test this hypothesis, we measure the number *m* of co-tags. Furthermore, if co-tags are very popular, we would expect a stronger effect because they would provide a larger audience. We therefore measure the popularity of co-tags in terms of numbers of tweets *T* and adopters *A* during the observation period. And if co-tags are about diverse topics, this may further boost the effect by extending the audience to many groups with small overlap. In conclusion, we hypothesize that co-occurring with many popular tags about diverse topics should be a sign that a hashtag will grow popular.

Given an emergent hashtag *h*, we track other tags that co-occur with *h* within *t* hours after *h* is born and measure the topical diversity H_2 of these co-tags. We observe a positive correlation between the diversity of early co-tags and the future popularity of the tag (see Fig. 4b). Then we apply the same method as in the prediction via the user diversity to test the predictive power of different traits associated with early co-tags. In this case, the prediction features for each target hashtag are computed based on early co-tags instead of adopters. Again, the goal of our experiment is to compare the predictive power of different features, thus we examine the relative differences in AUC values generated by the various traits. The results are reported in Table 5. The number *m* of co-tags observed in the early stage is the best single predictor of virality. When we combine *m* with a second feature, co-tag diversity provides the best results irrespective of the threshold or the duration of the early observation window. Interestingly, *m* and H_2 are both about diversity and perform better than the popularity-based features *T* and *A*.

In the discussion above, we evaluate the predictive powers of two categories of features for identifying future viral hashtags. These two sets of features, based on early adopters and co-tags, respectively, have different effectiveness. By comparing the AUC values in Tables 4 and 5, we find that adopter features yield better results. However, they also require additional

Table 5. AUC of prediction results using different co-tag features within t early hours. Prediction features include the number of tweets containing the co-tags (T), the number of co-tag adopters (A), the diversity of co-tags (H_2), and the number of observed co-tags (m). The threshold is expressed as a top percentile of most popular hashtags that are deemed viral for evaluation purposes. Best results for each column are bolded.

Threshold	50%			10%			1%			0.1%		
t (hour)	1	6	24	1	6	24	1	6	24	1	6	24
T	0.50	0.51	0.52	0.51	0.53	0.55	0.58	0.62	0.66	0.66	0.72	0.75
A	0.50	0.50	0.52	0.50	0.52	0.54	0.58	0.62	0.65	0.65	0.71	0.74
H_2	0.50	0.51	0.53	0.52	0.54	0.57	0.61	0.66	0.70	0.70	0.77	0.82
m	0.52	0.53	0.55	0.55	0.58	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + T^\dagger$	0.52	0.53	0.55	0.54	0.57	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + A^\dagger$	0.52	0.53	0.55	0.55	0.57	0.61	0.64	**0.70**	**0.75**	0.72	**0.81**	**0.86**
$m + H_2^\dagger$	**0.55**	**0.55**	**0.57**	**0.58**	**0.60**	**0.63**	**0.66**	**0.70**	**0.75**	**0.74**	**0.81**	**0.86**

$\dagger$ A linear combination with coefficients determined by regression fitting using least squared error.

doi:10.1371/journal.pone.0118410.t005

prerequisite knowledge: in addition to tracking hashtag co-occurrences for building the topic network, we also need to record user-generated content. The features built upon early co-tags are less expensive, but the performance is slightly worse; a possible interpretation for this is that few tweets in the observation window may contain co-occurring tags, while they all have associated users. Therefore co-tag features are more sparse. Depending on what type of information is available, one might choose either approach or a combination of both.

Social influence

High topical diversity of adopters and co-occurring tags is a positive sign that a hashtag is growing popular, as shown in the previous section. However, does high topical diversity also signal a growth in individual influence? On one hand, when an individual talks about various topics, she may have contact with many others through shared interests or hashtags, thus attracting more attention (see Fig. 3). On the other hand, focused interest may enhance expertise in specific fields, thus increasing the content interestingness and shareability. In this light, low diversity triggered by expertise might help people become popular. In this section we evaluate these two contradictory hypotheses.

Some people are more influential than others in persuading friends to adopt an idea, an action, or a piece of information. The concept of *social influence* has been discussed extensively in social media research. Most of the studies in the literature have considered user activity [44], number of followers [44, 47], capability to trigger large cascades [41, 42], or to be retweeted or mentioned a lot [44, 45, 47] as signals of high social influence. Which user characteristics make people popular and influential? Does the diversity of individual topical interests play a role in the social influence processes? Let us consider several individual properties:

Number of retweets (RT) How many times an individual is retweeted during the observation time period. We consider RT as a direct indicator of social influence, since it quantifies how many times the user succeeds in making others adopt and spread information. Due to the settings of the Twitter API, the number of retweets per user that we collect includes all the retweeters in every cascade. That is, suppose user B retweets user A and then C retweets B; both tweets are counted in RT for A, even though C did not directly retweet A. However, since the majority of information cascades are very shallow [42], RT is a good approximation of the direct retweet count. The number of retweets is dependent on the length of the

observation window, because we believe that social influence is accumulated in time and requires long-term endeavor [44].

Number of followers (*fol*) The number of followers suggests how many people can potentially view a message once the user posts it; the more followers a user has, the higher the chances that his content can spread out to a large audience.

Number of tweets (*twt*) The number of tweets generated by the user; the higher the number, the more active the user.

Content interestingness (β) The interestingness of the content posted by a given user is defined as "the probability it will get retweeted when viewed" [52]. To measure β in the Twitter context, we assume that the value of *RT* for an individual is proportional to the number of tweets *twt* she produced, the number of followers *fol*, and the appeal β of the content:

$$\beta \propto \frac{RT}{twt \cdot fol}. \tag{4}$$

Diversity of interests (H_1) See the definition of diversity of user interests, H_1, in Methods (Equation 2).

Table 6 lists the results of a linear regression estimating how many times a user is retweeted according to several user features. Intuitively, users with many followers are more likely to spread their messages and thus get retweeted more frequently, because they have many more potential viewers. The number of followers is the most important factor, as supported by the largest positive coefficient in the regression. The number of generated tweets also has a positive coefficient in the regression, implying that being active helps users get retweeted more. The result confirms several existing studies suggesting that high social influence requires long-term, consistent effort [44, 45]. The interestingness of the story is positively correlated with social influence as well, although not as strongly as the other factors. Finally, the negative coefficient of diversity in Table 6 suggests that users with diverse interests tend to have low influence. This supports the hypothesis that social influence is topic-sensitive, requiring expertise in a specific field [21]; posting about the same topic is more effective for gaining social influence, compared to commenting on many different subjects. In summary, people can acquire social influence by having a big audience group, being productive, creating interesting content, and staying focused on a field. Unfortunately, it seems that there is no simple recipe of success.

We illustrate how several user properties are related to the number of followers and the topical diversity of user interests in Fig. 5. Most users have a small number of followers and low entropy (Fig. 5a). Active users tend to have high diversity, unsurprisingly (Fig. 5b). The number

Table 6. Linear regression estimating how many times a user is retweeted. For efficiency, the regression is based on a random sample of 10% of the users (N = 2,171,624).

	Coefficient	SE
(Intercept)	20.9	0.5
Num. followers (*fol*)†	193.0 ***	0.5
Num. tweets (*twt*)†	51.1 ***	0.5
Content interestingness (β)†	3.9 ***	0.5
Diversity of interests (H_1)†	-9.1 ***	0.5

† Variables are normalized by *Z*-score.

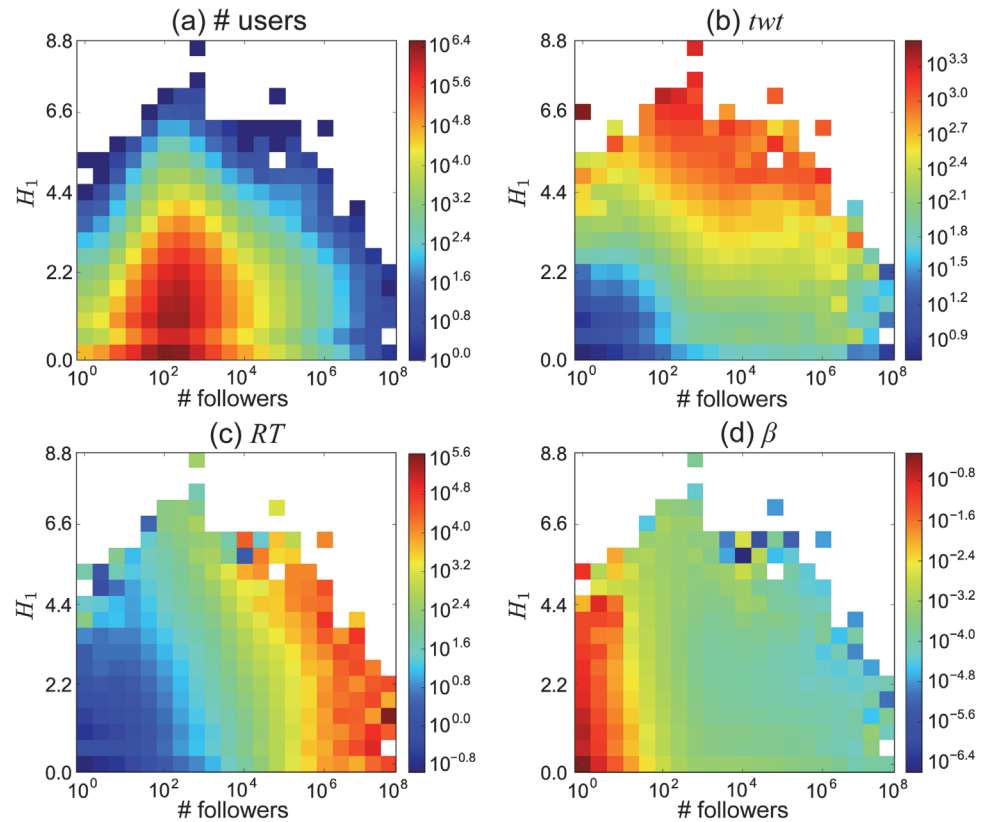*** $p < 0.001$

doi:10.1371/journal.pone.0118410.t006

Fig 5. Heatmaps of different metrics as a function of topical diversity and the follower count.
Heatmaps of (a) the number of users, (b) the number of tweets generated, (c) how many times a user is retweeted, and (d) the content interestingness of a user, as a function of the diversity of topical interests, H_1, and the number of followers in the observation window.

doi:10.1371/journal.pone.0118410.g005

of followers is shown in Fig. 5c to be a powerful factor to get retweeted more often, consistently with the regression results in Table 6. Finally, the content interestingness appears to be anti-correlated with the number of followers but not correlated with user diversity (Fig. 5d).

Active vs. inactive users. Let us explore how the number of followers a user can attract is affected by the diversity of topical interests. The entropy measure for diversity is biased by user activity: generating more tweets with more hashtags tends to yield higher entropy (see Fig. 5b). Thus we group users by productivity, so that individuals in the same group have comparable values of topical diversity. For users in the same group, we compute the Spearman rank correlation between the number of followers and diversity. We use Spearman because, unlike Pearson, it does not require that both variables be normally distributed. According to the plot in Fig. 6, low-engagement users attract followers by talking about various topics, while active users tend to obtain many followers by maintaining focused topical interests. For the most active users, topical diversity is not relevant; many of these accounts are spammers and bots.

Celebrities vs ordinary users. When looking into the effect of interest diversity on content appeal, we need to control for the number of followers, since our interestingness measure is strongly correlated with the number of followers (see Fig. 5d). The negative correlations shown in Fig. 7 suggest that in general, focused posts promote content appeal. One possible interpretation is that people follow someone for a reason and the provided content has to be consistent in order to match such expectations; one is less likely to share a tip on cosmetics from a
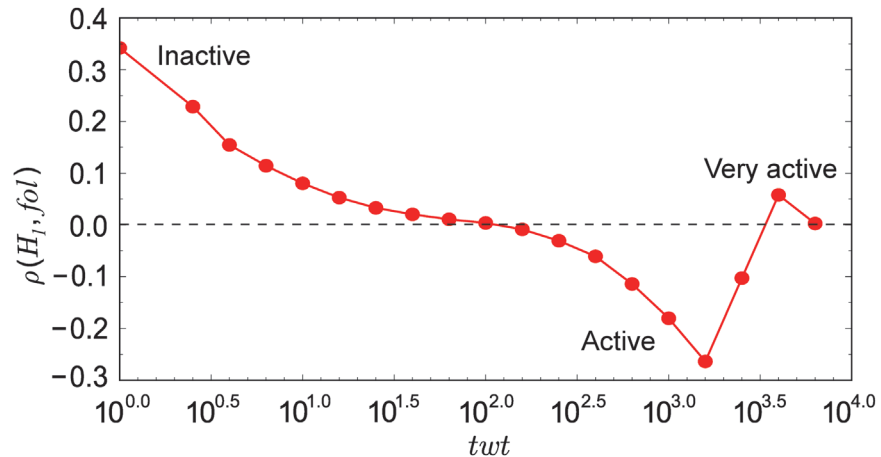
Fig 6. Spearman rank correlation between the number of followers and the topical diversity of user interests as a function user activity. All shown correlation values are significant ($p < 0.05$).
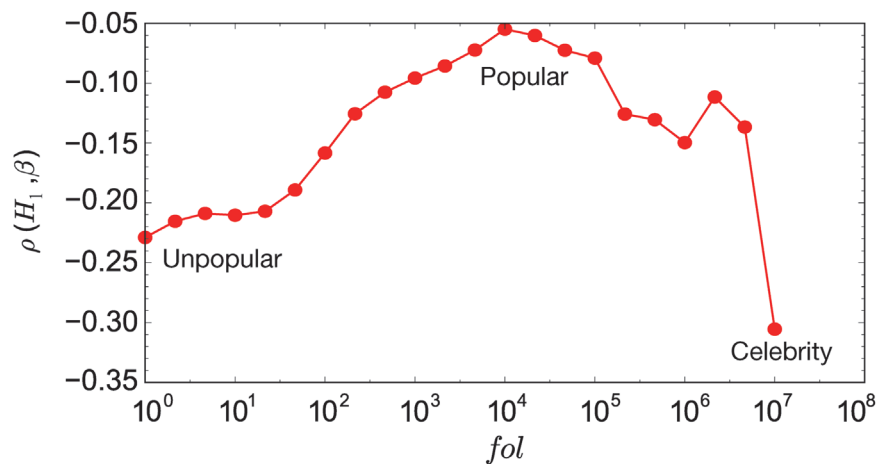
Fig 7. Spearman rank correlation between content interestingness and the topical diversity of user interests as a function of how many followers users have. All shown correlation values are significant ($p < 0.05$).

politician. This effect is stronger for users with few followers and celebrities; people with moderate popularity generate retweets with focused and diverse content.

Discussion

We proposed methods to identify topics using Twitter data by detecting communities in the hashtag co-occurrence network, and to quantify the topical diversity of user interests and content, defined by how tags are distributed across different topic clusters.

One might wonder if topical locality, and thus the quality of the topical diversity measures based on hashtag communities, are hindered by the widespread use of generic, non-topical tags. For example, some people use #rt when retweeting a message from someone else, to show agreement or to suggest content to their followers; the message that gets retweeted in this scenario can be about any topic. This kind of generic hashtags can be applied in various contexts and

spread across many topic clusters, diluting the coherence of each topic theme. To check if this is the case, we examined the top 10 generic hashtags. They are #rt, #ff, #np, #oomf, #teamfollowback, #retweet, #tbt, #followback, #follow, and #openfollow in our dataset. Eight of them actually belong to the same community (the "Twitter-specific" cluster in Table 2), the other two in another. The average entropy of these generic tags, in terms of how they are distributed across clusters, is $H = 0.9$, much lower than expected for 10 randomly sampled tags ($\langle H \rangle = 2.8$ averaged across 100 trials). We conclude that generic tags are often employed together; the community structure in the hashtag co-occurrence network interprets generic hashtags as similar topics.

We found that popular hashtags tend to have adopters who care about various issues and to co-occur with other tags of diverse themes at the early stage of diffusion. One practical application evaluated in this paper is to predict viral hashtags using features built upon the topical diversity of early adopters or co-tags. In the prediction using information on early adopters, the performance of topical diversity is competitive with other user features while combined with the number of early adopters. In the prediction via early co-occurring hashtags, features about diversity, including the number of early co-tags and their topical diversity, outperform popularity-based features. However, high topical diversity is not a positive factor for individual popularity. High social influence is more easily obtained by having a big audience group, producing lots of interesting content, and staying focused. In short, diverse messages and focused messengers are more likely to generate impact.

The interesting observation that high diversity helps a hashtag grow popular but does not help develop personal authority originates from the different mechanisms by which a hashtag and a user attract attention. In the diffusion process of a hashtag, adopters with diverse interests play a role as bridges connecting different groups and thus positively improve the visibility of the tag. These results are consistent with Granovetter's theory [53], as well as our recent findings on the strong link between community diversity and virality [37, 38]. On the other hand, a user gains social influence through expertise or authority within a cohesive group with common interests.

Topical diversity provides a simple yet powerful way to connect the social network topology with the semantic space extracted from online conversation. We believe that it holds great potential in applications such as predicting viral hashtags and helping users strengthen their online presence.

Acknowledgments

Author Contributions

Conceived and designed the experiments: LW FM. Performed the experiments: LW. Analyzed the data: LW FM. Wrote the paper: LW FM.

References

1. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, Brewer D, et al. Computational social science. Science. 2009; 323: 721–723. doi: 10.1126/science.1167742 PMID: 19197046

2. Vespignani A. Predicting the behavior of techno-social systems. Science. 2009; 325: 425–428. doi: 10.1126/science.1171990 PMID: 19628859

3. Leskovec, J, Backstrom, L, Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In: Proc. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD); 2009. pp. 497–506.

4. Xie, L, Natsev, A, Kender, JR, Hill, M, Smith, JR. Visual memes in social media: tracking real-world news in youtube videos. In: Proc. ACM Intl. Conf. on Multimedia; 2011. pp. 53–62.

5. Simmons, MP, Adamic, LA, Adar, E. Memes online: Extracted, subtracted, injected, and recollected. In: Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM); 2011. pp. 17–21.

6. Agarwal MK, Ramamritham K, Bhide M. Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments. Proc of VLDB Endowment. 2012; 5: 980–991. doi: 10.14778/2336664.2336671

7. Ferrara, E, JafariAsbagh, M, Varol, O, Qazvinian, V, Menczer, F, Flammini, A. Clustering memes in social media. In: IEEE/ACM Intl. Conf. on Advances in Social Networks Analysis and Mining (ASONAM); 2013. pp. 548–555.

8. Aggarwal, CC, Subbian, K. Event detection in social streams. In: SIAM Intl. Conf. on Data Mining (SDM); 2012. pp. 624–635.

9. Tang, J, Sun, J, Wang, C, Yang, Z. Social influence analysis in large-scale networks. In: Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD); 2009. pp. 807–816.

10. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008; 10: P10008. doi: 10.1088/1742-5468/2008/10/P10008

11. Newman MEJ. Communities, modules and large-scale structure in networks. Nature Physics. 2012; 8: 25–32. doi: 10.1038/nphys2162

12. Davison, BD. Topical locality in the web. In: Proc. ACM SIGIR Intl. Conf. on Information retrieval; 2000. pp. 272–279.

13. Menczer F. Lexical and semantic clustering by web links. Journal of the American Society Information Science and Technology. 2004; 55: 1261–1269. doi: 10.1002/asi.20081

14. Menczer F, Belew RK. Adaptive information agents in distributed textual environments. In: Proc. Intl. Conf. on Autonomous Agents; 1998. pp. 157–164.

15. Goldberg D, Nichols D, Oki B, Terry D. Using collaborative filtering to weave an information tapestry. Communications of ACM. 1992; 35: 61–70. doi: 10.1145/138859.138867

16. Goldberg K, Roeder T, Gupta D, Perkins C. Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval. 2001; 4: 133–151. doi: 10.1023/A:1011419012209

17. Schifanella R, Barrat A, Cattuto C, Markines B, Menczer F. Folks in folksonomies: social link prediction from shared metadata. In: Proc. ACM Intl. Conf. on Web search and data mining (WSDM); 2010. pp. 271–280.

18. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F. Friendship prediction and homophily in social media. ACM Trans Web. 2012; 6: 9. doi: 10.1145/2180861.2180866

19. Haveliwala TH. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans on Knowledge and Data Engineering. 2003; 15: 784–796. doi: 10.1109/TKDE.2003.1208999

20. Michlmayr E, Cayzer S. Learning user profiles from tagging data and leveraging them for personal (ized) information access. In: Proc. of Intl. World Wide Web Conf; 2007. pp. 1–7.

21. Weng J, Lim EP, Jiang J, He Q. Twitterrank: finding topic-sensitive influential twitterers. In: Proc. ACM Intl. Conf. on Web search and data mining (WSDM); 2010. pp. 261–270.

22. Weng L, Lento T. Topic-based clusters in egocentric networks on Facebook. In: Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM); 2014. pp. 623–626.

23. Java A, Song X, Finin T, Tseng B. Why we twitter: Understanding microblogging usage and communities. In: Proc. 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis; 2007. pp. 56–65.

24. Phelan O, McCarthy K, Smyth B. Using Twitter to recommend real-time topical news. In: Proc. ACM Intl. Conf. on Recommender Systems; 2009. pp. 385–388.

25. Michelson M, Macskassy SA. Discovering users' topics of interest on Twitter: a first look. In: Proc. 4th ACM Workshop on Analytics for noisy unstructured text data; 2010. pp. 73–80.

26. Chen J, Nairn R, Nelson L, Bernstein M, Chi E. Short and tweet: experiments on recommending content from information streams. In: Proc. ACM Intl. Conf. on Human factors in computing systems (CHI); 2010. pp. 1185–1194.

27. Xu Z, Lu R, Xiang L, Yang Q. Discovering user interest on Twitter with a modified authortopic model. In: Proc. IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology (WI-IAT). 2011; 1: pp. 422–429.

28. Weng L, Flammini A, Vespignani A, Menczer F. Competition among memes in a world with limited attention. Scientific Reports. 2012; 2: 335. doi: 10.1038/srep00335 PMID: 22461971

29. An J, Cha M, Gummadi PK, Crowcroft J. Media landscape in Twitter: A world of new conventions and political diversity. In: Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM); 2011. pp. 18–25.

30. Berger J, Milkman KL. What makes online content viral? Journal of Marketing Research. 2009; 49: 192–205. doi: 10.1509/jmr.10.0353

31. Salganik M, Dodds P, Watts D. Experimental study of inequality and unpredictability in an artificial cultural market. Science. 2006; 311: 854–856. doi: 10.1126/science.1121066 PMID: 16469928

32. Daley DJ, Kendall DG. Epidemics and rumours. Nature. 1964; 204: 1118–1119. doi: 10.1038/2041118a0 PMID: 14243408

33. Goffman W, Newill VA. Generalization of epidemic theory: An application to the transmission of ideas. Nature. 1964; 204: 225–228. doi: 10.1038/204225a0 PMID: 14212412

34. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Physical Review Letters. 2001; 86: 3200–3203. doi: 10.1103/PhysRevLett.86.3200 PMID: 11290142

35. Moreno Y, Nekovee M, Pacheco AF. Dynamics of rumor spreading in complex networks. Physical Review E. 2004; 69: 066130. doi: 10.1103/PhysRevE.69.066130

36. Aral S, Walker D. Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management Science. 2011; 57: 1623–1639. doi: 10.1287/mnsc.1110.1421

37. Weng L, Menczer F, Ahn YY. Virality prediction and community structure in social networks. Scientific Reports. 2013; 3: 2522. doi: 10.1038/srep02522 PMID: 23982106

38. Weng L, Menczer F, Ahn YY. Predicting meme virality in social networks using network and community structure. In: Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM); 2014. pp. 535–544.

39. Ienco D, Bonchi F, Castillo C. The meme ranking problem: Maximizing microblogging virality. In: IEEE Intl. Conf. on Data Mining Workshops (ICDMW); 2010. pp. 328–335.

40. Yang L, Sun T, Mei Q. We know what @you #tag: does the dual role affect hashtag adoption? In: Proc. Intl. World Wide Web Conference (WWW); 2012. pp. 261–270.

41. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nature Physics. 2010; 6: 888–893. doi: 10.1038/nphys1746

42. Bakshy E, Hofman JM, Mason WA, Watts DJ. Everyone's an influencer: quantifying influence on twitter. In: Proc. ACM Intl. Conf. on Web search and data mining (WSDM); 2011. pp. 65–74.

43. Romero DM, Tan C, Ugander J. On the interplay between social and topical structure. In: Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM); 2013. pp. 516–525.

44. Cha M, Haddadi H, Benevenuto F, Gummadi PK. Measuring user influence in Twitter: the million follower fallacy. In: Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM); 2010. pp. 10–17.

45. Suh B, Hong L, Pirolli P, Chi EH. Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In: Proc. IEEE Intl. Conf. on Social Computing; 2010. pp. 177–184.

46. Weng L, Ratkiewicz J, Perra N, Gonçalves B, Castillo C, Bonchi F, et al. The role of information diffusion in the evolution of social networks. In: Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD); 2013. pp. 356–364.

47. Romero DM, Galuba W, Asur S, Huberman BA. Influence and passivity in social media. In: Proc. 20th Intl. Conf. on World Wide Web (WWW Companion Volume); 2011. pp. 113–114.

48. Quigley JM. Urban diversity and economic growth. Journal of Economic Perspectives. 1998; 12: 127–138. doi: 10.1257/jep.12.2.127

49. Eagle N, Macy M, Claxton R. Network diversity and economic development. Science. 2010; 328: 1029–1031. doi: 10.1126/science.1186605 PMID: 20489022

50. Lin YR, Margolin D, Keegan B, Baronchelli A, Lazer D. #Bigbirds never die: understanding social dynamics of emergent hashtag. In: Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM); 2013. pp. 370–379.

51. Serrano MA, Krioukov D, Boguná M. Self-similarity of complex networks and hidden metric spaces. Physical Review Letters. 2008; 100: 078701. doi: 10.1103/PhysRevLett.100.199902 PMID: 18352602

52. Lerman K. Social information processing in news aggregation. IEEE Internet Computing. 2007; 11: 16–28. doi: 10.1109/MIC.2007.136

53. Granovetter M. The strength of weak ties. American Journal of Sociology. 1973; 78: 1360–1380. doi: 10.1086/225469

54. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, et al. Structure and tie strengths in mobile communication networks. Proc Nat Acad Sci (PNAS). 2007; 104: 7332–7336. doi: 10.1073/pnas.0610245104

55. Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. Can cascades be predicted? In: Proc. Intl. Conf. on World Wide Web (WWW); 2014. pp. 925–936.

56. Szabo G, Huberman BA. Predicting the popularity of online content. Communications of the ACM. 2010; 53: 80–88. doi: 10.1145/1787234.1787254