Scalable Detection of Viral Memes from Diffusion Patterns

Pik-Mai Hui, Lilian Weng, Alireza Sahami, Yong-Yeol Ahn, and Filippo Menczer

Abstract Social media and social networking platforms have flourished with the rapid development of mobile technology and the ubiquitous use of the Internet. As a result, *memes*, or pieces of information spreading from person to person, can be reshared among users quickly and gain huge popularity. As viral memes have tremendous social and economic impact, detecting these viral memes at their early stages of spread is a worthy, yet challenging problem. Here we review the literature on predicting viral memes, and present empirical results from Twitter and Tumblr datasets. We demonstrate how diffusion patterns of memes, in the context of network communities, play an important role in predicting virality. We show that it is feasible to obtain predictive features based on community structure even at the massive scales that common social media services need to process. Our results may not only enable practitioners to make predictions about meme diffusion, but also help researchers understand how and why different factors, in particular diffusion patterns in communities, affect online virality.

1 Introduction

A *meme* is a distinct piece of information that replicates among people, like biological genes replicating through reproduction [1]. Memes resemble infectious diseases, in the sense that both travel through social ties between people [2, 3]. As blooming online social media services facilitate online social interactions, they also change how memes spread through society. Most importantly, social media platforms such

Pik-Mai Hui*, Lilian Weng, Yong-Yeol Ahn, and Filippo Menczer
Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, IN, USA
*e-mail: huip@indiana.edu

Alireza Sahami
Yahoo! 701 1st Ave, Sunnyvale, CA, USA

as Facebook, Google Plus, Twitter, and Tumblr connect billions of users into a network that can spread a meme to the whole world instantly. At the same time, these services allow us to directly observe and study the spreading of memes and user behaviors by recording detailed data about user activities.

A vast number of memes is created every day. However, only a tiny fraction goes viral. This raises the most fundamental question in information diffusion research: *what makes something viral?* This question has attracted attention across disciplines including marketing and advertisement, as well as machine learning and network science. One shall agree that the question is meaningful but too broad. Here we focus on a more specific and well-defined question: *How can we predict the virality of a meme early?*

There are roughly two general approaches to the problem of meme virality prediction: time series analysis and feature-based classification. What follows in this chapter focuses on feature-based classification [4, 5, 6]. Readers who are interested in the approach of time series analysis are referred to a different literature [7, 8, 9, 10]. The feature-based classification approach aims to discover distinguishing features of viral memes and to apply supervised machine learning techniques using these features. As in standard feature-based machine learning problems, a general saying is *garbage in, garbage out,* implying that if inputs to a model are not informative, its output will neither be meaningful. Therefore the most critical step is to identify and extract useful features from datasets at hand.

We study a set of useful features from our theoretical and conceptual understanding of network structure and social information diffusion processes. In particular, we discuss the features of the diffusion patterns based on dense subgroups (communities) in underlying networks. We will demonstrate that diffusion pattern can be extracted at scale, which preserves its strength in virality prediction in two massive datasets from Twitter and Tumblr.

2 What Makes It Viral?

Although we do not address this question directly, understanding the potential reasons why memes go viral is nevertheless crucial for identifying useful features and for any discussion about viral memes. From literature we identify three key aspects of viral spreading, namely innate attractiveness of memes, user characteristics, and properties of the underlying social network. Motivated readers are recommended to query the references for more details on these aspects of virality.

2.1 Innate Attractiveness

The innate appeal of a meme may be the most basic factor contributing to its virality. It is intuitive that users are more likely to reshare memes with better “quality”. Quality can be defined in different contexts. For example, Berger and Milkman studied the emotional constituents in news articles and their impact on the articles' virality. They find that news articles that actively evoke arousal become more viral later on [11]. Many studies presuppose virality as an intrinsic trait of memes. Since a meme is represented by its content, it justifies the search for content features that correlate with quality. For one, Guerini *et al.* characterized various aspects of virality and how they indicate the future virality of text-based content [12].

Although innate attractiveness is an intuitive explanation of virality, it does not paint the whole picture. The attractiveness of a meme is highly dependent on many contextual features, such as other existing memes and the culture of surrounding population. Studies have also demonstrated that quality alone does not explain virality well. In fact, agent-based simulation showed that highly skewed distribution of meme popularity can arise even if we do not assume any difference in innate quality of memes [13]. Moreover, the success of online content, such as songs from online music downloads and social news filtering, depends significantly on provided social cues [14, 15]. This suggests that factors other than innate quality, such as visibility and reachability of the memes, may as well contribute to virality.

2.2 User Characteristics

The importance of social influence leads us to the concept of influencers and the roles of user characteristics in general. Although there are seemingly countless memes available, the scarcity of user effort in consuming information leads to limited individual *attention* in any social networks. Similar to biological organisms (and genes) striving for resources to reproduce, all memes strive for the attention of people. Since user consumes meme at a limited rate, only the memes that are seen within a short time period have a chance to propagate. Memes originating at an isolated location in the social network may not have any chance to spread because no one can see them in the first place. Such memes quickly go extinct in the system. Meanwhile, a meme that happens to be reshared by a user with many followers will have a significantly higher chance to reproduce across the followers' minds.

When user B reads user A's post, the likelihood of user B resharing the information depends on his/her evaluation of user A. That is, the influence that one exerts on others varies across the actors. Content by a well-respected celebrity such as a founder of a famous organization naturally generates a stronger influence on others than that by a normal person, despite that they are two copies of the same content. In addition, each user has a specific set of topical interests. Some care more about global politics and wars in the Middle-East, while others may only want to know about new French recipes. Since users consume and share information according to their own interests, it is more likely for meme to spread between users with similar interests, when one shares and one consumes closely relevant contents. These effects are further exacerbated by a combination of limited user attention and abundant supply of memes. Weng *et al.* showed that limited individual attention in the competition among memes induces strong heterogeneity in meme popularity and longevity [13]. In deciding which meme to consume, each user prioritizes based on their interests and this alters meme popularity [16].

In other words, the spreading of viral memes favors users of specific characteristics. We call them influential users. Many methods have been proposed for quantifying user influence and identifying these influential users. In general, these methods use relevent observables of user characteristics, such as high degree or retweetability [17, 18], topical similarity [19, 20, 21], information forwarding activity [22, 18], or size of cascades [23, 24], to infer the strength of user influence over other.

2.3 Properties of Underlying Social Network

The characteristics of social ties in the underlying social network, through which memes spread, also affect the success of memes. Strong and homophilous ties are considered more effective than weak ties for spreading messages [25], while weak ties are thought of as transmitting novel information [26]. These theories are commonly used in viral marketing and consumer studies, where researchers actively apply network approaches to analyze and model local and global patterns of social network structure [27, 28, 29]. In addition, the existence of hubs, namely nodes with extremely large degree, is known to affect the persistence of infections, the distribution of cascade sizes, and the vulnerability of the system [30, 31]. Intuitively, hubs provide pathways through which memes can teleport to distant parts of the network instantly, facilitating the development of meme popularity on the whole network.

Another important network structure feature in most social networks is the presence of dense subgraphs called *communities* [32, 33, 34, 35]. Communities are characterized by internal cohesion (more internal edges than expected) and external isolation (fewer outgoing edges than expected). While communities naturally contraints information flow across their borders, they may be necessary for providing initial critical mass before a meme can spread broadly [36]. In addition, the theory of complex contagion [37, 38, 39, 40, 41] suggests that we may expect an even stronger constraining effect from community structure [4]. Therefore, information extracted from the network structure and early spreading patterns is valuable to predict the virality of a meme. Further discussion on extracting features from community structures of social networks follows in a later section.

3 Data and Methods

In this section we present details of the datasets used in our experiments, and explain the methods we applied to extract network communities and to predict virality. We begin with a brief introduction to the online social media platforms from which our data was collected, and the networks that we constructed using each of these platforms.

3.1 Social Media Platforms

Online social media platforms enable people to share information and subscribe to updates from other users. The information can be of any type, ranging from short text messages and blog posts, to images and video clips. On these platforms, users typically choose others to whom they pay attention by 'following' them. Most platforms also provide users with multiple mechanisms of information sharing, which serve different purposes.

Twitter is one of the most popular social media platforms. On Twitter, users post short messages called *tweets*. Between a pair of users (u, v), we consider three main types of interactions: (i) u can *follow* v to subscribe to tweets from v; (ii) u can *retweet* v's messages to re-broadcast them to u's followers; and (iii) u can *mention* v's screen name in tweets by using the "@" symbol (e.g., '@potus'). Users can also explicitly attach indexable topic identifiers to a tweet by using *hashtags*, terms with the "#" symbol as a prefix (e.g., `#news`).

Tumblr is another popular social networking and microblogging platform supported by Yahoo! since 2013, hosting hundreds of millions of monthly active users and blogs. Tumblr features many functions similar to Twitter, such as hashtags, resharing, liking, and replies.

On both Twitter and Tumblr, hashtags can be used to operationalize the concept of memes, thanks to multiple characteristics of hashtags that accord with the definition of a meme [1]. First, hashtags are concretely defined by user consensus and uniquely identifable through searches; second, hashtags reproduce through imitation by users; third, hashtags mutate, compete, and dominate in the same system over time. For example, `#ows` rapidly suppressed several similar hashtags to become the reference label for the Occupy Wall Street movement among hundreds of thousands of people who participated in related public discourse [42]. The usage of hashtags also makes the application of our methods straightforward and our findings easily comparable to results based on other platforms.

3.2 Community Detection

Communities contain rich information about the structure of a social network. These communities can be extracted by applying different algorithms. The results in this chapter are based on communities detected by two methods, namely InfoMap [33] and Louvain's method [43]. We have chosen these two methods, based on contrasting principles, to evaluate the robustness of the results under different choices of community detection algorithm. InfoMap and Louvain's method optimize for different objective functions and are therefore expected to produce distinct results, particularly regarding community size and resolution [44]. Another difference is that InfoMap considers the direction of edges, while Louvain's method treats all edges as undirected. Therefore the results may provide insight about the usefulness of edge directionality as signals for virality prediction.

Nowadays, the sizes of online social networks and the volume of information traffic on them is so large that analysis requires distributed storage and computing environments. Algorithms running on single computers do not scale well to such large networks, say with tens of million nodes. Additionally, moving large volumes of data stored on different storage nodes to a single machine is costly. Although the original implementations of the InfoMap and Louvain's algorithms were not designed for parallel computation, distributed implementations of these algorithms have been developed to better utilize resources in multiple-machine clusters. These scalable methods optimize execution speed and resource efficiency without sacrificing accuracy. We use *distributed Louvain* [45] and *RelaxMap* [46], parallel implementations of Louvain's and InfoMap methods respectively, to extract communities from large Twitter and Tumblr networks.

3.3 Twitter Information-Sharing Network

In prior work, virality was predicted using community features extracted from a Twitter follower network [5]. While constructing such a follower network is desirable, it poses some challenges. Some social media platforms, such as Facebook, regard friendship data as private, and therefore do not make it available for research. Furthermore. collecting complete follower information among many users can be forbiddingly expensive. The APIs provided by popular online social platforms restrict the rate at which such data can be queried without payment, making even moderate-size experiment difficult. This motivates an alternative approach.

We can extract communities based on an information-sharing network rather than a follower network. The links in such a network represent how memes spread through, e.g., retweets and replies. This can be used as a proxy for the social network that captures the process of meme diffusion. Since people typically retweet messages from users they follow, an information-sharing network has a significant overlap with the follower network. Let us consider two networks constructed in this fashion, using high-volume streams of Twitter and Tumblr posts.

In our experiment, the Twitter information-sharing network is constructed using a 10% sample of public tweet stream. The tweets used in our study are from July to September 2015 (Table 1). We divide the collected tweets into two temporal parts: a one-month *observation period* followed by a two-month *experiment period*.

In the observation period we collect existing hashtags and information-sharing activities. These activities are used to construct a directed information-sharing network. Each edge in the network is formed by retweets and mentions of one user by another, and is weighted by the frequency of information flow from source to destination user. When user A is retweeted by user B, or when user A mentions user B, information flows from A to B. Communities in this network are extracted by RelaxMap and Louvain's algorithms. To reduce noise, only the largest weakly-connected component of network is used in community detection.

In the experiment period, we consider only newly-born hashtags, which did not occur in the observation period. Each new hashtags is tracked for a period of thirty days, starting from its first occurrence. If a hashtag first occurs within thirty days of the end of the experiment period, so that we do not have thirty days of data in the experiment period, we do not consider it in our study. For each tracked hashtag, we record the sequence of users who share it (adopters).

This setup has some desirable properties. Since the networks are constructed using only information from the observation period and evaluation is done strictly over content in the experiment period, there is no information leak between training and evaluation. Moreover, every hashtag in the evaluation is observed for exactly thirty days after its first use, avoiding a bias against late hashtags.

In summary, tweets from the observation period are used to construct the directed network from which communities are extracted. The experiment period is used to construct meme adoption histories and run the prediction experiments.

3.4 Tumblr Information-Sharing Network

We also collected posts from the Tumblr firehose, a database with the complete history of user posts. On Tumblr, a user can create and own multiple blogs with one account. Tumblr identifies the same user posting in distinct blogs as different *persona*. However, each user is identified by one primary blog while reacting to posts from other users, such as when replying and liking posts. Therefore we consider a user's primary blog as their identity. We focus on text posts, excluding other types of content such as pictures and video clips.

We divided this dataset the same way we did for the Twitter network (Table 1). A directed network is constructed by scanning all text posts in November 2015 (the observation period), and its largest weekly connected component is used to extract communities. An edge is generated when a user likes or replies to a post by another user, and edges are weighted by the frequencies of interaction. Edges are directed from user A to user B when B likes or replies to posts by A. Text posts in December 2015 and January 2016 (the experiment period) were collected to run the predictions.

The Tumblr dataset contains a very diverse set of hashtags. Tumblr hashtags are case sensitive, can contain spaces and emoji, and have no length limit. As a result, they can be very long (full sentences) and have duplication, for instance "Cute cat" and "cute_cat." To limit the noise caused from these degenerate cases, we filtered out hashtags that are longer than twenty characters and trimmed all emoji, common phrase separators (space, underscore, etc.) and repeated expressions, then lower-cased all characters.

	Twitter	Tumblr
Type of edge	Retweets & mentions	Replies & likes
Observation period	2015-07	2015-11
Experiment period	2015-08/09	2015-12/2016-01
# Nodes	29,224,842	19,701,097
# Edges	169,685,133	711,573,645

Table 1 Information and basic statistics about the network datasets in the study.

4 Network Community Features

In this section we present the features extracted from the networks. The features are a subset of the ones used in our prior work [5]. In particular, we focus on features that are motivated by the community structure of the underlying social networks. These network features are computed based on the locations of the first n adopters of each hashtag, where the parameter n is set to be a relatively small number compared to the final number of tweets generated by viral hashtags. In our experiment, $n = 25$.

Let us start by defining a few key concepts and mathematical notations. Some of the information is mentioned in previous sections, but is included below for the sake of completeness.

Definition 1. Meme: We consider each hashtag h as a meme. The popularity of meme h is quantified by the number of adopters. $A(h)$ is the set of all adopters who posted about h and $A_n(h) \subseteq A(h)$ is the set of early adopters who posted at least one of the first n posts. We define the *popularity* of h as $|A(h)|$.

Definition 2. Adopter sequence: For a given meme h, we consider the sequence of meme adopters, $\langle a_1^h, a_2^h, \ldots, a_n^h \rangle$, where $a_i^h \in A(h)$ is the creator of the i-th post containing h. A user may appear multiple times in the sequence if the user posts about h more than once.

Definition 3. Community: A community $c \in \mathscr{C}$ is a dense subgraph of nodes (users) in the network. Given information about which nodes belong to which communities, $A(h|c)$ is the set of adopters of a meme h in community c. $A_n(h|c)$ is the similar set that only considers the first n relevant tweets. $C(h)$ denotes the *infected communities* of h, which are communities with at least one tweet containing h. Similarly, the infected communities with early posts are $C_n(h)$.

Community structure is useful in predicting meme virality because of how memes travel among users who are socially connected. This process is commonly called *social contagion*. It has been argued that social contagions are *complex* contagions, in contrast to *simple* contagions like epidemic spreading. To explain the connection between complex contagion and community structure in the context of social network analysis, we note that complex contagion is known to possess two distinctive characteristics:

Social reinforcement. Until a certain point, each additional exposure drastically increases the probability of adoption [47, 48, 49].

Homophily. Social relationships are more likely to be formed between people who share certain characteristics, captured in the sayings "birds of a feather flock together" and "similarity breeds connection" [50, 51].

Community structure has been shown to help quantify the strength of both social reinforcement and homophily by the following mechanisms [4]. First, dense connectivity inside a community increases the chances of multiple exposures, thus enhancing the contagion that is sensitive to social reinforcement. Second, groups with similar tastes naturally establish more edges among them, forming communities. Therefore members of the same community are more likely to share similar interests. We thus expect that, if these two effects are strong, communities will facilitate the internal circulation of memes while preventing diffusion across communities, causing strong concentration or low community diversity.

Unpopular memes tend to be concentrated in a small number of communities, while a few viral memes have high community diversity, spreading widely across communities like epidemic outbreaks [4]. This can be explained by trapping of information flow in communities. Viral memes are able to breach the borders of communities and out-survive other memes. Therefore, features that quantify the community diversity should help predict future meme virality. As an illustration, Fig. 1 is a visualization of the early diffusion patterns of a few memes based on the first 30 tweets, `#TheWorseFeeling` and `#IAdmit` clearly exhibit more community diversity than non-viral memes, e.g. `#ProperBand` and `#FollowFool`.

Based on the above analysis, we define a key feature of diffusion patterns based on community structure as follows:

Definition 4. Adopter entropy, $H_n^A(h)$. The measurement of entropy describes how adopters of a given meme are scattered or concentrated across communities. Large entropy indicates low concentration or high diffusion diversity:

$$H_n^A(h) = -\sum_{c \in C(h)} \frac{|A_n(h|c)|}{|A_n(h)|} \log \frac{|A_n(h|c)|}{|A_n(h)|}.$$

5 Experiment

Let us present the details of our experiment on virality prediction using the diffusion features extracted from the network community structure. We first define a virality prediction task. We will show that diffusion diversity is a strong predictor of virality.

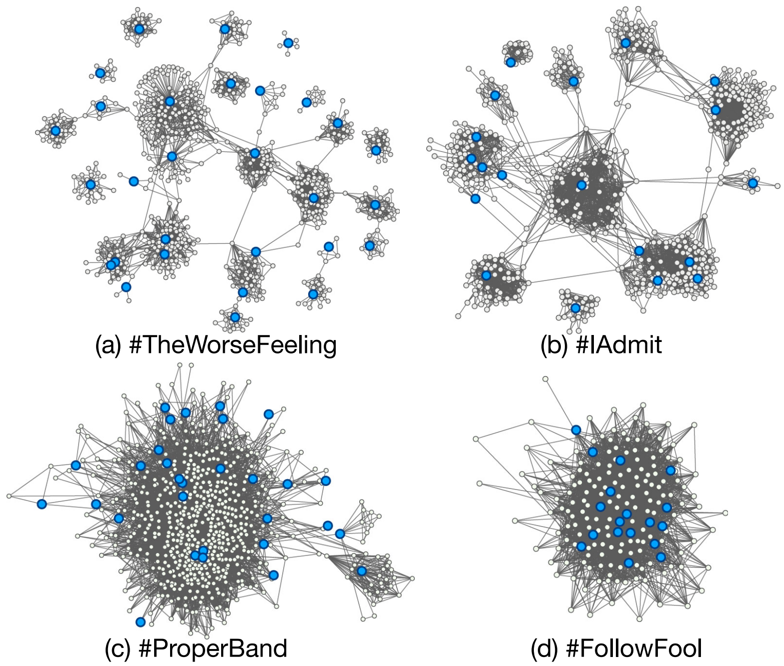

Fig. 1 Visualizations of diffusion patterns of viral (a,b) and non-viral (c,d) memes on Twitter. Early adopters among the first 30 tweets (in blue) and their neighbors in the same communities are shown. Each node represents a user and each link indicates the reciprocal follow relationship between two users. Figure reproduced with permission [5].

5.1 Task Specification

Each new hashtag is associated with a series of adopters within the experiment period. We only compute features using the positions of the first $n = 25$ adopters in the network. Our method therefore requires that a new hashtag has been used at least 25 times within the experiment period.

Meme popularity exhibits a broad and skewed distribution, as observed in many previous studies [52, 13]. Our key questions are whether the diffusion diversity feature based on community structure provides a predictive signal, and whether this signal is informative at the large scales of our information-sharing networks. The following recipe defines a meme virality prediction task:

1. Each hashtag is given either viral (1) or not (0) as its ground-truth class; the most-frequent 50% of the inspected hashtags within a month of usage are defined as viral.
2. All hashtags are ranked by adopter entropy $H_n^A(h)$, from the highest to the lowest.
3. The top 50% of hashtags based on the ranking in step 2 are predicted as viral.

	Twitter	Tumblr
RelaxMap	0.67	0.60
D-Louvain	0.68	0.60

Table 2 Prediction accuracy (AUC) from evaluation on each of the datasets.

4. Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) are used to evaluate prediction accuracy.

We note that this balanced binary classification task is simpler than the more realistic scenario in which only a small fraction of memes go viral.

ROC curves are drawn by first ranking the scores of the hashtags, then evaluating each sample point as a true positive or false positive in the ranked order. If the true positive data points are among the top ranks, the curve will bounce up, hence the AUC will be close to one. On the other hand, if false positive sample points are ranked high, the AUC will be close to zero. A random ranking will spread true and false positives evenly, and therefore yield an AUC close to 0.5.

5.2 Evaluation

The ROC curves in Figure 2 and AUC values in Table 2 show that community entropy of adopters $H_n^A(h)$ alone provides a useful signal in predicting which memes will go viral in large-scale social media. The AUC values around 0.7 and 0.6 for Twitter and Tumblr networks, respectively, represent significant improvements upon the random baseline. Naturally, the results could be improved further by combining entropy with other features in the literature [5, 6].

The RelaxMap and distributed Louvain's methods perform similarly on the same data. Recall that Louvain's method ignores the direction of edges, while RelaxMap does not — InfoMap is based on directed random walks. To investigate the contribution of edge directionality, we ran RelaxMap on an undirected version of the Twitter information-sharing network. This was done by adding weights for reciprocal edges, similarly to the way this is done in the distributed implementation of Louvain's method. The resulting AUC is not significantly different from the random baseline. This suggests that RelaxMap makes use of both weights and directionality of the edges while extracting communities, and this affects the signal we use for virality prediction.

The diffusion patterns are informative in the prediction task on both Twitter and Tumblr platforms. Despite the simplicity of the task, the results of our evaluation demonstrate that for meme virality prediction, diffusion patterns are robust against source platforms and network construction, and scale up to very large networks.

Compared to Twitter, virality prediction in Tumblr seems to be much more challenging. The difficulty may be attributed to different ways in which the platform is used and the data is collected. First, hashtags on Tumblr tend to be used dif-

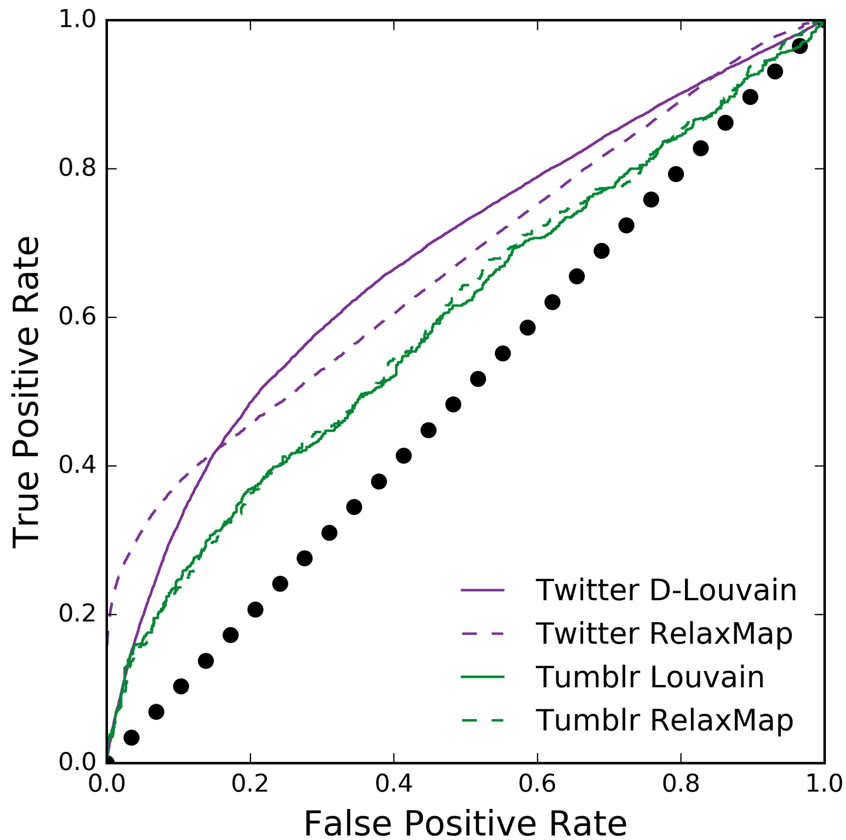

Fig. 2 A plot of ROC curves using diffusion diversity (adopter entropy) as the ranking criterion. Different curves correspond to different information-sharing networks and community detection algorithms.

ferently due to the lack of strong limitations on the set of characters. People use hashtags with more characters and diverse types of expression styles, such as irony and sarcasm. As the possible space of hashtags grows, it becomes less clear if the assumption of hashtags as proxies of memes is appropriate. Further, unlike Twitter, Tumblr encourages users to create blog posts without length limitation, giving rise to distinct meme consumption and diffusion patterns.

Another potential difference between the two platforms is the sampling of posts in the Twitter stream, which is biased toward active users who are responsible for most of the tweets. The Tumblr firehose includes barely active and less predictable users.

6 Conclusion

In this chapter, we explore the question of virality of online content and its prediction on large social media platforms. We summarize three perspectives on driving factors of virality — innate attractiveness of the content, user characteristics, and the network structure of the underlying social network. We present a simple, yet effective community feature that captures the diffusion patterns of memes in the network. We show that the communities, from which the entropy feature is derived, can be extracted in large-scale information-sharing networks such as Twitter and Tumblr. We also find that diffusion diversity provides a predictive signal across platforms.

There are multiple future directions for this line of research. A noteworthy challenge in deploying the methods in any real-time system is the computational complexity of updating the required features as the social network evolves. Although community structures can be assumed to be fairly stable over time, it is unclear for how long this assumption of static network holds. Consensus clustering [53] could be applied to explore this question.

Another potential direction is to investigate the effect of groups with different characteristics, for instance cultures, religions, and genders, on meme consumption. There has been little work on feature-based models that are aware of group-level characteristics. One can imagine that a meme will gain attention in a particular group while being ignored in others. If early adopters of the meme are in relevant groups of users who are motivated to share it, the meme is more likely to go viral. Such content-aware approach, accompanied with powerful community features, may lead to the development of more powerful prediction algorithms.

References

1. Richard Dawkins. *The Selfish Gene*. Oxford Univ. Press, 1989.
2. Daryl J Daley and David G Kendall. Epidemics and rumours. *Nature*, 204(4963):1118–1118, 1964.
3. William Goffman and Vaun A. Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204:225–228, 1964.
4. Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific Reports*, 3(2522), 2013.
5. Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Predicting successful memes using network and community structure. In *Proc. Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.
6. Justin Cheng, Lada Adamic, Alex Dow, Jon Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proc. Intl. World Wide Web Conf. (WWW)*, 2014.
7. Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proc. Intl. Conf. on web information systems and mining (WISM)*, pages 32–38, 2009.
8. Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, and Chunyan Wang. Trends in social media: Persistence and decay. In *Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2011.
9. Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 177–186, 2011.

10. Stanislav Nikolov. Trend or no trend: A novel nonparametric method for classifying time series. Technical report, MIT, 2012.
11. Jonah Berger and Katherine L. Milkman. What makes online content viral? *J. Marketing Research*, 49(2):192–205, 2009.
12. Marco Guerini, Carlo Strapparava, and Gözde Özbal. Exploring text virality in social networks. In *Proc. AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 506–509, 2011.
13. Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2(335), 2012.
14. Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.
15. Lev Muchnik, Sinan Aral, and Sean J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.
16. Lei Yang, Tao Sun, and Qiaozhu Mei. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *Proc. Intl. World Wide Web Conf. (WWW)*, pages 261–270, 2012.
17. M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, pages 10–17, 2010.
18. Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proc. IEEE Intl. Conf. on Social Computing*, pages 177–184, 2010.
19. Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proc. ACM Intl. Conf. on Knowledge discovery and data mining (KDD)*, 2009.
20. Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, 2010.
21. Lilian Weng and Filippo Menczer. Topicality and impact in social media: diverse messages, focused messengers. *PloS one*, 10(2):e0118410, 2015.
22. Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *Proc. Intl. World Wide Web Conf. (Companion Volume)*, pages 113–114, 2011.
23. Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernan A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.
24. E. Bakshy, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proc. ACM Intl. Conf. on Web search and data mining (WSDM)*, pages 65–74, 2011.
25. Jacqueline Brown and Peter Reingen. Social ties and word-of-mouth referral behavior. *J. Consumer Research*, 14(3):350–362, 1987.
26. MS Granovetter. The strength of weak ties. *Am. J. Sociology*, 78(6):1360–1380, 1973.
27. Jure Leskovec, Lada Adamic, and Bernardo Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.
28. W.A. Mason, A. Jones, and R.L. Goldstone. Propagation of innovations in networked groups. *J. Experimental Psychology: General*, 137(3):422, 2008.
29. Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Sci.*, 57(9):1623–1639, 2011.
30. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.
31. Duncan J. Watts. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci.*, 99(9):5766–5771, 2002.
32. M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103(23):8577–8582, 2006.
33. Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.*, 105(4):1118–1123, January 2008.
34. Yong-Yeol Ahn, James Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

35. S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
36. Azadeh Nematzadeh, Emilio Ferrara, Alessandro Flammini, and Yong-Yeol Ahn. Optimal network modularity for information diffusion. *Physical review letters*, 113(8):088701, 2014.
37. Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.
38. Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.
39. Damon Centola and Michael Macy. Complex contagions and the weakness of long ties 1. *American journal of Sociology*, 113(3):702–734, 2007.
40. Damon Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.
41. Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
42. Michael D. Conover, E Ferrara, Filippo Menczer, and Alessandro Flammini. The digital evolution of occupy wall street. *PLOS ONE*, 8(3):e55957, 2013.
43. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
44. Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.
45. Sotera. spark-distributed-louvain-modularity. https://github.com/Sotera/spark-distributed-louvain-modularity.
46. Seung-Hee Bae, Daniel Halperin, Jevin West, Martin Rosvall, and Bill Howe. Scalable flow-based community detection for large-scale network analysis. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 303–310. IEEE, 2013.
47. Eytan Bakshy, Brian Karrer, and Lada Adamic. Social influence and the diffusion of user-created content. In *Proc. ACM Conf. on Electronic Commerce*, pages 325–334, 2009.
48. Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc. Intl. World Wide Web Conf. (WWW)*, 2011.
49. Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.
50. Miller McPherson, Lynn Lovin, and James Cook. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociology*, 27(1):415–444, 2001.
51. Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.
52. Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proc. Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, pages 90–97, 2010.
53. Andrea Lancichinetti and Santo Fortunato. Consensus clustering in complex networks. *Scientific reports*, 2:336, 2012.