

# A Survey of Information Cascade Analysis: Models, Predictions and Recent Advances

FAN ZHOU, University of Electronic Science and Technology of China, China

XOVEE XU, University of Electronic Science and Technology of China, China

GOCE TRAJCEVSKI, Iowa State University, USA

KUNPENG ZHANG, University of Maryland, USA

The deluge of digital information in our daily life – from user-generated content such as microblogs and scientific papers, to online business such as viral marketing and advertising – offers unprecedented opportunities to explore and exploit the trajectories and structures of the evolution of information cascades. Abundant research efforts, both academic and industrial, have aimed to reach a better understanding of the mechanisms driving the spread of information and quantifying the outcome of information diffusion. This article presents a comprehensive review and categorization of information popularity prediction methods, from *feature engineering and stochastic processes*, through *graph representation*, to *deep learning-based approaches*. Specifically, we first formally define different types of information cascades and summarize the perspectives of existing studies. We then present a taxonomy that categorizes existing works into the aforementioned three main groups as well as the main subclasses in each group, and we systematically review cutting-edge research work. Finally, we summarize the pros and cons of existing research efforts and outline the open challenges and opportunities in this field.

CCS Concepts: • **Applied computing** → **Sociology**; • **Information systems** → **Data mining**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

Additional Key Words and Phrases: popularity prediction, information diffusion, social network analysis

## ACM Reference Format:

Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2020. A Survey of Information Cascade Analysis: Models, Predictions and Recent Advances. *ACM Comput. Surv.* 1, 1, Article 111 (May 2020), 41 pages. <https://doi.org/10.xxxx/xxxxxxx.xxxxxxx>

## 1 INTRODUCTION

The rapid development of wireless communication technologies and the Internet and miniaturization and availability of mobile devices have dramatically changed the way people obtain data and information and use it when interacting with each other. Understanding how information is spread – along with what kind of factors drive the success of information diffusion and making predictions about the population size that information can affect – is a challenging but critical necessity question for many real-world application domains, e.g., viral marketing [114], advertising [89], scientific impact quantification [195], recommendation [209], campaign strategy [22], and epidemic prevention [241].

Authors' addresses: Fan Zhou, fan.zhou@uestc.edu.cn, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, 610054; Xovee Xu, xovee@live.com, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China, 610054; Goce Trajcevski, gocet25@iastate.edu, Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA, 50011; Kunpeng Zhang, kpzhang@umd.edu, Department of Decision, Operations and Information Technologies, University of Maryland, College Park, MD, USA, 20742.

© 2020 Association for Computing Machinery.

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Computing Surveys*, <https://doi.org/10.xxxx/xxxxxxx.xxxxxxx>.

The trajectories and structures of information diffusion, as well as the adopters/participants in information spreading, form the so-called *information cascades*. Many researches and enterprises have put much effort into modeling and learning information diffusion in cellular networks, online social networks, paper citation networks, and content sharing networks, among many others [2, 8, 193, 197, 240] demonstrating that the ability to predict information items or cascades is of interest from both academic and business perspectives. The word *prediction* could refer to different meanings in different applications, e.g., it could be predicting the popularity of tweets/hashtags in microblogs [33, 96, 141, 242], the number of ‘likes’ for a photo/video in Facebook [38], views of videos in YouTube [39, 74, 211], ratings of movies in IMDB [150], citations of academic papers in Google Scholar or DBLP [172, 195], votes on stories in Digg [113], comments on news articles [182, 187], or social influence [157].

In this work, we provide a comprehensive review of existing information cascade prediction methods, which is a challenging task both because of the huge amount of publications in this field as well as the lack of uniform standards to classify the existing works. To begin with, the specification of the prediction problem itself varies in different works. According to the *problem formulation* in different applications, prediction may refer to either binary/multi-class classification or regression. For example, it could be predicting the exact size of a cascade at a future moment [3, 11, 106, 179] or just estimating whether a cascade would grow beyond a threshold [38, 45]. Moreover, *strategically*, based on the range of observed information, the prediction can be made prior to the publication [135] or by peeking into the early cascade evolution [173]. From the perspective of *analytical levels*, popularity prediction can be made by focusing on different diffusion levels. For example, *macro-level* models [119] learn the collective behavior of cascades, while *micro-level* models [157, 217] focus more on individual user actions/responses to specific information items.

From the perspective of *methodology*, there are plentiful choices of algorithmic approaches for modeling and predicting information cascades as well as data types associated with the cascades. Traditionally, various features (e.g., temporal and structural) associated with information items can be extracted with feature engineering – while typical machine learning models (e.g., linear/logistic regressions, naïve Bayes, SVM and decision trees) or stochastic process models (e.g., Poisson and Hawkes point processes) are readily applicable for modeling and predicting information popularity. With the recent advances in deep neural networks, especially the techniques for learning graph representation (e.g., DeepWalk [153], node2vec [68], and graph convolutional network (GCN) [94, 238]) and sequential data (e.g., recurrent neural network (RNN) and its many variants [41, 79]), various deep learning-based information diffusion models have emerged [38, 119, 242].

Several works have reviewed the information diffusion models in recent years [62, 70, 143, 183]. Earlier works [70, 183] mainly focus on various feature engineering approaches and classical machine learning methods for modeling and predicting the popularity of information items. More recent surveys [62, 143] concentrate on web content and/or microblog information diffusion, while emphasizing different aspects of information cascade modeling. For example, [143] focuses on web content popularity prediction and review *a priori* and *a posteriori* prediction approaches, as well as evaluation protocols and classification/regression methods. However, the authors also give an overview of feature-based approaches and time series modeling methods. Similarly, [62] focuses on microblogs – more specifically, Twitter and Weibo information popularity prediction – while prioritizing generative models that rely on modeling time series events with stochastic processes.

In comparison with previous surveys, this article makes the following distinct contributions:

- (1) A more general perspective: we extend the realm of modeling online web content, e.g., user-generated content microblogs [62], to a general definition of information, i.e., *any* measurable

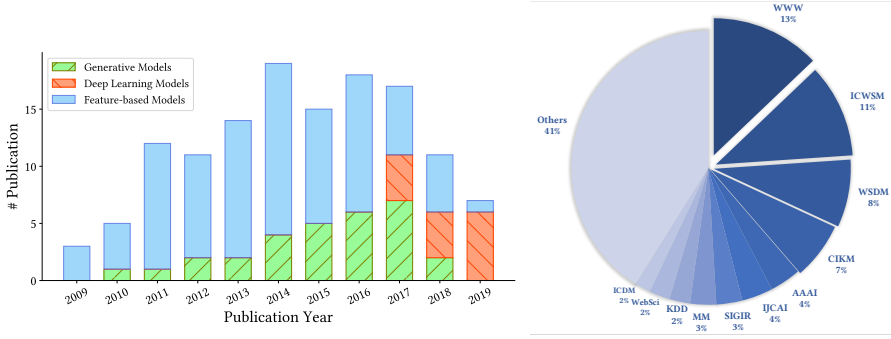


Fig. 1. Left: Number of publications over the last decade. Right: Distribution of venues of publications.

entities that can be propagated in *any* networks, including scientific publications and citation networks, user published text/photos/videos and online sharing networks, etc.

- (2) A wider range of networks and cascades: as opposed to existing surveys that focus on a single type of social network (e.g., Twitter and Weibo in [62]), this work reviews information spreading in various networks, including but not limited to Facebook, Flickr, YouTube, Netflix, IMDB, Instagram, Wikipedia, Reddit, and DBLP.
- (3) A broader (and balanced, in-depth) survey of the methodologies for predicting the outcome/popularity of cascades: complementing the previous surveys' composite of methods at different analytical levels [70, 143, 183], ours provides finer-grained analysis of trade-offs, advantages and limitations of existing methods. We also consider a wider spectrum of features, methodologies and interpretations, compared to a recent survey that also focused on popularity prediction [62].
- (4) More comprehensive recent literature: we currently provide the most comprehensive literature review of methods of modeling information diffusion, which not only spans a broad range of traditional feature engineering approaches and generative models, but also contains recent advances in modeling and predicting information popularity with graph representation learning and deep learning techniques.

Fig. 1 shows summary statistics of the papers included in this survey. Clearly, information cascade modeling and prediction have consistently spurred great interest in recent decades, although the methodologies employed have changed over time, e.g., deep learning-based methods have emerged as the more popular techniques in recent years. Research works that we reviewed were published in high-ranking conference proceedings/journals related to data mining, machine learning, social networks and information management, such as WWW, ICWSM, KDD, CIKM, WSDM, SIGIR, MM, TKDE, TKDD, ICDM, AAAI, and IJCAI. We note, though, that more than 40% of papers come from a variety of conference proceedings/journals outside of the typical venues, further attesting that this is an interdisciplinary topic spanning computer science, artificial intelligence, sociology, economics, marketing, statistics, etc.

We build a taxonomy of methods from three different aspects, as illustrated in Fig. 2. *First*, popularity prediction can be considered as either a *classification* problem or a *regression* problem, according to the problem formulation and target application tasks. For example, information outbreak prediction and users' joint behavior prediction perform the classification tasks, while predicting the exact volume of items/cascades at a specific future time requires regression learning. *Second*, prediction can be made before or after the publication of the information. According to the prediction time, existing methods can be categorized as *ex-ante* prediction or prediction after

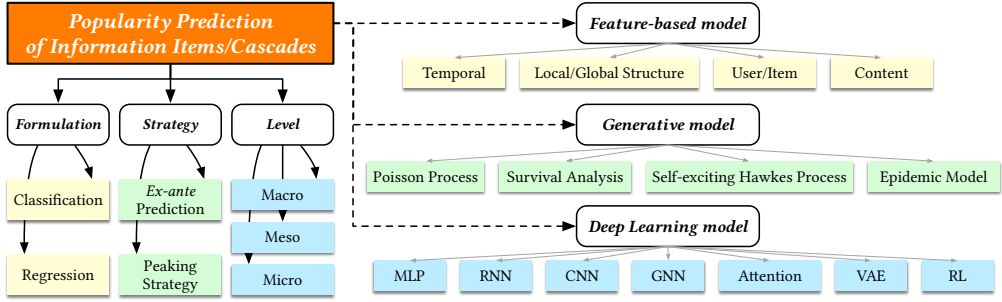


Fig. 2. Taxonomy of the information diffusion models.

peeking into the early stage of information cascades. *Third*, information prediction works can be further classified as *micro-level*, *macro-level* and *meso-level* predictions by the granularity of the tasks. *Methodologically*, existing works are grouped into three categories based on the methods modeling information cascades, including feature-based methods, generative models and deep learning approaches.

The rest of this survey is structured according to the hierarchies of the method taxonomy. In [Section 2](#), we introduce the basic formulations, problem definitions, and the commonly used evaluation protocols and benchmark datasets. In [Section 3](#), we illustrate basic characteristics of information cascades and review corresponding feature-based models. In [Section 4](#), we review another major group of models relying on time series of cascades and stochastic processes. [Section 5](#) summarizes recent advances in modeling information cascades with deep learning-based techniques. We discuss the open challenges and opportunities and conclude this survey in [Section 6](#).

## 2 PROBLEM DEFINITION, EVALUATION, DATASETS AND TAXONOMY

Modeling information items/cascades has long been an active field in both academia and industry [161, 183]. Due to the natural diversity and broad scope of information, existing works vary in their item/cascade definitions. In this section, we first review the frequently studied information items and cascades – examples ranging from traditional news articles and academic papers to newly emerged microblogs and online photos/videos. Then we group the problems into three categories: (i) predictions via classification or regression; (ii) predictions before or after publication, and (iii) the level/scope of the prediction. Along with introducing problem formulations, we present necessary preliminaries associated with information items/cascades.

### 2.1 Types of Information Items/Cascades

In this survey, we consider information items as *any* measurable entities in terms of their *popularity*, whereas an information cascade is constituted by a propagation sequence of information items. The best-known and most heavily studied information items are user-generated content (UGC), made possible by the prevailing Web 2.0 services and mobile devices. UGC is no longer produced by editors or publishers only, as traditional audiences have become the consumers and creators of content, such as posts in social networking platforms, threads in discussion forums, and photos and videos on content sharing websites. Such content has thoroughly changed the way users interact with information and other users and how information is created, presented, disseminated, and dies out. Even in conventional information fields, content such as magazines, newspapers, and academic journals are digitalized and consumed in online platforms. Understanding the inner drive

that determines the spread of information items is nontrivial and important for many real-world applications such as advertising, decision making, and caching strategy.

The spread and diffusion of information items are featured by various *endogenous* or *exogenous* factors. For example, in social networking platforms like Twitter and Weibo, users spontaneously post tweets, follow other users, “like” or comment on tweets, and most importantly, retweet other users’ (re)tweets, which will show in the feed of her/his followers (potential future retweeters of this tweet). In contrast, user-driven resharing mechanisms can also be triggered by social influence [5] and by external influences like burst events [44].

## 2.2 Problem Definitions

Information items/cascades are transient, sparse, and biased data. The prediction of information items/cascades is challenging and in some cases, they are in fact unpredictable [32, 143]. In order to better characterize the complex nature of the information items/cascades prediction problem and to explore under which circumstances the problem can be tackled, here we apply threefold categorization: (i) classification/regression; (ii) before/after publication; and (iii) granularity of the prediction, i.e., a macro-level prediction (collective behavior) or micro-level prediction (individual behavior).

**2.2.1 Classification vs. Regression.** The popularity prediction of information items/cascades can be defined as either a classification problem or a regression problem.

**Definition 2.1. Popularity Prediction.** Given  $M$  information item  $\{I_1, I_2, \dots, I_M\}$  and a specific measurement of popularity for each item  $I_i$  published by author  $u_0$  at time  $t_0$ , popularity prediction aims at predicting the popularity  $P_i(t_p)$  of  $I_i$  at a future prediction time  $t_p$ .

When it comes to the *classification* tasks, the exact value of  $P_i(t_p)$  is not required. Instead, one aims to predict whether an information item will be popular or not, given a predefined absolute/relative threshold. For example, some models predict whether an item gains reshares [148, 154], or whether an item will burst in the near future [38, 45] – both of which can be treated as a *binary* classification problem. Another line of works predefine several popularity intervals, and then predict which interval an item is most likely to fall into [59, 81, 133, 134], converting the problem to a classical *multi-class* classification task. As for *regression* models, they primarily predict the exact popularity value  $P_i(t_p)$  (e.g., retweets and citations) that an item will gain in the future [3, 11, 106, 179].

Generally speaking, classification tasks are relatively easier than regression tasks. [12] showed that with only four types of features, the model classification accuracy can reach 84% but perform poorly on regression task. Although the intuition behind formulating regression is more natural and provides a fine-grained scope to analyze which factors affect the future popularity and lead information items to success, precise regression prediction often requires far more information regarding items and users, which therefore implies a higher complexity [58]. In addition, it suffers from undesirable issues such as overfitting, inductive bias, and prediction error accumulations [212]. We note that many of the models we surveyed can be easily modified from regression to classification, or vice versa.

**2.2.2 Ex-ante prediction vs. peeking strategy.** Popularity prediction can also be made *before* or at the time of its publication (a.k.a. *ex-ante* prediction [135]), or *after* its publication (a.k.a. *peeking strategy* [173]). Obviously, prediction in advance is a more challenging task, since only limited information about the items/users is available. Take tweets as an example: in *ex-ante* prediction, the available information includes: tweet’s content (e.g., purely text, emojis, embedded images/videos, URLs/hashtags), profile of the user who posts this tweet (e.g., name, bio, location, join date, number of followers/followees), the time when this tweet was published, etc. Prediction before publication

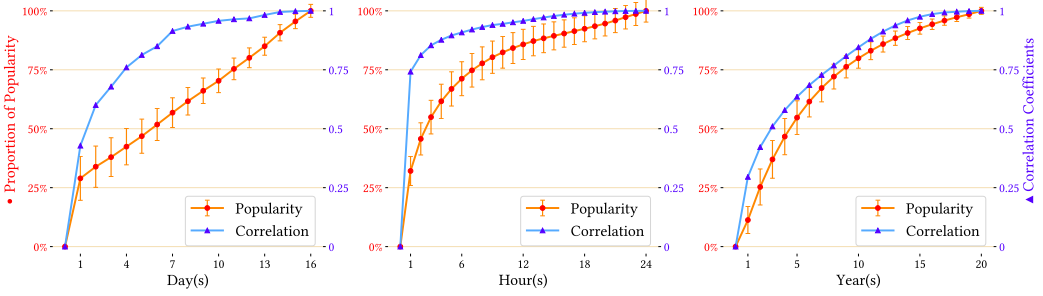


Fig. 3. Left axis: growth trends of cascades over time, the points are means and the bars on each point are variances. Right axis: Pearson correlation coefficients between observed popularity and ground truth popularity over time. Plots are based on three datasets: Twitter hashtags (left), Weibo tweets (middle), and APS papers (right), each have 1.3M, 119K, and 452K information cascades, respectively.

is beneficial in many downstream tasks such as advertising and marketing, as one prefers to know how much attention and/or profit a specific advertisement would have in the near future. Based on predicted results, one can invest in items with highest (potential) influence to spread information to larger audiences, thus maximizing the business interests of corresponding products.

Although *a priori* popularity prediction is attractive, reliable prediction requires an exhaustive set of features, which is often unavailable, sensitive, and hard to obtain in real-world scenarios. Instead, if one could *peek* into the early stage of a cascade’s evolving process, prediction would become much easier [179]. Consider a time interval  $[t_0, t_p]$  of the life-cycle of an information item, and let  $[t_o, t_p]$  denote the interval we can observe, where  $t_o < t_p$  is the observation time. We calculated the proportion of overall observed popularity  $\sum_i^N P_i(t_o)$  to the overall ground truth popularity  $\sum_i^N P_i(t_p)$  (values increased from 0 to 1 monotonically), and Pearson correlation coefficients between every observed  $P_i(t_o)$  and prediction target  $P_i(t_p)$ . Fig. 3 shows the results for three large scale datasets: Twitter hashtags [205], Weibo tweets [28] and APS papers [172]. Clearly, for all three datasets with the increase of observation time  $t_o$ , the prediction task becomes simpler – coinciding with earlier findings [179] that a strong correlation exists between early and future popularity. Fig. 3 also reveals different collective increasing patterns among the three different scenarios of information cascades: for Weibo tweets, the faster they gain attention from users, the faster they faded owing to the newly emerged competitors; for Twitter hashtags, the popularity last longer among users than it does for Weibo tweets; as for APS papers, the item evolving speeds are in the middle.

Predicting the popularity after an item’s publication indicates how it is possible for an item to be successful in spreading. The rationale is that information items that have successfully propagated in their early stages tend to become popular – i.e., early patterns are indicative for long-term popularity [179]. Previous works found that the similar, or even the same, content information may vary significantly in terms of popularity [38]. This also raises questions about the limitations of prediction models – e.g., is it a problem of insufficient data, or is it inherently unpredictable [38, 135, 168, 173]? Is this because the innate quality of content has a minor impact, while the social influence or other external factors govern the eventual popularity of items [144, 168]? Unfortunately, there is no explicit answer applicable to all situations. Such limitations of *ex-ante* prediction drive researchers to design various peeking strategies, and/or investigate the evolution patterns of information items after the publication.



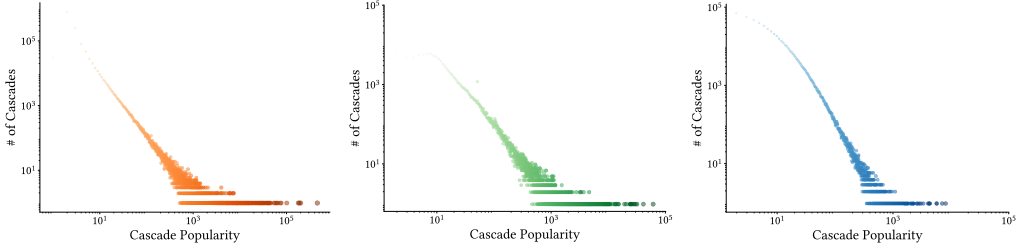


Fig. 4. Popularity distributions of Twitter (left), Weibo (middle), and APS (right) in log-log scales. By fitting the tails of power-law distribution, the exponents  $\alpha$  are 1.916, 2.044 and 2.788, with minimum popularity  $P_{\min}$  greater than or equal to 34, 45 and 58, for Twitter, Weibo and APS datasets, respectively.

Observing who participates in forwarding (or adopts) a piece of an information item enables consideration a sequence of participants/adopters as an *information cascade*, often defined as:

**Definition 2.2. Information Cascade.** Given an information item  $I_i$  and its corresponding  $N$  participants/adopters, an information cascade can be defined as  $C_i = \{(u_j, t_j) | j \in [1, N], t_j \leq t_p\}$ , where each tuple  $(u_j, t_j)$  represents user  $u_j$  participating in this cascade  $C_i$  at time  $t_j$ .

As an example, consider an information item  $I_i$  of a tweet, and use the number of retweets as its popularity. Then, within time window  $[t_0, t_o]$ , we can observe a sequence of retweets which form a cascade – framing the prediction task as: given item  $I_i$ 's observed cascade at observation time  $t_o$ , predict the future popularity  $P_i(t_p) = |C_i|$  at prediction time  $t_p$ .

**2.2.3 Macro-, Micro-, and Meso-level.** The problem of information cascades – especially the prediction of cascades – is interdisciplinary, involving various expertise: text/image/video processing, social influence analysis, temporal and topological modeling, etc. The prediction methods can be further classified according to the *granularity* of tasks: *macro-level* prediction aims at the collective behavior of cascades, whereas a *micro-level* prediction focuses on the individual status/actions of users/items. Macro-level prediction [28, 38, 119] models the cascade from a holistic and global perspective. The output of macro-level prediction is how much attention an information item will get in the future. Instead of taking a cascade as a whole, *micro-level* prediction concerns the behavior of individual users/items, e.g., predicts the activation probability of a specific user, by giving the current cascade statuses, individual/group characteristics, and neighboring relationships [56, 61, 121, 156, 157, 163, 181, 196, 202, 209, 217, 222, 235].

Another scope, referred to as *meso-level* prediction [62], concerns the behaviors of communities/clusters [9, 51, 78, 82, 116, 164, 189, 205]. For example, [205] studied information spreading by considering community structures; [78] proposed a group-level prediction model. Models combining two levels of scope are also proposed, e.g., [230] utilizes micro-level user behavioral dynamics (one-hop subcascades) to help in predicting the macro-level cascade popularity. In this survey, most methods are macro-level popularity prediction of information items/cascades – i.e., given an information item or partially observed cascade, the goal is to predict the future popularity of this item/cascade. Nonetheless, features and methods that can be used for micro-level and/or meso-level predictions are also reviewed, as they can be easily incorporated into macro-level models while providing insights into macroscopic cascade behaviors, e.g., through multi-task learning [36] or multi-scale reinforcement learning [218].

Table 1. A review of frequently used model evaluation metrics.

Metric	Abbr.	Reference
Accuracy	ACC	[20, 34, 38, 47, 48, 58, 59, 73, 81, 85, 90, 92, 123, 130, 137, 141, 164, 170, 171, 173, 185, 187, 198, 213, 222, 237]
Accuracy with tolerance $\tau$	ACC@ $\tau$	[15, 16, 29, 46, 60, 62, 73, 172, 212, 230, 231], [243]*
Area under the ROC Curve	AUC	[23, 38, 47, 48, 58, 59, 85, 87, 148, 157, 192, 198, 226]
Coefficient of Determination	$R^2$	[11, 48, 108, 109, 134, 135, 162, 175, 184, 210], [1, 12, 30, 83] <sup>†</sup>
Coefficient of Correlation	CC	[77, 85, 150, 169, 184, 186, 203, 242], [20, 64, 131, 190, 207, 208] <sup>†</sup>
$F_1$ or $F_\beta$ Score	$F_1, F_\beta$	[4, 19, 23, 38, 45, 47, 48, 56, 58, 62, 71, 72, 81, 85, 86, 98, 101–103, 123, 127, 154, 157, 164, 171, 177, 187, 189, 198, 206, 214, 216, 223, 225, 226, 237], [52, 59, 97, 125, 133, 134] <sup>‡</sup>
Mean Absolute Error	MAE	[48, 96, 150, 198, 210], [214]*, [14, 129, 131, 159, 199, 200, 207, 239] <sup>†</sup>
Mean Absolute Percentage Error	MAPE	[15, 16, 29, 40, 46, 57, 60, 62, 72, 74, 78, 88, 100, 122, 124, 128, 132, 141, 169, 172, 201, 203, 212, 216, 234, 242], [214, 243]*, [92] <sup>†</sup> , [245]* <sup>†</sup>
Mean Square Error	MSE	[3, 62, 72, 160], [36]*, [35, 105, 106, 131, 179, 190, 199, 200, 218, 229, 239] <sup>†</sup> , [28, 37, 83, 118, 119, 245]* <sup>†</sup>
Mean Square Percentage Error	MSPE	[29, 132, 155], [243]*, [188] <sup>†</sup>
Precision	Prec.	[1, 4, 7, 19, 23, 45, 47, 48, 56, 58, 62, 71, 72, 81, 90, 102, 103, 157, 164, 171, 189, 205, 206, 213, 214, 216, 223, 225, 226, 228, 230, 237, 244], [59, 125, 133, 134] <sup>‡</sup>
Recall	Rec.	[1, 4, 7, 19, 23, 45, 47, 48, 56, 58, 62, 71, 72, 81, 90, 102, 103, 157, 164, 171, 189, 205, 206, 213, 214, 216, 223, 225, 226, 237, 244], [59, 125, 133, 134] <sup>‡</sup>
Root Mean Square Error	RMSE	[72, 74, 113, 132, 136, 184], [14, 72, 101, 102, 150, 230, 231] <sup>†</sup>

\* Some models use incremental popularity, i.e.,  $P \rightarrow \Delta P = P(t_p) - P(t_o)$ .

<sup>†</sup> Some models use logarithmic popularity (error), i.e.,  $P \rightarrow \log P$ , or other similar transformations/normalizations.

<sup>‡</sup> Some models use macro/micro-Precision, macro/micro-Recall, or macro/micro- $F_1$ .

### 2.3 Evaluation Metrics

The most commonly used classification metrics are Accuracy, Precision, Recall, and  $F$ -measure. Given a predefined threshold, an item/cascade can be classified as *outbreak* or *viral* if its popularity (size) exceeds that threshold. One known limitation of accuracy metrics is the class imbalance problem, due to the highly skewed popularity distribution [99]. For example, in one Twitter dataset [205], more than 92.8% tweets have retweets  $\leq 10$ , while only 0.114% tweets have retweets  $> 1,000$ . They follow heavy-tailed distributions, which are consistent with the findings of many previous studies [28, 38], as shown in Fig. 4. To address such limitations, researchers reformulate the problem definitions or adopt other types of evaluation metrics, e.g., [38] adopts a class balanced binary classification, by predicting whether a cascade will exceed the median size of all cascades; [157] and [99] filter out a majority of cascades in order to obtain a balanced dataset (undersampling). Most of the *peeking strategy* models requiring a sufficient number of early adopters, implicitly ignore those small samples, consequently yielding balanced classes.

As for the regression task, mean square error (MSE) and its variants are the most popular metrics. Popularity is often shown in logarithmic scale, e.g., MSLE or RMSLE, in order to prevent the loss functions or metrics being affected by extreme values and assure numerical stability [28, 37].



Table 2. Frequently used scenarios in popularity prediction literature.

Platform	Category	Reference
Digg	News Aggregator	[3, 85, 90, 113, 157, 179]
Facebook	Social Networking	[11, 30, 38, 108, 177, 181, 186, 191]
Flickr	Image Sharing	[64, 77, 91, 132, 137, 173, 207, 208, 239]
Twitter	Social Networking	[1, 4, 7, 11, 12, 23, 30, 34, 44, 52, 53, 56, 62, 72, 78, 86, 96, 100–103, 105–107, 118, 119, 130, 132–136, 141, 142, 149, 150, 154, 155, 157–159, 163–165, 167, 169, 170, 173, 175, 177, 189, 190, 192, 200, 205, 206, 214, 216, 218–223, 228, 233–235, 242, 244]
Weibo	Social Networking	[14–16, 19, 28, 29, 36, 37, 45, 57–60, 62, 71, 72, 83, 97, 98, 128, 129, 157, 198, 201, 226, 230, 231, 237, 245]
YouTube	Video Sharing	[1, 3, 38, 44, 46, 52–55, 74, 77, 142, 150, 159, 171, 179, 186, 192, 210, 228, 229]
-	Academic	[16, 28, 36, 37, 47, 48, 83, 118, 119, 124, 172, 194, 212, 215, 245]
-	News Articles	[12, 30, 99, 127, 182, 184, 187, 188, 220, 225]

Table 3. Basic statistics of three benchmark datasets used in this survey.

Dataset	# Cascades	# Nodes	# Edges	Avg. Depth	Avg. $P(t_o)$	Avg. $P(t_p)$
Twitter	1,345,913	595,460	14,430,254	2.198	21.37 ( $t_o = 1d$ )	118.02 ( $t_p = 16d$ )
Weibo	119,313	6,738,040	15,249,636	2.337	56.49 ( $t_o = 1h$ )	117.51 ( $t_p = 24h$ )
APS	514,609	616,316	54,614,609	4.096	4.56 ( $t_o = 3y$ )	7.75 ( $t_p = 20y$ )

Coefficient of determination/correlation and their variants, rankings [182],  $k$ -top coverage [149, 242], are also frequently used metrics in some specific scenarios.

The choice of a metric is often subjective, even under a well-defined problem formulation. Previous works found that a model might perform well in one metric but significantly drop in another [72], making it difficult to do a fair comparison between various approaches. We summarize frequently used metrics as well as their adopters in Table 1.

## 2.4 Datasets

In Table 2, we list widely used and publicly available datasets. The scope of information items ranges broadly, from news articles, academic papers, posted images, music and videos, and these diverse scenarios of popularity prediction cause difficulties in the design of prediction models. Whether for feature extractions, problem formulations, evaluation selections, devising of generative processes, or deep learning architecture designs, it is difficult and sometimes even impossible to fully generalize one model from one specific platform to another. The datasets used throughout this survey contain two microblog datasets, Twitter hashtags [205] and Weibo tweets [28], and one scientific dataset APS papers [172]. The basic statistics of the three datasets are shown in Table 3.

## 3 CHARACTERISTICS OF CASCADES AND FEATURE ENGINEERING APPROACHES

Feature extraction plays a crucial role in predicting the popularity of items/cascades, whether formulated as classification or regression, prior or posterior, micro- or macro-level. Various features and machine learning techniques have been used in a variety of subjects, e.g., illustrating diffusion patterns in network, analyzing time-series evolution trends, and building prediction models.

Table 4. Frequency of five feature groups used by feature-based models.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Temporal	2	3	5	6	5	10	6	10	3	1	0	22.8%
Casc. struct.	1	0	1	2	6	6	5	7	2	2	0	14.3%
Glob. graph	1	1	2	3	7	8	5	7	2	2	1	17.4%
User/item	3	1	7	7	3	10	6	8	3	4	1	23.7%
Content	3	1	7	9	5	7	3	6	4	3	1	21.9%
	4.46%	2.68%	9.82%	12.1%	11.6%	18.3%	11.2%	17.0%	6.25%	5.26%	1.34%	

Table 5. Strategies and formulations of feature-based models.

Strategy	Formulation	Reference
<i>Ex-ante</i>	Classification	[7, 47, 86, 87, 90, 130, 137, 148, 154, 171, 185, 187, 225, 244]
	Regression	[11, 35, 64, 91, 106, 109, 131, 135, 165, 175, 190, 208, 210, 215]
	Both	[12, 48, 105, 108]
Peeking	Classification	[4, 19, 23, 38, 45, 52, 71, 81, 87, 92, 97, 98, 103, 127, 133, 134, 164, 170, 173, 189, 205, 206, 213, 237]
	Regression	[3, 14, 30, 59, 74, 77, 78, 99, 106, 122, 125, 150, 155, 160, 167, 179, 182, 184, 186, 188, 192, 198, 226, 228, 229]
	Both	[1, 48, 58, 72, 73, 85, 101, 102, 105, 113, 141]

We now discuss the characteristics of information items/cascades, categorized into four groups: *temporal*, *structure*, *user/item attributes*, and *content features*. Table 4 and Table 5 summarize the strategies, formulations, and feature frequency of feature-based models in the last decade.<sup>1</sup>

### 3.1 Temporal Features

Temporal features in information items/cascades have been among the most crucial factors in popularity prediction [38, 58]. We discuss the temporal features in the context of feature engineering.

**3.1.1 Observation time.** Temporal features are usually extracted based on peeking strategy, i.e., observing a small number of early participants and their action time, so as to retrieve a sequence of timestamps that can be utilized for feature selection. Since the length of time series are highly irregular, e.g., in a fixed time interval, some cascades may have tens of thousands of participants, while most participants receive little attention, directly utilizing timestamps as a feature is ineffectual in practice. Calculation of time series often needs transformations in advance [221], e.g., dividing the time span in evenly distributed intervals and then calculating the cumulative/incremental popularity, or only observing a fixed number of early participants [38].

In order to predict the popularity  $P_i(t_p)$  at prediction time  $t_p$  based on the information observed at time  $t_o$ , [179] analyzed the relationships between (log-transformed) popularity  $P_i(t_p)$  and  $P_i(t_o)$ . They found a high correlation between early stage and future popularity and used a simple linear prediction model taking the observed popularity in early stage as input to speculate the future popularity.

**3.1.2 Publication time.** Another important temporal feature is the publication time  $t_0$ . As reported in previous works [109, 136, 179, 184, 188], the popularity of a certain item is strongly related to its

<sup>1</sup>We refer readers to the supplementary material for the details per individual work, including strategy, formulation, and incorporated features.

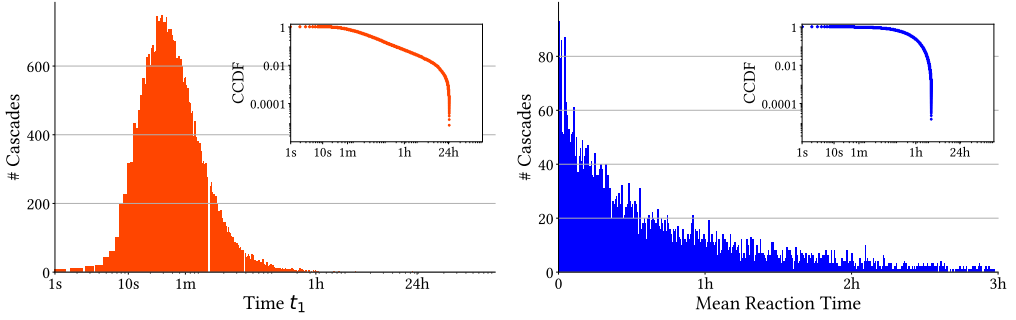


Fig. 5. Left: Histogram of arrival time  $t_1$  of the first retweet in Weibo cascades. Inset: CCDF of time  $t_1$ . Right: Mean reaction time  $\frac{1}{M-1} \sum_{j=2}^M (t_j - t_{j-1})$ . Inset: CCDF of mean reaction time.

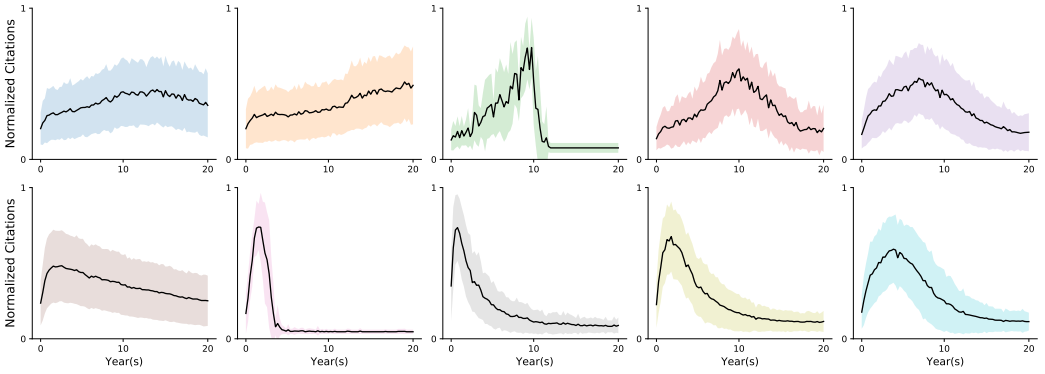


Fig. 6. Ten evolving patterns characterizing the citation cascades in the APS dataset. We use hierarchical clustering implemented by *sicpy* package in Python to cluster the APS papers. For each cluster, we show 20 years of evolving trends of citations, by use of mean values and  $\pm$  standard deviations.

publication time, e.g., items published at midnight are less likely to be viewed while items published during daytime are generally more popular (though they have more competition) [13, 60, 208, 216]. To mitigate the effect of user activity periodicity, several solutions have been proposed. [154] designed 24 local models, each of which is trained with the samples published in a specific hour during the day. The *tweet time* was used in [59] to eliminate the imbalanced diurnal effect of user activities, while other temporal factors such as *digg time* [179], *source time* [188], and *user activeness variability* [208] have been employed to improve model robustness. More simply, some works only explore the items published in daytime to train their models [28, 37].

**3.1.3 First participation time.** The time  $t_1$  when the first participant arrives is also considered as an important temporal feature. As shown in the left graph of Fig. 5, most tweets with at least 10 retweets in 24 hours received the first retweet no later than 1 hour after publication. Similarly, more sophisticated temporal features have been considered in the literature, including *mean arrival time*  $\frac{1}{M} \sum_{j=1}^M t_j$ , *mean reaction time*  $\frac{1}{M} \sum_{j=1}^M (t_j - t_{j-1})$  (right of Fig. 5), *change rate*, *dormant period*, *peek fraction*, etc. For instance, it has been demonstrated that the human reaction time often follows log-normal distribution, e.g., people's reactions on calls, mails, and social networks [17, 234].

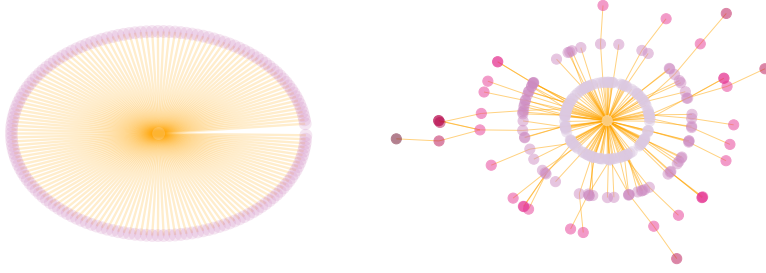


Fig. 7. Two cascade graphs retrieved from APS dataset with popularity  $P_i \approx 200$ . All the nodes of left cascade graph are directly connected to the root node (Wiener index is 1.994). In contrast, the structure of the right cascade is more viral, i.e., a large amount of nodes are connected to the root node indirectly (index is 2.605).

**3.1.4 Evolving trends.** Characterizing the evolving trends of information cascades has been shown to yield informative signals for popularity prediction [8, 74, 115, 221]. Such temporal patterns of time series can be categorized into several types, e.g., *smoothly* increasing or *suddenly* bursting/decaying, dependent on different clustering algorithms. In Fig. 6, we show 10 evolving trends of APS paper citations spanning 20 years by applying an agglomerative hierarchical clustering algorithm [75].

**3.1.5 Discussion.** Despite their importance of temporal features, recent studies also suggested that they may not work in certain scenarios – i.e., their advantages diminish over time [213], and their effect is not always comparable to other features [206]. Note that many works try to model and analyze the temporal evolution trends of cascades, e.g., based on time series patterns [220, 221], survival analysis [111, 112], and point processes [172, 242]. A special class of works, *generative* models, will be systematically reviewed in Section 4.

## 3.2 Structural Features

Cascade structure, sometimes also referred to as *information diffusion*, has been studied extensively [14, 38, 56, 58, 234, 237], and works can be categorized according to their way of modeling cascades: (i) *participants only*, i.e., only *cascade graphs* are involved; (ii) *global graph*, i.e., both participants and non-participants are considered; and (iii) *r-reachable graph*, i.e., a compromise, extending the cascade graph within the scope of global graph.

**3.2.1 Cascade graph.** A cascade graph is constructed based on its participants and their interactions:

**Definition 3.1. Cascade Graph.** Given an information item  $I_i$  and the corresponding cascade  $C_i$ , a cascade graph is defined as  $\mathcal{G}_c = \{\mathcal{V}_c, \mathcal{E}_c\}$ , where nodes  $\mathcal{V}_c = \{u_0, u_1, \dots, u_N\}$  are all participants of cascade  $C_i$ , and matrix  $\mathcal{E}_c \subseteq \mathcal{V}_c \times \mathcal{V}_c$  contains a set of edges representing all the immediate relationships between  $\mathcal{V}_c$  in a cascade (e.g., retweeting or citing relationship).

Cascade graph  $\mathcal{G}_c$  characterizes the process of information diffusion of an item  $I_i$  such as the *spreading directions* and *graph topology*. One pioneering study employed structural features into popularity prediction to predict which Twitter users are likely to mention a specific URL [56]. The correlation between cascade popularity and two structural features (edge density and depth) of microblogging network among early participants was analyzed in [14]. It found that with lower value of edge density and higher value of depth – implying a diverse group of early participants – the popularity of cascades is more likely to be large.

The topological structure (shape) of different cascade graphs may vary significantly from each other, even if the number of nodes is the same [226]. As shown in Fig. 7, two cascade graphs

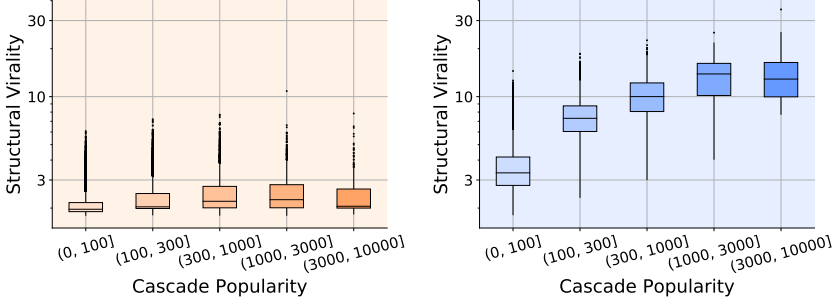


Fig. 8. Box plot of structural virality (Wiener index) by cascade popularity for Weibo (left) and APS (right). The lines in boxes are median value, whereas boxes are quartile values of structural virality from lower to upper. The relationship between cascade popularity and structural virality for Weibo tweets is neutral; on the contrary, for APS papers, the correlation is quite positive.

from APS have distinct shapes. Although both cascades have  $\approx 200$  nodes, eventually, the structure, depth, and structural virality (quantified by Wiener index [67]) of the two cascades are totally different. This means that simple structure measurements, such as node degree, depth, PageRank, etc., are limited in explaining whether an information item would be popular or not. For example, non-viral cascades (or broadcast cascades, e.g., left graph in Fig. 7) could have considerably greater popularity in the end, although the virality value is very small at the beginning. In contrast, a viral cascade graph (e.g., right graph in Fig. 7) not always results in a larger cascade – in fact, the Pearson correlation coefficient between cascade popularity and their structural virality is relatively low. This phenomenon implies that the initial structural features become less important as cascades grow over time [38, 237].

To further investigate the relationship between cascade popularity and structural virality, we draw box plots over the distribution of structural virality on different ranges of cascade popularity for both Weibo and APS (cf. Fig. 8), from which we make the important observation that the popularity of microblogs is dominated by two different propagation mechanisms, i.e., *broadcast diffusion* and *word-of-mouth*, dependent on the cascades' eventual size. For cascades smaller than 3,000, the median value of structural virality grows slightly with cascade popularity (i.e., from 1.967 to 2.258). However, for tweets with popularity  $> 3,000$ , their median structural virality decreased to 2.053, which means one of the most important drivers for cascades to succeed comes from the promotion of organizations or celebrities who have a large number of followers [11, 67]. In other words, the dominant factor for very large cascades is the *broadcast diffusion* mechanism, in which case users are more likely to directly interact with information sources (or important intermediates) and thus the growth of cascade does not rely on viral spreading [49]. Interestingly, the same trend can also be observed in scientific paper cascades, though their structural virality is generally larger than Weibo tweets (3.779 vs. 2.164).

Other features of the cascade graph include node degree/eigenvector/closeness/betweenness centrality, authority/hub score, graph depth/density/diameter, structural diversity/virality, and various variants of the same such as mean, median, percentile, entropy, standard deviation, etc. More examples can be found in the supplementary materials.

**3.2.2 Global graph.** Apart from the cascade graph which presents the *local* spread of information, many works studied global graphs [38, 58, 66, 146, 163], the definition of which depends on the application scenarios. For example, in social networks, the most common global graph is the

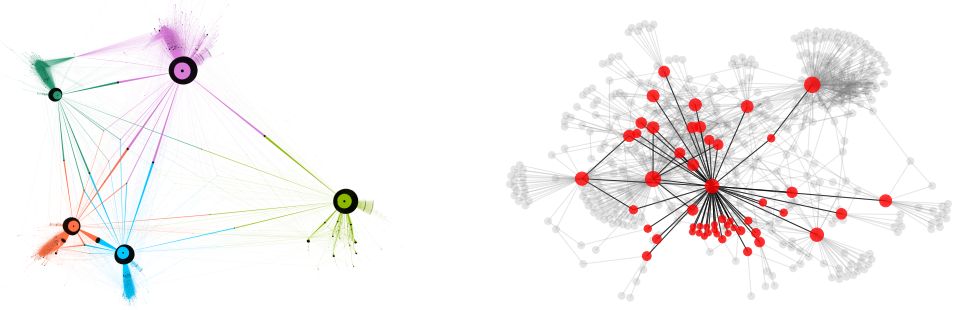


Fig. 9. Left: a global graph retrieved from the five largest cascades in Weibo dataset. It contains 262,458 nodes and 324,540 edges. Edges from different cascades are in different colors. We can clearly see that these cascades are not isolated. Right: an example of 1-reachable graph retrieved from Weibo dataset. Dark nodes (red) come from a cascade graph  $\mathcal{G}_c$ , where pale nodes are neighbors of the red nodes.

follower/followee graph or friendship graph [84]. There also exist many other types of global graphs – e.g., [84] retrieved a sparse hidden friend graph from the dense follower/followee graph and [85] defined a co-participation graph from Digg comments. Based on the interactions between users (retweet, comment, mention, like/dislike, etc.), from Twitter an interaction global graph can be constructed, e.g., a global graph based on historical mentioning relationships between users [58, 59, 120, 164, 219]. From APS, based on the interactions between authors (cite, collaborate, etc.), a citing global graph or collaborating global graph was constructed in [47]. These interaction global graphs can be useful in various prediction tasks, especially when the explicit social graph is unavailable, and a suitable choice to represent the actual diffusion of information from historical behaviors [201].

**Definition 3.2. Global Graph** A global graph  $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$  is a collection of  $\mathcal{V}_g$  nodes and a set of  $\mathcal{E}_g \subseteq \mathcal{V}_g \times \mathcal{V}_g$  edges representing the relationships between nodes, e.g., friends or collaborators.

Based on this simple definition, other types of graph can be defined, e.g.: (i) a directed graph with edges indicating the direction from one to another, e.g., unilateral relationships such as follower/followee in social networks; (ii) a weighted graph whose nodes or edges are associated with assigned weights, e.g., multiple times of interactions between users [222]; (iii) a heterogeneous graph whose nodes or edges have more than one type of attribute, e.g., a graph with *authors*, *papers*, *venues* as nodes, and *publish*, *cite*, *co-authorship* as edges; (iv) an attributed graph whose nodes or edges have associated features, e.g., the metadata of information items, the texts of tweets, the titles/abstracts of papers, the embeddings of images, etc.

A global graph provides us a *macro* perspective to analyze how information disseminates to individuals and to communities, and how information cascades grow their popularity in the context of their positions in the global graph. Compared to information diffusion in the cascade graph only – which shows the local spread patterns for one specific cascade – the global graph describes all the relationships between users and potential routes for diffusion. For example, in social networking platforms, besides various external stimuli such as recommendations from the system, key words search from users, etc., information items are primarily discovered and disseminated through the users' social networks [179].

The left portion of Fig. 9 shows an example of global graph, which is composed of the five largest cascades in the Weibo dataset. Nodes in the graph are individual users while edges represent the retweeting relationships between users. As it shows, a large number of users not only participate



in one cascade but also act as bridges between separate cascades. The historical behaviors, personal preferences, and communities of users can further help us to identify the roles of users in the cascade graphs [189, 206]. A complete global graph can be very large, e.g., a retweeting global graph retrieved from Weibo has more than 6M nodes (users) and 15M edges (retweet actions), while the citing global graph retrieved from APS has 422K nodes (authors) and 54M edges (cite actions). It is noteworthy that nodes in cascade graph may not necessarily appear in the global graph. For example, in Twitter, a user can retweet posts from other users who are not her/his followers or followees. This limits the usability of models which rely only on the global graph [77, 215].

Previous studies have also highlighted the influence of social community, e.g., social reinforcement and homophily of social contagions [6, 31, 138, 163]. As shown in [205, 206], the viral cascades often diffuse across more communities than non-viral cascades, i.e., they are less likely to be trapped into a low number of communities. However, the analysis of the inter-community diffusion patterns of tweets in [189] demonstrated that the effects of community features are relatively small compared to other features such as past success and user degrees.

**3.2.3  $r$ -reachable graph.** Based on the definition of cascade graph and global graph, we can define a sub-graph retrieved from the global graph and named  $r$ -reachable graph:

**Definition 3.3.  $r$ -reachable Graph** Given a global graph  $\mathcal{G}_g$ , and its cascade sub-graph  $\mathcal{G}_c$ , an  $r$ -reachable graph of  $\mathcal{G}_c$  is defined as  $\mathcal{G}_c^r = \{\mathcal{V}_c^r, \mathcal{E}_c^r\}$ , where  $\mathcal{V}_c^r$  contains (1) all nodes in  $\mathcal{V}_c$ , and (2) all nodes in  $\mathcal{V}_g$  are within  $r$ -hops of nodes in  $\mathcal{V}_c$ . For example, when  $r = 1$ ,  $\mathcal{V}_c^r$  contains all  $\mathcal{V}_c$  and all immediate neighbors of nodes in  $\mathcal{V}_c$ .

The right portion of Fig. 9 shows an example of  $r$ -reachable graph ( $r = 1$ ), where dark nodes and their interactions form the cascade graph.  $r$ -reachable graph tells us how many nodes are exposed to the active nodes and their topology. The rationale behind modeling an  $r$ -reachable graph is that the highly exposed nodes would potentially bring more nodes into this cascade in the future [59].

Previous works have utilized  $r$ -reachable graph to facilitate the prediction of cascades' popularity [38, 134]. Global graph  $\mathcal{G}_g$  and several 1-reachable graphs  $\mathcal{G}_c^1$  (a.k.a. border graph) from Weibo retweet cascades were constructed in [58, 59], extracting structural features from  $\mathcal{G}_g$  and  $\mathcal{G}_c^1$  to predict the popularity of tweets. A comparison of several structural and content features in [134] revealed that the number of nodes in the cascade graph and 1-reachable graph are the two most predictive of 53 features in predicting the popularity of Twitter hashtags. However, the construction and calculation of an  $r$ -reachable graph can be costly, e.g., for a cascade graph with dozens of nodes (on average), its 2-reachable graph may contains tens of thousands of nodes.

**3.2.4 Discussion.** The influence of the cascade graph's early structure on the final item popularity remains largely unclear, with neither of the standard structural feature selection approaches making an accurate prediction. In addition, different platforms have unique diffusion mechanisms that may cause dynamics to differ from the well studied social network scenarios [38, 103, 114, 117, 120, 157, 242]. For example, the findings in [6] show that the correlation between the popularity of LindedIn signup cascades and structural virality is remarkably high, which shows a very similar structural behavior to APS citation cascades. Finally, even for a particular platform, the types of items (events) and the item contents may affect the information diffusion behavior [67, 237].

### 3.3 User/Item Features

Temporal and structural features require probing into early observations, which is sometimes impractical. Therefore, a number of works have focused on the features associated with users and information items, which have unique properties and innate attractiveness that make them especially useful in predicting popularity before publication.

**3.3.1 User features.** User behaviors play crucial roles in information dissemination and consumption (e.g., viewing, commenting, sharing, and preferring). One of the most common user features is the number of followers as a proxy of user influence, which implies the speed and timing of future popularity [234]. Those who have a large number of followers (audiences), e.g., celebrities and news organizations, are more likely to generate large cascades than normal users, since their messages are more visible in the network [11, 86, 178]. However, large cascades are not only produced by influential users and it is of interest to study large cascades originated by normal users instead of celebrities [49]. Many other user features have been extensively studied and explored for analyzing and predicting the popularity of information items/cascades, e.g., profiles (name, age, region, education, employment, account creation date, etc.) [232], historical behaviors (frequency of publish items and interact with other users, active time, etc.) [11, 81, 178], user interests [223], collectivity [129], similarity [173], past success [11], activity/passivity [222, 228], discoveries [139], affinities and responsiveness [232], etc.

**3.3.2 Item features.** In addition to the effects of user features, the effects of item characteristics on diffusion have also been studied in the literature. For example, [113] analyzed how Digg users' interfaces affect the visibility of items; [187] showed that metadata of news articles affects the volume of comments (e.g., the publication date of the article, number of articles published at the same time, number of articles with similar content, etc.); [85] proposed to use entropy computed among information categories and topics; and [208] used prevalence variability to analyze how popularity of different types of items changes over time.

**3.3.3 Discussion.** Most of the user/item features are self-explanatory, and it is not too hard to understand their correlations to popularity. For instance, users who have a larger audience, who live in a densely-populated area, who speak a widely-used language, who often discuss trending topics, etc., would have a greater chance to make their information popular. However, some features require more complicated algorithms and calculations to obtain, e.g., user influences, preferences, and similarities. Many prior works studied how individual users affect the diffusion of information. In [107], number of followers, PageRank, and number of retweets were combined to rank influential users. [19] studied three types of influence during information diffusion, i.e., interest-oriented influence, social-oriented influence, and epidemic-oriented influence, and showed that each of these three surrogates of influence can contribute to the popularity prediction. However, some previous works found social influence that, merely measured by user topology in graph, reveals little about the actual influence [32, 45]. Leveraging historical evidence of user behaviors, an orthogonal sparse logistic regression (OSLOR), which predicts cascade outbreak via selecting users who are more powerful and less redundant, was proposed in [45].

## 3.4 Content Features

Content is recognized as the inner drive and one of the key factors that leads items to success [225], e.g., breaking news, rumors/fake news, hot spots, controversial/peculiar topics, disinformation and misinformation, etc., attract significantly more attention than normal content.

**3.4.1 Text Content.** Text features are widely adopted by existing popularity prediction models, since text is ubiquitous information appearing in articles, microblogs, image/audio/video captions/descriptions, and can even be retrieved from images/audio/videos.

Existing works analyze the user-generated textual content with various language models, such as Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) [21], combined with typical machine learning models such as naïve Bayes, SVM and linear regression to predict item popularity. For example, TF-IDF and LDA are used to learn the topic distributions

of tweets in [81]. TF-IDF was also used in [223] to estimate the importance of keywords in the content of user tweets, and further calculated the mutual correlation between all the historical content of her/his tweets and the content of a specific item to measure how much a user likes a piece of content, i.e., if the correlation is high, then this user may have a higher probability of adopting this item. In [90], authors analyzed several semantic and statistical content features of Digg comments, including comment length, number of verbs/nouns, content entropy, readability, subjectivity/objectivity, polarity, etc., and they found that people prefer to retweet short, simple, and readable content. The impact of textual and semantic features on the volume of news article comments was discussed in [187]. Predicting whether a tweet would be retweeted purely based on the content of tweets (including marks, terms, emoticons, sentiments and topics) from [148] was later adopted in [105] to predict the size of retweet cascades, and for sentiment impact on retweetability of political tweets in [175]. The design of content features for predicting scientific impact of academic papers, including LDA topic models which extract probability distribution from paper titles and abstracts, topic diversity calculated by entropy, and paper novelty, as well authority of authors [95], etc., in [47, 48, 215] showed that the content features of papers play critical roles in prediction. The sentiment and self-disclosure of tweets analysis, along with calculations of topic affinity of tweets using a Twitter-based LDA for better topic discovery and extraction, was presented in [232]; another study embedded sentences in vectors to combine sentiments of the descriptions of short videos [35]. Combining domain, spam score, tweet topic, user topic, topic interactions between tweet and user, and tweet category as the content features were studied in [135]. The comparative study of transactional (TF-IDF or LDA) and semantic (sentiment, controversy, content richness, hotness and trend momentum) models and their inter-relationships in [244] considered three kinds of diffusion mechanism: cascades of URL, hashtag, and retweet. It was found that semantic features have better performance on hashtag cascades (with higher content complexity), while transactional features are more effective on URL and retweet cascades (with lower content complexity).

**3.4.2 Content features of image.** The characteristics of images are significantly different from texts, in terms of modeling and retrieving image features, which often require techniques from computer vision learning. Some basic features regarding an image can be retrieved easily if available, e.g., in [137] the authors analyzed the basic attributes associated with images, including size, time, date, season, orientation (landscape/portrait), device, dominant color, and even whether taken on flash, as well as image resolution, location (coordinates), caption, tag, etc.

One seminal work [91] studied the correlations between features of Flickr images and their normalized popularity (i.e., view count). In particular, the content features of images – categorized as simple human-interpretable features, low- and high-level image features – help in improving the prediction performance significantly, e.g., images with striking colors tend to be popular. The low-level image features, such as gist, texture, color patches, gradients, along with representative features are extracted by local binary pattern [151] and convolutional neural networks (CNNs) [110], while the high-level image features are retrieved from an ImageNet classifier [104] among 1,000 object categories, each indicating presence or absence of a certain category in the image. The results were later extended in [64] by taking sentiment features into consideration.

In [185], authors analyzed a wide spectrum of visual features of Pinterest images and evaluated their influence on popularity. Except for frequently used individual features, visual features were grouped into two sets: aesthetic and semantic features, the former accounting for the “beautiffulness” of an image (e.g., dominant colors, saturation, brightness, contrast, texture, background area, region focus and focus centrality/density, etc.), while the latter are extracted from various computer vision techniques, e.g., SIFT local descriptor [126]. Other works include: [38] which estimates the

likelihood scores of categories as the content features of Facebook photos; [19] which preprocessed images of tweets into a bag of visual words as content features; and [131] which, in addition to extracting features like texture and color from images, also adopted VGG19 [174] to extract deep features.

**3.4.3 Other content features.** For online videos, basic features include the video length, resolution, and number of frames. Using several key frames of a short video to represent the video, [35] added color histogram and aesthetic features, quality features, object features retrieved from CNNs, and sentiment features trained by SentiBank [25], and used them as visual features of videos (acoustic features of videos were also considered – including Mel-Frequency Cepstral Coefficients and Audio-Six). Detection of faces and texts in video frames in [186] combined scene dynamics, clutter, rigidity, thumbnail, and features extracted from ResNet [76].

There have been some interesting experiments to collect features using crowd-sourcing. The controlled experiments on songs in [168] found that social influence increases the inequality and unpredictability of a song's popularity. The experiments on Amazon Mechanical Turk [53] required the respondents to watch several YouTube videos and choose which video they were willing to share and predict which video would be the most popular in the future. Although the perceptions of users are very subjective and consensus among them is rare, the popularity of the target video became significantly higher once the respondents reached an agreement.

**3.4.4 Discussion.** The content of information items has proven to be a qualified predictor, and the prediction of popularity can be made on and before the items' publication when no posterior information is observed (cold start). Many existing studies have analyzed the relationships between content of items and their popularity, including linguistic characteristics of texts [180], color and objects in images [91], etc. However, in analyzing the root causes of an item's final popularity, it is still hard to disentangle the effects of such descriptive factors from the effects of the item's intrinsic content [176].

Opinions on the effectiveness of content features differ. Some previous works consider content features as *weak* predictors compared to other features such as temporal, structural, and/or individual features. For example, [38] found that content features get less important when observing more participants, and [11] found that the model was not improved by addition of content features, which is also confirmed in [106]. Authors of [135] argued that content features explain the variance of popularity poorly. Content based methods suffer from issues that hinder their performance – e.g., despite the recent success of deep learning, natural language processing and computer vision, it is still challenging to effectively and efficiently identify, retrieve, and model the content of items, and the results are far from satisfactory. In addition, previous works found that even for items with identical content, their popularity varies greatly [24, 38, 65, 109], raising questions about whether, if one relies only on content features, the popularity of items/cascades is inherently unpredictable or cannot be predicted *a priori*.

## 3.5 Prediction Methods

Since the main challenge of feature-based models lies in the feature engineering, improving the capability of prediction models is not the focus in related literature. For example, [38] demonstrated that most of the machine learning methods have similar performance, despite time/space complexity. For completeness, we summarize common machine learning methods, or what are adopted as their main building blocks, as the prediction methods in Table 6. In addition to the methods listed in Table 6, a few learning paradigms such as inductive/transductive learning, early feature fusion, and multi-view learning approaches have been investigated to predict the popularity of information

Table 6. Machine learning methods.

Method	Abbr.	Reference
Autoregressive (–moving-average)	AR (MA)	[46, 73, 74, 122, 136, 220, 233]
Decision Tree	DT	[12, 38, 47, 48, 58, 59, 62, 73, 85, 98, 101–103, 105, 106, 133–135, 137, 192, 198, 215, 226]
$k$ -nearest Neighbors Algorithm	$k$ -NN	[12, 46, 59, 62, 73, 85, 98, 99, 122, 133, 134, 215]
Linear Regression	LR	[1, 3, 12, 14, 24, 30, 38, 46, 48, 62, 72, 73, 77, 90, 99, 101, 102, 122, 131, 135, 155, 159, 167, 179, 182, 184, 190, 206, 210, 215, 225, 229, 230, 234, 242]
Logistic Regression Classifier	LRC	[38, 45, 47, 48, 59, 81, 86, 87, 98, 103, 134, 137, 148, 150, 157, 164, 173, 198, 213, 223, 237]
Multilayer Perceptron	MLP	[47, 62, 101, 102, 198]
naïve Bayes classifier	Bayes	[38, 47, 48, 59, 62, 73, 86, 133, 134, 170, 171, 225, 226]
Random Forests	RF	[4, 23, 38, 47, 48, 62, 71, 72, 87, 98, 103, 135, 141, 160, 173, 185, 187, 189, 198, 205, 206, 213, 226, 237]
Support Vector Machine	SVM	[12, 35, 38, 47, 48, 59, 62, 64, 72, 73, 85, 87, 91, 97, 98, 101–103, 108, 127, 130, 131, 133, 134, 137, 157, 165, 173, 186, 190, 198, 215, 222, 223, 228, 237]

items/cascades [35, 78, 131, 208]. In general, the selection of prediction methods in existing works is to some extent casual, and a systematic evaluation of these models is still lacking.

### 3.6 Global Overview of Pros & Cons

Feature-based models are often reported as competitive and explainable compared to others [72]. However, the main bottleneck of feature-based models preventing their implementation in real application is the hand-crafted feature engineering. Some features are hard to obtain due to privacy concerns, such as preferences and viewing histories, and some (e.g., user classification and clustering) are computationally intensive, which limits the models' scalability. In this spirit, given an exhaustive set of features, how to select a relatively small portion of representative features to maximize the marginal benefits of prediction between effectiveness and efficiency, becomes a critical consideration for designing practical prediction models [58, 226, 242]. Also, the existing models require more and more features like history view counts and diffusion paths, which in most cases are unavailable and, most importantly, are not generalizable to different scenarios.

We reviewed four categories of features and corresponding prediction models. However, it is impossible to mention all features and models, or to evaluate their performance in all feature combinations. A comprehensive study on evaluating different features in different conditions would be beneficial in standardizing the feature engineering and feature selection/combination.

## 4 GENERATIVE MODELS

Many real-world phenomena – e.g., information retweeting, hospital admissions, and citations of scientific papers, – can be formulated as event sequences in the continuous temporal domain. Modeling the arrival/occurrence of event sequences or the participation time series is a fundamental step towards understanding the underlying dynamics of the information diffusion. The spreading of information items is therefore widely characterized by probabilistic statistical generative approaches such as epidemic models, survival analysis, and various stochastic point processes. In this section, we review these generative models for popularity prediction.

#### 4.1 Poisson Processes

Models based on point processes distinguish themselves from feature-based models because of their statistical, probabilistic, and generative forms. Point processes are often used when modeling time series, e.g., the arrival rate of customers, phone calls, and mechanical failures, in queuing theory and operational research.

A generative probabilistic model using reinforced Poisson process (RPP) was proposed in [195] (followed by [172]) to predict items' popularity (in their case the scientific impact of papers) using three key ingredients: (i) fitness or attractiveness of an item; (ii) temporal decay function; and (iii) reinforcement mechanism (e.g., *rich-get-richer*). Specifically, the rate function of *inhomogeneous* Poisson process for information item  $I_i$  is defined as:

$$\lambda_i(t) = \alpha_i \phi_i(t) P_i(t), \quad (1)$$

where  $\alpha_i$  is the attractiveness of item  $I_i$  represented as a single value evaluated by maximum likelihood estimation [195], or follows posterior distribution from a conjugate prior [172];  $\phi_i(t)$  is the relaxation function or decay function to characterize the aging effect;  $P_i(t)$  is the total number of popularity  $I_i$  received at time  $t$ . With the above defined rate function, given  $(j-1)$ -th participant arrives at  $t_{j-1}$ , the probability of  $j$ -th participant arrives at  $t_j$  is given by:

$$\Pr(t_j | t_{j-1}) = \alpha_i \phi_i(t_j) P_i(t_j) e^{-\int_{t_{j-1}}^{t_j} \alpha_i \phi_i(t) P_i(t) dt}. \quad (2)$$

Based on RPP model, [60] proposed an extended model PETM. Unlike RPP, which is designed to quantify the long-term scientific impact, PETM is more applicable to the scenario of microblogging platforms. Specifically, PETM uses a power-law time decay function  $\phi_i(t) = t^{-\gamma_i}$  ( $\gamma_i > 0$ ) to substitute the log-normal distribution  $\phi_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} \exp\left(-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}\right)$  in [172], since they observe a power-law distribution in retweeting dynamics. The linear reinforcement mechanism in RPP was replaced by an exponential decay reinforcement function  $p_i(t_j) = \sum_{j=0}^{P_i(t_j)} \exp(-\delta_{ij})$  ( $\delta_i > 0$ ) to regularize the large popularity, and a time mapping function  $f: t_j \rightarrow \hat{t}_j$  to mitigate the influence of diurnal rhythm of user activities following their earlier works [58, 59].

A dynamic activeness model based on Poisson process was introduced in [124], aiming to predict the intensity (cascade size), coverage (number of users involved), and duration of a trend in DBLP co-authorship network. The RepostTree model [128] decomposes a sequence of Weibo retweets into tree-structures based on users' follower relationships and subsequently computes a composite Poisson distribution based on early observation of retweets with maximum likelihood estimation. Another extension of RPP model [57] decomposed the Weibo retweet cascade graph as  $k$  sub-processes (in their case  $k = 3$ ), and kept the attractiveness  $\alpha_{i,j}$  and log-normal distribution relaxation function  $\phi_i(\tau)$  (cf. Eq. 1) unchanged. Then the rate function was defined as:

$$\lambda_i(t) = \sum_{j=1}^k \alpha_{i,j} \phi_i(t, t_j) P_i(t), \quad (3)$$

where  $t_j$  is the time when the  $j$ -th sub-process starts. Recently, learning the collective user behaviors in cascades by utilizing Poisson process with a latent user interest layer was proposed in [129].

#### 4.2 Survival Analysis

Survival analysis is a branch of statistics that is widely used in engineering, economics, and sociology [140]. Borrowing the idea from survival analysis to predict the popularity of online contents in [111, 112] a model based on Cox proportional hazard regression [43] was implemented, consisting of two components: (i) a set of explanatory risk factors  $\{x_j\}_j$ ; and (ii) a baseline Weibull



distribution function  $h_0(t) = \frac{\gamma}{\lambda}(\frac{t}{\lambda})^{\gamma-1}$ . After fitting the parameters  $\{\beta_j\}_j$  of Cox regression and parameters  $\gamma$  and  $\lambda$  of Weibull distribution, the hazard function is approximated by:

$$h_i(t) = h_{i0}(t)^{\sum_j \beta_{ij} x_{ij}}. \quad (4)$$

In similar spirit, a dynamic egocentric model of citation networks based on a counting process was proposed in [194], with intensity function  $\lambda_i(t) = h_{i0}(t) \exp(\sum_j \beta_{ij} s_{ij}(t))$ , where  $h_{i0}(t)$  is the baseline hazard function,  $\{s_{ij}(t)\}_j$  is a set of risk factors dependent on item  $I_i$ , and  $\beta_{ij}$  are parameters need to estimate. Risk factors are calculated by citation graph and LDA.

The Weibull distribution in survival analysis was also adopted in [230, 231], introducing a networked Weibull regression model (NEWER) to characterize the information diffusion in a Weibo global graph. By modeling the retweeting event that happened on a node as the survival process, the density function is defined as  $\frac{\gamma_i}{\lambda_i}(\frac{t}{\lambda_i})^{\gamma_i-1} \exp(-(\frac{t}{\lambda_i})^{\gamma_i})$ , where  $\lambda_i$  and  $\gamma_i$  are parameter vectors linearly associated with the user's representation vector containing user and structural features extracted from user historical behaviors and global graph, respectively.

There are a few other survival analysis based models, e.g., [177] and [214], with the fundamental difference between these models lying in the design of hazard functions and survival probabilities, which should be appropriate for the target application with empirical data.

### 4.3 Self-exciting Hawkes Point Processes and Epidemic Models

Self-exciting point process based models predict the rate of events (e.g., retweets or citations) as a function of time and the previous history of events. One of the seminal works was reported in [44], describing the dynamics of viewing behavior of YouTube videos by two factors: (i) a response function representing a power-law distribution of human activity waiting time; and (ii) an epidemic process which can be modeled by self-exciting Hawkes conditional Poisson process. The conditional intensity was defined as:

$$\lambda_i(t) = V(t) + \sum_{t_j < t} \mu_j \phi(t - t_j), \quad (5)$$

where  $V(t)$  is the exogenous source,  $\mu_j$  represents the number of potential participants that will be influenced by  $u_j$  to join in this cascade  $C_i$  at time  $t$ , and  $\phi(t) \sim 1/t^{1+\theta}$  ( $0 < \theta < 1$ ) is a memory kernel. The idea was extended in [233], utilizing self-exciting Hawkes process to characterize the tweet popularity in Twitter. The number of followers and a Pareto distribution of the kernel  $\phi(\tau)$  were used to model the magnitude  $\mu_j$  of each event.

The SpikeM model [136] fits the exponential rising and power-law falling patterns of information diffusion by incorporating (addressing) both the advantages (disadvantages) of epidemic model and self-exciting Hawkes processes. The base model of SpikeM can be represented as:

$$P(t_{j+1}) = |\mathcal{V}_u| \cdot \sum_{t=t_b}^{t_{j+1}} \mu(t) \phi(t_{j+1} - t) + \epsilon, \quad (6)$$

where  $|\mathcal{V}_u|$  is the number of unaffected users,  $\mu(t) = P(t_j) + S \cdot 3(t = t_b)$  is the available stimuli at time  $t$  ( $3(\cdot)$  is the indicator function), and  $S$  is the influence of external shock. SpikeM takes periodicity of human behavior into consideration by multiplying Eq. (6) with an additional time-dependent periodicity factor  $p(t_{j+1})$ . Extensions can be found in recent works [100, 158], which share the same idea of linking the epidemic model and Hawkes process.

Dual sentimental Hawkes process (DSHP) [46] is another generative prediction model built on Hawkes process. DSHP considers sentimental impact of information items, and two self-exciting

processes, i.e., self-excitation and cross-excitation. The intensity function of DSHP is:

$$\lambda_i(t) = V(t) + \underbrace{\sum_{t_{j,k} < t} \mu_{ik} \phi_k(t - t_j)}_{\text{self-excitation}} + \underbrace{\sum_{t_{j,l} < t} \mu_{il} \phi_l(t - t_j)}_{\text{cross-excitation}}, \quad (7)$$

where the exogenous source is formulated as a Pareto distribution  $V(t) = \alpha/t^{\alpha+1}$ , and kernels ( $\phi_k$  and  $\phi_l$ ) are formulated as two Rayleigh distributions, i.e.,  $\phi(\tau) = \frac{\tau}{\sigma^2} \exp\left(-\frac{\tau^2}{2\sigma^2}\right)$ .

The SEISMIC (self-exciting model of information cascades) in [242] aims to predict the retweet counts. It first fits a memory kernel  $\phi(\tau)$  of human reaction time – a constant early on, then following a power-law decay later – and then measures the retweetability of tweet  $I_i$  at time  $t$  with a time dependent tweet infectiousness  $p_i(t)$ . The intensity is described as:

$$\lambda_i(t) = p_i(t) \cdot \sum_{t_j < t} \mu_j \phi(t - t_j), \quad (8)$$

where  $\mu_j$  is the number of followers of user  $u_j$  who retweet the original tweet  $I_i$ . Self-excited Hawkes process (SEHP) proposed in [15] is another example, which defined the intensity function as:  $\lambda_i(t) = V_i \phi(t) + \mu \sum_{t_j < t} \phi(t - t_j)$ , where  $\phi(\tau)$  is an exponential decay memory kernel.

Time-dependent Hawkes process (TiDeH) [96] is yet another extension of SEISMIC taking into account the circadian oscillations and aging of information. Later, leveraging a combination of advantages of feature-based models and generative models, [141] proposed a hybrid model consisting of Hawkes process with a predictive layer trained from random forest and a feature-based predictor. A self-exciting Hawkes process model to predict the individual paper citation count was introduced in [212], with intensity defined as:

$$\lambda_i(t) = V_i \phi_{ik}(t) + \mu_i \sum_{t_j < t} \phi_{il}(t - t_j), \quad (9)$$

where  $V_i$  captures the intrinsic popularity (or quality) of each paper, which is defined by paper/author-specific covariates, and where  $\phi_{ik}(\tau)$  and  $\phi_{il}(\tau)$  are exponential aging kernels.

A combination of Hawkes intensity processes (HIP) with *exogenous* stimuli and *endogenous* triggering effect from Twitter and YouTube was proposed in [159], to predict the popularity of videos. In particular, the self-exciting Hawkes process was extended by defining the expectation of intensity of the observed events:

$$\xi_i(t) = \mathbb{E}[\lambda_i(t)] = V(t) + \mu_i \int_0^t \xi(t - \tau) \phi_i(\tau) d\tau, \quad (10)$$

where  $V(t)$  accounts for the unobserved external influence from Twitter and the number of video shares, and  $\phi_i(\tau)$  is a power-law memory kernel. HIP models the video view volumes directly rather than taking individual events one after another.

#### 4.4 Other Generative Models and Discussion (Pros & Cons)

The spatial and temporal heterogeneous Bass model (STH-Bass) proposed by [216] borrowed the idea of the Bass model [18] that describes the process of how new products get adopted in a population to predict popularity of tweets. Compared to the original Bass model, there is an additional consideration of spatial-temporal heterogeneity, which is more applicable for tweet popularity prediction rather than predicting the sales of product. Subsequently, [40] used a topic-oriented feature combined Bass model with association analysis to make the popularity prediction.

The Bayesian approach to predict the popularity of tweets in [234] takes features into consideration, including time-series and retweet cascade graphs. In [132], the popularity growth was approached as a sequence of linear and non-linear phases and the proposed model does not rely

on microscopic information, while the prediction can be made efficiently without training. Other related works include: [169], proposing LMPP model to predict the hashtag popularity by modeling the hashtag-tweet reinforcement and inter-hashtag competitions; CHESS [181], predicting the popularity (watch time) of Facebook videos through an efficient and scalable Hawkes process; [16], using Hawkes process with survival theory to predict the popularity of tweets and papers; and [203], which devised MIC2MAC to link microscopic event data to macroscopic popularity inference, where Hawkes process and jump stochastic differential equation are used for prediction.

Generative models generally do not need heavy feature engineering and are inherently interpretable. They mainly rely on time-series data, and their predictions can be made in real-time, once the model has been prepared and the parameters have been estimated [242]. However, their performance has been questioned [28, 72], and they are often easily influenced by outliers [141]. In addition, generative models usually make strong assumptions on fixed parameters, which limits their generality and model expressiveness [50]. Moreover, the complex underlying mechanisms governing the success of cascades are underestimated and simplified, in order to simulate/reproduce the cascade diffusion processes. Finally, most of generative models are network-agnostic – i.e., they fail to model the important structural information that could help understand the process/paths of information diffusion. Thus, despite their efficiency and interpretability, generative models are less powerful in making precise predictions [28].

## 5 DEEP LEARNING MODELS

The renaissance of neural networks in recent years has spurred a number of deep learning based prediction models. Deep neural networks are often shown to be more effective than linear models [181]. For example, models based on RNNs do not rely on explicit assumptions of the diffusion of cascades and are more flexible to capture the temporal dependencies in cascades [202]. Models based on graph representation learning do not require laborious hand-crafted features from the underlying graphs of cascades, e.g., the specific designs of node influence and community detection.

One of the early approaches using representation learning and deep learning techniques is [91], focusing on predicting the number of views of Flickr images. On one hand, several hand-crafted low- and high-level social and visual features were designed (e.g., past success, number of contacts, colors, and textures). On the other hand, CNNs were utilized to learn the representations of images with the last hidden layer of the model.

DeepCas [119] is the first graph representation learning based method for modeling and predicting the popularity of information cascades. It borrows the idea of DeepWalk [153] to sample the cascade graphs with random walks. The sampled node sequences are then fed into a bidirectional gated recurrent units (Bi-GRU) [41], along with attention mechanism [10], to obtain the node embeddings. The prediction was made in an end-to-end manner. Subsequently, DeepCas was extended to DCGT [118] by incorporating the contents associated with the nodes in the cascade.

DeepHawkes [28] attempts to unify the advantages of generative models and deep learning techniques. In particular, three crucial concepts of Hawkes process – user influence, self-exciting mechanism, and time decay effect – are analogously transformed into the DeepHawkes model. Similar to DeepCas, user representations are learned in an end-to-end way. However, rather than directly modeling the structural patterns of cascade graphs – given a piece of information and its diffusion trajectories – DeepHawkes uses GRU, sum pooling, and non-parametric time kernel to aggregate the contributions of early adopters.

The attention based neural popularity prediction (ANPP) model, which utilizes GloVe [152] to embed the words of tweets, and leverages node2vec [68] to encode user graph, was presented in [34]. GRU with attention mechanism is used to aggregate the learned embeddings, as well as the time series feature vectors. Deep temporal context networks (DTCN) [207] predict the popularity

Table 7. Deep learning models.

Reference	Venue	Model	Building blocks
Li et al. [119]	WWW '17	DeepCas	Random Walks, GRU, Attention
Chen et al. [34]	ISI '17	ANPP	node2vec, GloVe, GRU, Attention
Wu et al. [207]	IJCAI '17	DTCN	ResNet, LSTM, Attention
Cao et al. [28]	CIKM '17	DeepHawkes	GRU, Pooling, Non-Parametric Time Kernel
Zhang et al. [239]	WWW '18	UHAN	VGGNet, LSTM, Hierarchical Attention
Wang et al. [199]	DASFAA '18	MOOD	Memory Network, Attention, Tensor Factorization
Wang et al. [200]	IJCAI '18	UMAN	Convolution, LSTM, Attention, User Memory
Bielksi et al. [20]	ACCESS'18	B&T	ResNet, Self-Attention, Pooling, GloVe, LSTM
Liao et al. [123]	AAAI '19	DFTC	LSTM, 1-D CNN, Attention, HAN
Chen et al. [37]	ICDE '19	CasCN	GCN, LSTM, Time Decay, Sum Pooling
Zhao et al. [243]	PAKDD '19	KB-PPN	Knowledge Base, LSTM, Gate/Attention Mechanisms
Chen et al. [36]	SIGIR '19	DMT-LIC	Multi-Task Learning, GCN, LSTM, Shared Gate
Yang et al. [218]	IJCAI '19	FOREST	Reinforcement Learning, GRU, DeepWalk
Huang et al. [83]	ACCESS '19	Cascade2vec	Graph Perception Network, GRU, Attention, Pooling
Cao et al. [29]	WSDM '20	Coupled-GNN	GNNs, Influence/State Gate Mechanisms, DeepWalk
Zhou et al. [245]	INFOCOM '20	VaCas	Spectral Graph Wavelets, Hierarchical VAEs, GRU

of Flickr images by jointly learning the user/photo embeddings, temporal context of resharing sequences, and multiple time-scale attention mechanism. Specifically, ResNet [76] and LSTM [79] are utilized to model the visual dependencies and temporal dependencies, respectively.

User-guided hierarchical attention network (UHAN) [239] considered both visual and textual modalities of Flickr images by characterizing three different representations, i.e., visual representation pre-trained from VGGNet [174], textual representation encoded by LSTM, and user representation learned under supervision. Then user-guided intra- and inter-attention mechanisms were used to jointly learn the importance over the aforementioned two modalities. Similarly, UMAN [200] focuses on texts of Douban and Twitter events, and learns two sequential correlations of textual content similarities to predict popularity.

The deep fusion of temporal processes and content features (DFTC) model [123] predicts the popularity of WeChat online articles. It first uses an LSTM to model the long-term growth trends of a cascade, and then utilizes 1-D CNN to capture the short-term fluctuations. Article content is modeled by a hierarchical attention network [224]. The final prediction is made by an attentive fusion of previously learned representations.

Recurrent cascades convolutional network (CasCN) [37] samples a cascade graph as a series of sequential subcascades and adopts a dynamic multi-directional GCN [94] to learn structural information of cascades. In addition, LSTM and time decay effects are utilized to model the temporal information of cascades. The subsequent work [36] addresses the activation prediction and popularity prediction problems jointly within a deep multi-task learning framework. To that end, attention and gated mechanisms were utilized, combined with a shared-representation layer, to capture the spatial-temporal dynamics of cascades.

Cascade2vec [83] is another GCN based cascade prediction model, where the representations of cascades are learned by a specifically designed graph preception network with residual and attention components. Coupled-GNNs [29] leveraged two specifically designed GNNs – one for node states and the other one for influence spread – to capture the cascading effect.

More recently, variational cascade graph learning neural networks (VaCas) [245] integrated graph wavelets, hierarchical variational autoencoders (VAEs) [93], and Bi-GRUs to learn the structures of

cascade graphs. Both node- and cascade-level diffusion uncertainties, as well as the contextualized user behaviors, were modeled in an end-to-end deep learning framework.

There are some works focusing on different aspects of performance improvement. For example, KB-PPN [243] leverages the embeddings of knowledge base entities and their neighbors to enhance the popularity prediction based on a LSTM network. FOREST [218] first learns an RNN-based microscopic cascade prediction model, and then performs a series of simulations guided by reinforcement learning to predict the macroscopic popularity of observed cascades.

## 5.1 Discussion

Table 7 summarizes the deep learning based cascade models and their main building blocks. Generally, these works learn structural information of cascades with various graph representation techniques, e.g., unsupervised network embedding [27] or (semi-supervised) graph neural networks [246], while capturing the time series and temporal characteristics of participation/citations with an RNN or its variants. The user/item attributes, as well as their content, can be learned with various deep learning techniques according to the data types, e.g., texts and images. Compared to feature-based models, these works do not require heavy feature engineering and can capture non-linear features of both user/item content and popularity accumulation trend. Deep learning approaches, especially the graph representation learning techniques, are of particular interest for learning the evolving patterns (both microscopic and macroscopic) of cascades, compared to the generative models. However, one main drawback of deep learning models is the lack of model interpretability owing to the “black-box” nature of neural networks.

## 6 OPEN CHALLENGES AND OPPORTUNITIES

We now discuss some challenges and aspects that have not been sufficiently studied in cascade modeling and information diffusion prediction, specifically: (i) predictability and interpretability; (ii) external stimuli; and (iii) cooperation and competition.

### 6.1 Predictability and Interpretability

The prediction of popularity from information cascades – no matter how the problem is reformulated, which datasets are used for evaluation, and even what metrics are selected as evaluation protocols – is largely subjective to researchers’ design (a.k.a. researcher degrees of freedom [80]). Despite the many publications and significant progress in this field, some fundamental questions have yet to be answered. For example, can we really predict the popularity of cascades? If the answer is yes, then to what extent can a particular cascade be predicted? Or, what is the glass-ceiling performance a model can achieve? To effortlessly adapt to new emerging information (e.g., online fake news and rumors) and new information propagation platforms, we may want to exploit existing knowledge and methods. Then, what conclusions drawn from previous works are applicable to another dataset/scenario? Can we build general cascade datasets that can be used for learning basic knowledge/features regarding various types of information items, such as the ImageNet and Wikipedia in computer vision and natural language processing? Answering these questions is a non-trivial task and requires more research as well as more standardized open datasets – the latter, however, need to obey the privacy issues of many online social networks.

By reviewing different groups of prediction models, we have identified the advantages and limitations of existing methods for predicting popularity – e.g., the design and selection of features vary greatly for different cascade formulations and information items, while their effectivenesses and universalities are not generalizable from one to another. Furthermore, the assumptions of diffusion mechanisms made by generative models are subject to specific scenarios, while the prediction of deep learning models can not be properly explained. To better understand the predictions of existing

models, a standardized comparison environment should be emphasized, e.g., carefully selecting the appropriate evaluation protocols for classification/regression and using publicly available datasets, etc., in order to avoid overestimating (or underestimating) the performance of a particular model.

Other than purely pursuing prediction performance, or seeing at it in a new light, explaining the behavior of predictions may enable us to better understand the disciplines governing the success of information items, so as to maximize social and business benefits through mastering and even managing the information diffusion process. To that end, it is desirable to unveil the inner-context driven mechanisms that guide information to spread further. A few examples include: investigating which topology of the diffusion network expands faster [38], the temporal and textual characteristics of burst public events and fake news and why they are traveled fast [148, 193], and the visual explanations of why some images/videos become popular quickly and why others do not [91], etc. Successfully explaining these phenomena will both improve the popularity predictions, and help design more powerful decision making systems to detect and bound viral marketing/advertising, rumor/fake news spreading, public opinion control, epidemic prevention, etc.

## 6.2 External Stimuli

External stimuli are one of the primary causes that lead the prediction of information cascades to be unexplainable, unreliable, and finally maybe impossible. Compared to endogenous stimuli like in-site search and share [55], external (or endogenous) stimuli are uncertain and unforeseeable events. For example, a normal cascade with little fluctuations in its early stage may suddenly become popular due to some external stimuli which significantly increases the exposure/adoption rate of that information. As a previous work [147] shows, about 71% of the information in Twitter is diffused in the network without external influence, while the remaining 29% of tweets have been significantly affected and even manipulated by exogenous force.

Retrieval of knowledge and propagation patterns from one platform to help the prediction of items in another platform with cross-domain real-time transfer learning (e.g., learn tweet spreading to enhance the performance of predicting YouTube video views) was studied in [166]. Another example is SoVP [122], which learned the shared diffusion behavior in social networks to strengthen the ability to predict video popularity. The rationale behind these works is that the trends appearing on Twitter would later drive users to search and watch related videos in YouTube. Moreover, burst events are more likely to first appear on microblogging platforms like Twitter/Weibo and then spread to other kinds of platforms such as newspapers and video sharing sites. This inspired later works to predict popularity in one domain via resorting to other information sources and dissemination platforms to model external stimuli responsible for the popularity [1, 159, 228]. More such examples can be found in [150] and [192], which retrieve information from Twitter and YouTube to predict the “ratings” and “views” of movies in IMDB. In [30], information from Facebook and Twitter were collected to predict the views of articles. Another recent work [142] uses RNNs to model asynchronous media streams of YouTube videos.

Modeling external stimuli can significantly enrich data diversity and improve model robustness. Actually, many web sites now provide extra information that attracts external stimuli, e.g., the information of referrers/URLs that may bring new views/adoptions. For YouTube videos, a referrer can be a hyperlink embedded in a tweet pointing to a video. This provides clues to trace the sources of external stimuli. The top-10 referrers and the number of views each referrer brings as features to predict the evolving trends of videos were used in [52, 55]. This anatomy of sources of popularity gives us more insights into the future evolution of cascades [26], e.g., how many external sources are contributing to the popularity of cascades, which continuously existed and have already been diminished or vanished.



### 6.3 Cooperation and Competition

Information cascades are not isolated but affect each other and are often interwoven via interactions, cooperation and competition, when spreading among users. Cooperation means that the diffusion of an information item is accelerated by other item(s), e.g., if one thing was reported by many sources many times, then this information is more credible and has a higher probability of adoption [145]. In contrast, due to the limited attention span of users [204], information items (especially those with very similar content) could restrict each other's diffusion [42]. Thus, if cooperation happens, the expectation of an information item's future popularity should be larger than normal cases. Also in the case of competition, items with similar contents would inevitably face the problem of low popularity or at least reduced popularity as their competitors increase.

Existing works on this topic mainly focus on discovering cooperation/competition during diffusion, but rarely on analyzing reciprocity and its effect on popularity prediction. For example, interaction matrix between items to facilitate the prediction of tweets/hashtags was used in [98]. Hawkes processes were adopted to model cooperation and competition between cascades in social networks in [227, 236]. Modeling the interactions between information items/cascades, especially the cooperation and competition among them, can benefit popularity prediction and deepen our understanding of information diffusion in various social scenarios: e.g., figuring out when and how cooperation/competition happens and identifying the best timing and content of items to enable distributors to avoid competitions and meet more cooperation.

## 7 CONCLUSION

We provided a broad overview of the field of information diffusion modeling and popularity prediction and proposed a taxonomy to categorize related literature, highlighting a number of influential papers/approaches in each category. We also discussed the advantages and disadvantages of different types of methods over time and outlined some challenges and open research problems. There exists a vast literature and growing research on this topic – and we systematically covered, in certain detail, selected representative works, with a hope that this knowledge can be used as a stepping-stone for future work and/or evaluations in this field.

## ACKNOWLEDGMENTS

We thank all doctors, nurses and medical personnel who are on the front line of the war against COVID-19. This work was supported by National Natural Science Foundation of China under Grant No. 61602097 and No. 61472064, and NSF under Grant CNS 1646107.

## REFERENCES

- [1] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. 2014. Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In *WSDM*. ACM, 593–602.
- [2] Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. 2016. Information evolution in social networks. In *WSDM*. 473–482.
- [3] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A peek into the future: Predicting the evolution of popularity in user generated content. In *WSDM*. ACM, 607–616.
- [4] Sultan Alzahrani, Saud Alashri, Anvesh Reddy Koppela, Hasan Davulcu, and Ismail Toroslu. 2015. A network-based model for predicting hashtag breakouts in Twitter. In *SBP*. Springer, 3–12.
- [5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *KDD*. ACM, 7–15.
- [6] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, and Mitul Tiwari. 2015. Global diffusion via cascading invitations: Structure, growth, and homophily. In *WWW*. 66–76.
- [7] Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *HLT-NAACL*. Association for Computational Linguistics, 602–606.

- [8] Sitaram Asur, Bernardo A Huberman, Gabor Szabo, and Chunyan Wang. 2011. Trends in social media: Persistence and decay. In *ICWSM*.
- [9] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In *KDD*. ACM, 44–54.
- [10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.
- [11] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone’s an influencer: quantifying influence on twitter. In *WSDM*. ACM, 65–74.
- [12] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.
- [13] Peng Bao, Hua-Wei Shen, Wei Chen, and Xue-Qi Cheng. 2013. Cumulative effect in information diffusion: empirical study on a microblogging network. *PloS One* 8, 10 (2013), e76027.
- [14] Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. 2013. Popularity prediction in microblogging network: a case study on sina weibo. In *WWW*. ACM, 177–178.
- [15] Peng Bao, Hua-Wei Shen, Xiaolong Jin, and Xue-Qi Cheng. 2015. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes. In *WWW*. ACM, 9–10.
- [16] Peng Bao and Xiaoxia Zhang. 2017. Uncovering and predicting the dynamic process of collective attention with survival theory. *Scientific Reports* 7, 1 (2017), 1–8.
- [17] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207.
- [18] Frank M Bass. 1969. A new product growth for model consumer durables. *Management science* 15, 5 (1969), 215–227.
- [19] Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2014. Predicting trending messages and diffusion participants in microblogging network. In *SIGIR*. ACM, 537–546.
- [20] Adam Bielski and Tomasz Trzcinski. 2018. Understanding Multimodal Popularity Prediction of Social Media Videos With Self-Attention. *IEEE Access* 6 (2018), 74277–74287.
- [21] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3, Jan (2003), 993–1022.
- [22] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012).
- [23] Siddharth Bora, Harvineet Singh, Anirban Sen, Amitabha Bagchi, and Parag Singla. 2015. On the role of conductance, geography and topology in predicting hashtag virality. *SNAM* 5, 1 (2015), 57.
- [24] Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. 2012. The untold story of the clones: content-agnostic factors that impact YouTube video popularity. In *KDD*. ACM, 1186–1194.
- [25] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *MM*. 223–232.
- [26] Tom Broxton, Yannet Interian, Jon Vaver, and Mirjam Wattenhofer. 2013. Catching a viral video. *JHS* 40, 2 (2013), 241–259.
- [27] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *TKDE* 30, 9 (2018), 1616–1637.
- [28] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. DeepHawkes: Bridging the gap between prediction and understanding of information cascades. In *CIKM*. ACM, 1149–1158.
- [29] Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. Popularity Prediction on Social Platforms with Coupled Graph Neural Networks. In *WSDM*. 70–78.
- [30] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the life cycle of online news stories using social media reactions. In *CSCW*. ACM, 211–223.
- [31] Damon Centola and Michael Macy. 2007. Complex contagions and the weakness of long ties. *Amer. J. Sociology* 113, 3 (2007), 702–734.
- [32] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2009. Analyzing the video popularity characteristics of large-scale user generated content systems. *TON* 17, 5 (2009), 1357–1370.
- [33] Feng Chen, Wai Hong Tan, et al. 2018. Marked self-exciting point process modelling of information diffusion on Twitter. *The Annals of Applied Statistics* 12, 4 (2018), 2175–2196.
- [34] Guandan Chen, Qingchao Kong, and Wenji Mao. 2017. An attention-based neural popularity prediction model for social media events. In *ISI*. IEEE, 161–163.
- [35] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: predicting the popularity of micro-videos via a transductive model. In *MM*. 898–907.
- [36] Xueqin Chen, Kunpeng Zhang, Fan Zhou, Goce Trajcevski, Ting Zhong, and Fengli Zhang. 2019. Information Cascades Modeling via Deep Multi-Task Learning. In *SIGIR*. ACM, 885–888.
- [37] Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. 2019. Information Diffusion Prediction via Recurrent Cascades Convolution. In *ICDE*. IEEE, 770–781.

- [38] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *WWW*. ACM, 925–936.
- [39] Xu Cheng, Jiangchuan Liu, and Cameron Dale. 2013. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. *TMM* 15, 5 (2013), 1184–1194.
- [40] Quanquan Chu, Zhenhao Cao, Xiaofeng Gao, Peng He, Qianni Deng, and Guihai Chen. 2018. Cease with Bass: A Framework for Real-Time Topic Detection and Popularity Prediction Based on Long-Text Contents. In *CSoNet*. 25–29.
- [41] Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555* (2014).
- [42] Michele Coscia. 2014. Average is boring: How similarity kills a meme’s success. *Scientific Reports* 4 (2014), 6477.
- [43] David R Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 2 (1972), 187–202.
- [44] Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS* 105, 41 (2008), 15649–15653.
- [45] Peng Cui, Shifei Jin, Linyun Yu, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Cascading outbreak prediction in networks: a data-driven approach. In *KDD*. ACM, 901–909.
- [46] Wanying Ding, Yue Shang, Lifan Guo, Xiaohua Hu, Rui Yan, and Tingting He. 2015. Video popularity prediction by sentiment propagation via implicit network. In *CIKM*. ACM, 1621–1630.
- [47] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2015. Will this paper increase your h-index?: Scientific impact prediction. In *WSDM*. ACM, 149–158.
- [48] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2016. Can scientific impact be predicted? *IEEE Transactions on Big Data* 2, 1 (2016), 18–30.
- [49] P Alex Dow, Lada A Adamic, and Adrien Friggeri. 2013. The anatomy of large facebook cascades. In *ICWSM*.
- [50] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*. 1555–1564.
- [51] Milad Eftekhari, Yashar Ganjali, and Nick Koudas. 2013. Information cascade at group scale. In *KDD*. ACM, 401–409.
- [52] Flavio Figueiredo. 2013. On the prediction of popularity of trends and hits for user generated videos. In *WSDM*.
- [53] Flavio Figueiredo, Jussara M Almeida, Fabrício Benevenuto, and Krishna P Gummadi. 2014. Does content determine information popularity in social media?: A case study of youtube videos’ content and their popularity. In *CHI*. ACM, 979–982.
- [54] Flavio Figueiredo, Jussara M Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. 2014. On the dynamics of social media popularity: A YouTube case study. *TOIT* 14, 4 (2014), 24.
- [55] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M Almeida. 2011. The tube over time: characterizing popularity growth of youtube videos. In *WSDM*. ACM, 745–754.
- [56] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. 2010. Outtweeting the twitterers-predicting information cascades in microblogs. *WOSN* 10 (2010), 3–11.
- [57] Jinhua Gao, Huawei Shen, Shenghua Liu, and Xueqi Cheng. 2016. Modeling and predicting retweeting dynamics via a mixture process. In *WWW*. 33–34.
- [58] Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Effective and effortless features for popularity prediction in microblogging network. In *WWW*. ACM, 269–270.
- [59] Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Popularity prediction in microblogging network. In *Asia-Pacific Web Conference*. Springer, 379–390.
- [60] Shuai Gao, Jun Ma, and Zhumin Chen. 2015. Modeling and predicting retweeting dynamics on microblogging platforms. In *WSDM*. ACM, 107–116.
- [61] Sheng Gao, Huacan Pang, Patrick Gallinari, Jun Guo, and Nei Kato. 2017. A Novel Embedding Method for Information Diffusion Prediction in Social Network Big Data. *TII* 13, 4 (2017), 2097–2105.
- [62] Xiaofeng Gao, Zhenhao Cao, Sha Li, Bin Yao, Guihai Chen, and Shaojie Tang. 2019. Taxonomy and Evaluation for Microblog Popularity Prediction. *TKDD* 13, 2 (2019), 15.
- [63] Diego Garlaschelli and Maria I Loffredo. 2004. Patterns of link reciprocity in directed networks. *Physical Review Letters* 93, 26 (2004), 268701.
- [64] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *MM*. ACM, 907–910.
- [65] Eric Gilbert. 2013. Widespread underprovision on Reddit. In *CSCW*. ACM, 803–808.
- [66] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. 2007. Youtube traffic characterization: a view from the edge. In *IMC*. ACM, 15–28.
- [67] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.
- [68] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. 855–864.

- [69] Adrien Guille and Hakim Hacid. 2012. A predictive model for the temporal dynamics of information diffusion in online social networks. In *WWW*. ACM, 1145–1152.
- [70] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record* 42, 2 (2013), 17–28.
- [71] Ruocheng Guo, Elham Shaabani, Abhinav Bhatnagar, and Paulo Shakarian. 2015. Toward order-of-magnitude cascade prediction. In *ASONAM*. 1610–1613.
- [72] Ruocheng Guo and Paulo Shakarian. 2016. A comparison of methods for cascade prediction. In *ASONAM*. 591–598.
- [73] Manish Gupta, Jing Gao, ChengXiang Zhai, and Jiawei Han. 2012. Predicting future popularity trend of events in microblogging platforms. *ASIST* 49, 1 (2012), 1–10.
- [74] Gonca Gürsun, Mark Crovella, and Ibrahim Matta. 2011. Describing and forecasting video access patterns. In *INFOCOM*. IEEE, 16–20.
- [75] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- [76] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
- [77] Xiangnan He, Ming Gao, Min-Yen Kan, Yiqun Liu, and Kazunari Sugiyama. 2014. Predicting the popularity of web 2.0 items based on user comments. In *SIGIR*. ACM, 233–242.
- [78] Minh X. Hoang, Xuan-Hong Dang, Xiang Wu, Zhenyu Yan, and Ambuj K. Singh. 2017. GPPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks. In *WWW*. 725–733.
- [79] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- [80] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science* 355, 6324 (2017), 486–488.
- [81] Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting popular messages in twitter. In *WWW Companion*. ACM, 57–58.
- [82] Zhiting Hu, Junjie Yao, Bin Cui, and Eric Xing. 2015. Community level diffusion extraction. In *SIGMOD*. ACM, 1555–1569.
- [83] Zhenhua Huang, Zhenyu Wang, and Rui Zhang. 2019. Cascade2vec: Learning Dynamic Cascade Representation by Recurrent Graph Neural Networks. *IEEE Access* 7 (2019), 144800–144812.
- [84] Bernardo A Huberman, Daniel M Romero, and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *arXiv:0812.1045* (2008).
- [85] Salman Jamali and Huzefa Rangwala. 2009. Digging digg: Comment mining, popularity prediction, and social network analysis. In *WISM*. IEEE, 32–38.
- [86] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. 2013. Analyzing and predicting viral tweets. In *WWW*. ACM, 657–664.
- [87] Adele Lu Jia, Siqi Shen, Dongsheng Li, and Shengling Chen. 2018. Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features. *Computer Networks* 140 (2018), 112–125.
- [88] Xin Jin, Andrew Gallagher, Liangliang Cao, Jiebo Luo, and Jiawei Han. 2010. The wisdom of social multimedia: using flickr for prediction and forecast. In *MM*. ACM, 1235–1244.
- [89] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. ACM, 137–146.
- [90] Elham Khabiri, Chiao-Fang Hsu, and James Caverlee. 2009. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *ICWSM*.
- [91] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *WWW*. ACM, 867–876.
- [92] Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. 2011. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *CIT*. IEEE, 449–454.
- [93] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.
- [94] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.
- [95] Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *JACM* 46, 5 (1999), 604–632.
- [96] Ryota Kobayashi and Renaud Lambiotte. 2016. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *ICWSM*.
- [97] Qingchao Kong, Wenji Mao, Guandan Chen, and Daniel Zeng. 2018. Exploring trends and patterns of popularity stage evolution in social media. *IEEE Transactions on Systems, Man, and Cybernetics (SMC)* (2018).
- [98] Qingchao Kong, Wenji Mao, and Chunyang Liu. 2016. Popularity prediction based on interactions of online contents. In *CCIS*. IEEE, 1–5.
- [99] Qingchao Kong, Wenji Mao, Daniel Zeng, and Lei Wang. 2014. Predicting popularity of forum threads for public events security. In *2014 IEEE Joint Intelligence and Security Informatics Conference*. IEEE, 99–106.

- [100] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. 2020. Modeling Information Cascades with Self-exciting Processes via Generalized Epidemic Models. In *WSDM*. 286–294.
- [101] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. 2014. Predicting bursts and popularity of hashtags in real-time. In *SIGIR*. ACM, 927–930.
- [102] Shoubin Kong, Fei Ye, Ling Feng, and Zhe Zhao. 2015. Towards the prediction problems of bursting hashtags on Twitter. *Journal of the Association for Information Science and Technology* 66, 12 (2015), 2566–2579.
- [103] Siddharth Krishnan, Patrick Butler, Ravi Tandon, Jure Leskovec, and Naren Ramakrishnan. 2016. Seeing the forest for the trees: new approaches to forecasting cascades. In *WebSci*. 249–258.
- [104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
- [105] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. 2012. Prediction of retweet cascade size over time. In *CIKM*. ACM, 2335–2338.
- [106] Andrey Kupavskii, Alexey Umnov, Gleb Gusev, and Pavel Serdyukov. 2013. Predicting the audience size of a tweet. In *ICWSM*.
- [107] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *WWW*. ACM, 591–600.
- [108] Himabindu Lakkaraju and Jitendra Ajmera. 2011. Attention prediction on social media brand pages. In *CIKM*. ACM, 2157–2160.
- [109] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What’s in a name? understanding the interplay between titles, content, and communities in social media. In *ICWSM*.
- [110] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks* 3361, 10 (1995), 1995.
- [111] Jong Gun Lee, Sue Moon, and Kave Salamatian. 2010. An approach to model and predict the popularity of online contents with explanatory factors. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 623–630.
- [112] Jong Gun Lee, Sue Moon, and Kave Salamatian. 2012. Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. *Neurocomputing* 76, 1 (2012), 134–145.
- [113] Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *WWW*. ACM, 621–630.
- [114] Jure Leskovec, Lada Adamic, and Bernardo A Huberman. 2007. The dynamics of viral marketing. *TWEB* 1, 1 (2007), 5.
- [115] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*. ACM, 497–506.
- [116] Jure Leskovec, Kevin J Lang, and Michael Mahoney. 2010. Empirical comparison of algorithms for network community detection. In *WWW*. ACM, 631–640.
- [117] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2007. Cascading behavior in large blog graphs: Patterns and a model. *Society of Applied and Industrial Mathematics: Data Mining* (2007), 551–556.
- [118] Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei. 2018. Joint modeling of text and networks for cascade prediction. In *ICWSM*. 640–643.
- [119] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. Deepcas: An end-to-end predictor of information cascades. In *WWW*. 577–586.
- [120] Cheng-Te Li, Yu-Jen Lin, and Mi-Yen Yeh. 2015. The roles of network communities in social information diffusion. In *ICBD*.
- [121] Dong Li, Shengping Zhang, Xin Sun, Huiyu Zhou, Sheng Li, and Xuelong Li. 2017. Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction. *TKDE* 29, 9 (2017), 1985–1997.
- [122] Haitao Li, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, and Ke Xu. 2013. On popularity prediction of videos shared in online social networks. In *CIKM*. ACM, 169–178.
- [123] Dongliang Liao, Jin Xu, Gongfu Li, Weiqing Huang, Weiqing Liu, and Jing Li. 2019. Popularity Prediction on Online Articles with Deep Fusion of Temporal Process and Content Features. In *AAAI*. 200–207.
- [124] Shuyang Lin, Xiangnan Kong, and Philip S Yu. 2013. Predicting trends in social networks via dynamic activeness model. In *CIKM*. ACM, 1661–1666.
- [125] Weiwei Liu, Zhi-Hong Deng, Xiuwen Gong, Frank Jiang, and Ivor W Tsang. 2015. Effectively predicting whether and when a topic will become prevalent in a social network. In *AAAI*.
- [126] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60, 2 (2004), 91–110.
- [127] Xiaoyan Lu and Boleslaw Szymanski. 2017. Predicting viral news events in online media. In *IPDPSW*. 1447–1456.
- [128] Xinjiang Lu, Zhiwen Yu, Bin Guo, and Xingshe Zhou. 2014. Predicting the content dissemination trends by repost behavior modeling in mobile social networks. *JNCA* 42 (2014), 197–207.

- [129] Yunfei Lu, Linyun Yu, Tianyang Zhang, Chengxi Zang, Peng Cui, Chaoming Song, and Wenwu Zhu. 2018. Collective Human Behavior in Cascading System: Discovery, Modeling and Applications. In *ICDM*. IEEE, 297–306.
- [130] Zhunchen Luo and Xiao Liu. 2018. Real-time Scholarly Retweeting Prediction System. In *COLING: System Demonstrations*. 25–29.
- [131] Jinna Lv, Wu Liu, Meng Zhang, He Gong, Bin Wu, and Huadong Ma. 2017. Multi-feature fusion for predicting social media popularity. In *MM*. 1883–1888.
- [132] Ilias N Lymperopoulos. 2016. Predicting the popularity growth of online content: Model and algorithm. *Information Sciences* 369 (2016), 585–613.
- [133] Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will this# hashtag be popular tomorrow?. In *SIGIR*. ACM, 1173–1174.
- [134] Zongyang Ma, Aixin Sun, and Gao Cong. 2013. On predicting the popularity of newly emerging hashtags in Twitter. *JASIST* 64, 7 (2013), 1399–1410.
- [135] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. 2016. Exploring limits to prediction in complex social systems. In *WWW*. 683–694.
- [136] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. 2012. Rise and fall patterns of information diffusion: model and implications. In *KDD*. ACM, 6–14.
- [137] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. Nobody comes here anymore, it’s too crowded; predicting image popularity on flickr. In *ICMR*. ACM, 385.
- [138] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.
- [139] Matúš Medo, Manuel S Mariani, An Zeng, and Yi-Cheng Zhang. 2016. Identification and impact of discoverers in online social systems. *Scientific Reports* 6 (2016), 34218.
- [140] Rupert G Miller Jr. 2011. *Survival analysis*. Vol. 66. John Wiley & Sons.
- [141] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. 2016. Feature driven and point process approaches for popularity prediction. In *CIKM*. ACM, 1069–1078.
- [142] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *ICWSM*. 201–210.
- [143] Nuno Moniz and Luís Torgo. 2019. A review on web content popularity prediction: Issues and open challenges. *Online Social Networks and Media* 12 (2019), 1–20.
- [144] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
- [145] Seth A Myers and Jure Leskovec. 2012. Clash of the contagions: Cooperation and competition in information diffusion. In *ICDM*. IEEE, 539–548.
- [146] Seth A Myers and Jure Leskovec. 2014. The bursty dynamics of the twitter information network. In *WWW*. ACM, 913–924.
- [147] Seth A Myers, Chenguang Zhu, and Jure Leskovec. 2012. Information diffusion and external influence in networks. In *KDD*. ACM, 33–41.
- [148] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci*. ACM, 8.
- [149] Amandianeze O Nwana, Salman Avestimehr, and Tsuhan Chen. 2013. A latent social approach to youtube popularity prediction. In *GLOBECOM*. IEEE, 3138–3144.
- [150] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten De Rijke. 2012. Predicting imdb movie ratings using social media. In *ECIR*. Springer, 503–507.
- [151] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI* 7 (2002), 971–987.
- [152] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*. 1532–1543.
- [153] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *KDD*.
- [154] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2011. Rt to win! predicting message propagation in twitter. In *ICWSM*.
- [155] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In *WSDM*. ACM, 365–374.
- [156] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E Hopcroft. 2016. The lifecycle and cascade of wechat social messaging groups. In *WWW*. 311–320.
- [157] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. 2018. Deepinf: Social influence prediction with deep learning. In *KDD*. ACM, 2110–2119.
- [158] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. 2018. SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations. In *WWW*. 419–428.



- [159] Marian-Andrei Rizoio, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be hip: Hawkes intensity processes for social media popularity. In *WWW*. 735–744.
- [160] Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2016. Predicting news popularity by mining online discussions. In *WWW Companion*. 737–742.
- [161] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.
- [162] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. 2011. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 18–33.
- [163] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW*. ACM, 695–704.
- [164] Daniel M Romero, Chenhao Tan, and Johan Ugander. 2013. On the interplay between social and topical structure. In *ICWSM*.
- [165] Matthew Rowe, Sofia Angeletou, and Harith Alani. 2011. Predicting discussions on the social semantic web. In *ESWC*. Springer, 405–420.
- [166] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. 2013. Towards cross-domain learning for social video popularity prediction. *TMM* 15, 6 (2013), 1255–1267.
- [167] Yiye Ruan, Hemant Purohit, David Fuhry, Srinivasan Parthasarathy, and Amit P Sheth. 2012. Prediction of topic volume on twitter. In *WebSci*.
- [168] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.
- [169] Bidisha Samanta, Abir De, Abhijnan Chakraborty, and Niloy Ganguly. 2017. LMPP: A Large Margin Point Process Combining Reinforcement and Competition for Modeling Hashtag Popularity. In *IJCAI*. 2679–2685.
- [170] Zubair Shafiq and Alex Liu. 2017. Cascade size prediction in online social networks. In *Networking*. 1–9.
- [171] David A Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth F Churchill. 2011. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *ICWSM*.
- [172] Huawei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and predicting popularity dynamics via reinforced poisson processes. In *AAAI*.
- [173] Benjamin Shulman, Amit Sharma, and Dan Cosley. 2016. Predictability of popularity: Gaps between prediction and understanding. In *ICWSM*.
- [174] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
- [175] Stefan Stieglitz and Linh Dang-Xuan. 2012. Political communication and influence through microblogging—An empirical analysis of sentiment in Twitter messages and retweet behavior. In *HICSS*. IEEE, 3500–3509.
- [176] Greg Stoddard. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *ICWSM*.
- [177] Karthik Subbian, B. Aditya Prakash, and Lada Adamic. 2017. Detecting large reshare cascades in social networks. In *WWW*. 597–606.
- [178] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*. IEEE, 177–184.
- [179] Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Commun. ACM* 53, 8 (2010), 80–88.
- [180] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *ACL*. 175–185.
- [181] linpeng Tang, Qi Huang, Amit Puntambekar, Ymir Vigfusson, Wyatt Lloyd, and Kai Li. 2017. Popularity prediction of Facebook videos for higher quality streaming. In *USENIX ATC*. 111–123.
- [182] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. 2014. From popularity prediction to ranking online news. *SNAM* 4, 1 (2014), 174.
- [183] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 5, 1 (2014), 8.
- [184] Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. 2011. Predicting the popularity of online articles based on user comments. In *WIMS*. ACM, 67.
- [185] Luam Catao Totti, Felipe Almeida Costa, Sandra Avila, Eduardo Valle, Wagner Meira Jr, and Virgilio Almeida. 2014. The impact of visual attributes on online image diffusion. In *WebSci*. ACM, 42–51.
- [186] Tomasz Trzciński and Przemysław Rokita. 2017. Predicting popularity of online videos using support vector regression. *TMM* 19, 11 (2017), 2561–2570.
- [187] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2009. Predicting the volume of comments on online news stories. In *CIKM*. ACM, 1765–1768.
- [188] Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. 2010. News comments: Exploring, modeling, and online prediction. In *ECIR*. Springer, 191–203.

- [189] Sho Tsugawa. 2019. Empirical Analysis of the Relation between Community Structure and Cascading Retweet Diffusion. In *ICWSM*. 493–504.
- [190] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*. ACM, 643–652.
- [191] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *PNAS* 109, 16 (2012), 5962–5966.
- [192] David Vallet, Shlomo Berkovsky, Sebastien Ardon, Anirban Mahanti, and Mohamed Ali Kafaar. 2015. Characterizing and predicting viral-and-popular video content. In *CIKM*. ACM, 1591–1600.
- [193] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
- [194] Duy Q Vu, Arthur U Asuncion, Hunter David R, and Padhraic Smyth. 2011. Dynamic egocentric models for citation networks. In *ICML*. 857–864.
- [195] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.
- [196] Jia Wang, Vincent W Zheng, Zemin Liu, and Kevin Chen-Chuan Chang. 2017. Topological recurrent neural network for diffusion prediction. In *ICDM*. IEEE, 475–484.
- [197] Shoujin Wang, Liang Hu, Longbing Cao, Xiaoshui Huang, Defu Lian, and Wei Liu. 2018. Attention-based transactional context embedding for next-item recommendation. In *AAAI*. 2532–2539.
- [198] Senzhang Wang, Zhao Yan, Xia Hu, S Yu Philip, and Zhoujun Li. 2015. Burst time prediction in cascades. In *AAAI*.
- [199] Wen Wang, Wei Zhang, and Jun Wang. 2018. Factorization Meets Memory Network: Learning to Predict Activity Popularity. In *DASFAA*. 509–525.
- [200] Wen Wang, Wei Zhang, Jun Wang, Junchi Yan, and Hongyuan Zha. 2018. Learning Sequential Correlation for User Generated Textual Content Popularity Prediction. In *IJCAI*. 1625–1631.
- [201] Yongqing Wang, Huawei Shen, Shenghua Liu, and Xueqi Cheng. 2015. Learning User-Specific Latent Influence and Susceptibility from Information Cascades. In *AAAI*.
- [202] Yongqing Wang, Huawei Shen, Shenghua Liu, Jinhua Gao, and Xueqi Cheng. 2017. Cascade dynamics modeling with attention-based recurrent neural network. In *IJCAI*. 2985–2991.
- [203] Yichen Wang, Xiaojing Ye, Haomin Zhou, Hongyuan Zha, and Le Song. 2017. Linking micro event history to macro prediction in point process models. In *AISTATS*. 1375–1384.
- [204] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2 (2012), 335.
- [205] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2013. Virality prediction and community structure in social networks. *Scientific Reports* 3 (2013), 2522.
- [206] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting successful memes using network and community structure. In *ICWSM*.
- [207] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential prediction of social media popularity with deep temporal context networks. In *IJCAI*. 3062–3068.
- [208] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *AAAI*.
- [209] Qitian Wu, Yirui Gao, Xiaofeng Gao, Paul Weng, and Guihai Chen. 2019. Dual Sequential Prediction Models Linking Sequential Recommendation and Information Dissemination. In *KDD*. ACM, 447–457.
- [210] Siqi Wu, Marian-Andrei Rizoioiu, and Lexing Xie. 2018. Beyond views: Measuring and predicting engagement in online videos. In *ICWSM*. 434–443.
- [211] Siqi Wu, Marian-Andrei Rizoioiu, and Lexing Xie. 2019. Estimating Attention Flow in Online Video Network. *arXiv:1908.07123* (2019).
- [212] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. 2016. On Modeling and Predicting Individual Paper Citation Count over Time.. In *IJCAI*. 2676–2682.
- [213] Daniel Xie, Jiejun Xu, and Tsai-Ching Lu. 2017. What's trending tomorrow, today: Using early adopters to discover popular posts on Tumblr. In *BigData*. 2168–2176.
- [214] Wei Xie, Feida Zhu, Siyuan Liu, and Ke Wang. 2015. Modelling cascades over time in microblogs. In *ICBD*.
- [215] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *CIKM*. ACM, 1247–1252.
- [216] Yan Yan, Zhaowei Tan, Xiaofeng Gao, Shaojie Tang, and Guihai Chen. 2016. STH-Bass: A Spatial-temporal heterogeneous bass model to predict single-tweet popularity. In *DASFAA*. Springer, 18–32.
- [217] Cheng Yang, Maosong Sun, Haoran Liu, Shiyi Han, Zhiyuan Liu, and Huanbo Luan. 2019. Neural Diffusion Model for Microscopic Cascade Study. *TKDE* (2019).

- [218] Cheng Yang, Jian Tang, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. 2019. Multi-scale information diffusion prediction with reinforced recurrent networks. In *IJCAI*. 4033–4039.
- [219] Jiang Yang and Scott Counts. 2010. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*.
- [220] Jaewon Yang and Jure Leskovec. 2010. Modeling information diffusion in implicit networks. In *ICDM*. IEEE, 599–608.
- [221] Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *WSDM*. ACM, 177–186.
- [222] Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what@ you# tag: does the dual role affect hashtag adoption?. In *WWW*. ACM, 261–270.
- [223] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. 2010. Understanding retweeting behaviors in social networks. In *CIKM*. ACM, 1633–1636.
- [224] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*. 1480–1489.
- [225] Tae Yano and Noah A Smith. 2010. What’s worthy of comment? Content and comment volume in political blogs. In *ICWSM*.
- [226] Chengqi Yi, Yuanyuan Bao, and Yibo Xue. 2016. Mining the key predictors for event outbreaks in social networks. *Physica A: Statistical Mechanics and its Applications* 447 (2016), 247–260.
- [227] Eunae Yoo, Bin Gu, and Elliot Rabinovich. 2019. Competition and Coopetition among Social Media Content. In *HICSS*.
- [228] Honglin Yu, Lexing Xie, and Scott Sanner. 2014. Twitter-driven youtube views: Beyond individual influencers. In *MM*. ACM, 869–872.
- [229] Honglin Yu, Lexing Xie, and Scott Sanner. 2015. The lifecycle of a youtube video: Phases, content and popularity. In *ICWSM*.
- [230] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. 2015. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. In *ICDM*. IEEE, 559–568.
- [231] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. 2017. Uncovering and predicting the dynamic process of information cascades with survival model. *KAIS* 50, 2 (2017), 633–659.
- [232] Nicholas Jing Yuan, Yuan Zhong, Fuzheng Zhang, Xing Xie, Chin-Yew Lin, and Yong Rui. 2016. Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions. In *WSDM*. ACM, 3–12.
- [233] Amir Hassan Zadeh and Ramesh Sharda. 2014. Modeling brand post popularity dynamics in online social networks. *DSS* 65 (2014), 59–68.
- [234] Tauhid R Zaman, Emily B Fox, Eric T Bradlow, et al. 2014. A bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics* 8, 3 (2014), 1583–1611.
- [235] Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. 2010. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, Vol. 104. Citeseer, 17599–601.
- [236] Ali Zarezade, Ali Khodadadi, Mehrdad Farajtabar, Hamid R Rabiee, and Hongyuan Zha. 2017. Correlated cascades: Compete or cooperate. In *AAAI*.
- [237] Bolei Zhang, Zhuzhong Qian, and Sanglu Lu. 2016. Structure pattern analysis and cascade prediction in social networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 524–539.
- [238] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6, 1 (2019), 11.
- [239] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided Hierarchical Attention Network for Multi-modal Social Image Popularity Prediction. In *WWW*. 1277–1286.
- [240] Zi-Ke Zhang, Chuang Liu, Xiu-Xiu Zhan, Xin Lu, Chu-Xu Zhang, and Yi-Cheng Zhang. 2016. Dynamics of information diffusion and its applications on complex networks. *Physics Reports* 651 (2016), 1–34.
- [241] Liang Zhao, Jiangzhuo Chen, Feng Chen, Fang Jin, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Online flu epidemiological deep modeling on disease contact network. *GeoInformatica* (2019), 1–33.
- [242] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*. ACM, 1513–1522.
- [243] Wayne Xin Zhao, Hongjian Dou, Yuanpei Zhao, Daxiang Dong, and Ji-Rong Wen. 2019. Neural Network Based Popularity Prediction by Linking Online Content with Knowledge Bases. In *PAKDD*. Springer, 16–28.
- [244] Yunwei Zhao, Can Wang, Chi-Hung Chi, Kwok-Yan Lam, and Sen Wang. 2018. A comparative study of transactional and semantic approaches for predicting cascades on twitter. In *IJCAI*. 1212–1218.
- [245] Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, and Ting Zhong. 2020. Variational information diffusion for probabilistic cascades prediction. In *INFOCOM (In Press)*.
- [246] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv:1812.08434* (2018).

## A SUPPLEMENTARY MATERIAL

Here we list five tables corresponding to: feature-based models (Table 8 and 9), temporal (Table 10), structural (Table 11), and user/item features (Table 12), for the ease of reference.

Table 8. Feature-based models. Abbreviations: **Te.** – Temporal, **CS** – Cascade structure, **GG** – Global graph, **Us.** – User, **Co.** – Content.

Reference	Venue	Strategy	Formulation	Feature				
				Te.	CS	GG	Us.	Co.
Khabiri et al. [90]	ICWSM '09	Ex-ante	Classification				Y	Y
Tsagkias et al. [187]	CIKM '09	Ex-ante	Classification	Y			Y	Y
Jamali et al. [85]	WISM '09	Peeking	Both	Y	Y	Y	Y	Y
Lerman & Hogg [113]	WWW '10	Peeking	Both	Y		Y	Y	
Tsagkias et al. [188]	ECIR '10	Peeking	Regression	Y				
Yano & Smith [225]	ICWSM '10	Ex-ante	Classification					Y
Szabo & Huberman [179]	Comm. ACM '10	Peeking	Regression	Y				
Hong et al. [81]	WWW '11	Peeking	Classification	Y	Y	Y	Y	Y
Bakshy et al. [11]	WSDM '11	Ex-ante	Regression				Y	Y
Kim et al. [92]	CIT '11	Peeking	Classification	Y				
Gürsun et al. [74]	INFOCOM '11	Peeking	Regression	Y				
Tatar et al. [184]	WIMS '11	Peeking	Regression	Y				
Rowe et al. [165]	ESWC '11	Ex-ante	Regression				Y	Y
Naveed et al. [148]	WebSci '11	Ex-ante	Classification					Y
Petrovic et al. [154]	ICWSM '11	Ex-ante	Classification	Y			Y	Y
Shamma et al. [171]	ICWSM '11	Ex-ante	Classification				Y	
Lakkaraju & Ajmera [108]	CIKM '11	Ex-ante	Both				Y	Y
Yan et al. [215]	CIKM '11	Ex-ante	Regression			Y	Y	Y
Stieglitz et al. [175]	HICSS '12	Ex-ante	Regression				Y	Y
Tsur & Rappoport [190]	WSDM '12	Both	Regression	Y			Y	Y
Oghina et al. [150]	ECIR '12	Peeking	Regression	Y				Y
Artzi et al. [7]	HLT-NAACL '12	Ex-ante	Classification	Y			Y	Y
Bandari et al. [12]	ICWSM '12	Ex-ante	Both				Y	Y
Ruan et al. [167]	WebSci '12	Peeking	Regression	Y		Y	Y	Y
Ma et al. [133]	SIGIR '12	Peeking	Classification		Y	Y		Y
Gupta et al. [73]	ASIST '12	Peeking	Both	Y			Y	Y
Kupavskii et al. [105]	CIKM '12	Both	Both	Y	Y	Y	Y	Y
Figueiredo [52]	WSDM '13	Peeking	Classification	Y				Y
Ahmed et al. [3]	WSDM '13	Peeking	Regression	Y				
Pinto et al. [155]	WSDM '13	Peeking	Regression	Y				
Ma et al. [134]	JASIST '13	Peeking	Classification		Y	Y		Y
Bao et al. [14]	WWW '13	Peeking	Regression		Y	Y		
Jenders et al. [86]	WWW '13	Ex-ante	Classification				Y	Y
Kupavskii et al. [106]	ICWSM '13	Both	Regression	Y	Y	Y	Y	Y
Romero et al. [164]	ICWSM '13	Peeking	Classification		Y	Y		
Lakkaraju et al. [109]	ICWSM '13	Ex-ante	Regression					Y
Cui et al. [45]	KDD '13	Peeking	Classification			Y	Y	
Weng et al. [205]	Sci. Rep. '13	Peeking	Classification		Y	Y		
Li et al. [122]	CIKM '13	Peeking	Regression	Y	Y	Y		
Tatar et al. [182]	SNAM '14	Peeking	Regression	Y				
Castillo et al. [30]	CSCW '14	Peeking	Regression	Y			Y	Y
Abisheva et al. [1]	WSDM '14	Peeking	Both			Y	Y	
McParlane et al. [137]	ICMR '14	Ex-ante	Classification				Y	Y
Totti et al. [185]	WebSci '14	Ex-ante	Classification				Y	Y

Continued in next page.

Table 9. Continued: Feature-based models.

Reference	Venue	Strategy	Formulation	Feature				
				Te.	CS	GG	Us.	Co.
Gao et al. [58]	WWW '14	Peeking	Both	Y	Y	Y		
Cheng et al. [38]	WWW '14	Peeking	Classification	Y	Y	Y	Y	Y
Khosla et al. [91]	WWW '14	<i>Ex-ante</i>	Regression				Y	Y
Weng et al. [206]	ICWSM '14	Peeking	Classification	Y	Y	Y		
Bian et al. [19]	SIGIR '14	Peeking	Classification			Y	Y	Y
Kong et al. [101]	SIGIR '14	Peeking	Both	Y	Y		Y	Y
He et al. [77]	SIGIR '14	Peeking	Regression	Y		Y	Y	
Gao et al. [59]	APWeb '14	Peeking	Classification	Y	Y	Y		
Kong et al. [99]	JISIC '14	Peeking	Classification	Y	Y			
Yu et al. [228]	ACM MM '14	Peeking	Classification	Y		Y	Y	
Wang et al. [198]	AAAI '15	Peeking	Classification	Y	Y	Y	Y	
Liu et al. [125]	AAAI '15	Peeking	Classification	Y				
Dong et al. [47]	WSDM '15	<i>Ex-ante</i>	Classification			Y	Y	Y
Kong et al. [102]	JASIST '15	Peeking	Both	Y	Y		Y	Y
Yu et al. [229]	ICWSM '15	Peeking	Regression	Y				
Alzahrani et al. [4]	SBP '15	Peeking	Classification		Y	Y		
Guo et al. [71]	ASONAM '15	Peeking	Classification	Y	Y	Y		
Bora et al. [23]	SNAM '15	Peeking	Classification		Y	Y	Y	
Vallet et al. [192]	CIKM '15	Peeking	Classification	Y			Y	
Gelli et al. [64]	ACM MM '15	<i>Ex-ante</i>	Regression				Y	Y
Yi et al. [226]	Physica A '16	Peeking	Classification	Y	Y			Y
Dong et al. [48]	TBD '16	Both	Both	Y		Y	Y	Y
Wu et al. [208]	AAAI '16	<i>Ex-ante</i>	Regression	Y			Y	Y
Martin et al. [135]	WWW '16	<i>Ex-ante</i>	Regression				Y	Y
Rizos et al. [160]	WWW '16	Peeking	Regression	Y	Y	Y		
Shulman et al. [173]	ICWSM '16	Peeking	Classification	Y	Y	Y	Y	
Krishnan et al. [103]	WebSci '16	Peeking	Classification	Y	Y	Y	Y	
Kong et al. [98]	CCIS '16	Peeking	Classification	Y	Y	Y		
Guo et al. [72]	ASONAM '16	Peeking	Both	Y	Y	Y		
Zhang et al. [237]	ECML PKDD '16	Peeking	Classification	Y	Y	Y	Y	Y
Chen et al. [35]	ACM MM '16	<i>Ex-ante</i>	Regression				Y	Y
Mishra et al. [141]	CIKM '16	Peeking	Both	Y			Y	
Trzciński & Rokita [186]	TMM '17	Peeking	Regression	Y				Y
Hoang et al. [78]	WWW '17	Peeking	Regression	Y		Y	Y	
Lu & Syzdzanski [127]	IPDPSW '17	Peeking	Classification				Y	Y
Shafiq & Liu [170]	Networking '17	Peeking	Classification		Y			
Lv et al. [131]	ACM MM '17	<i>Ex-ante</i>	Regression				Y	Y
Xie et al. [213]	BigData '17	Peeking	Classification	Y	Y	Y		Y
Luo & Liu [130]	COLING '18	<i>Ex-ante</i>	Classification				Y	Y
Jia et al. [87]	CN '18	Both	Classification		Y	Y	Y	
Wu et al. [210]	ICWSM '18	<i>Ex-ante</i>	Regression				Y	Y
Zhao et al. [244]	IJCAI '18	<i>Ex-ante</i>	Classification					Y
Kong et al. [97]	SMC '18	Peeking	Classification	Y	Y	Y	Y	
Tsugawa [189]	ICWSM '19	Peeking	Classification			Y	Y	Y
Total: 89				51	32	39	53	49
				57%	36%	44%	60%	55%

Table 10. Temporal features.

Feature	Description
<i>change_rate</i>	The change rate of early popularity before observation time $t_o$ . [38, 52, 54, 103, 125, 160, 170, 192, 198, 213]
<i>dormant_period</i>	Dormant period refers to the time period before the item getting adopted. [62, 101, 102]
<i>local_peaks</i>	Local peak means in this time interval the increasing speed of popularity is larger than its neighbor intervals' speed. [198]
<i>maximum_interval</i>	Maximum time interval, i.e., $\max([t_j - t_{j-1}]_j)$ . [58, 62]
<i>mean_t, stddev_t</i>	Mean, median, sum and/or standard deviation of time series. [38, 62, 101–103, 111, 125, 160, 173, 213, 234, 237]
$P_i(t_j)$	Popularity of information item $I_i$ at time $t_j$ or incremental popularity at a time interval. [14, 23, 24, 30, 39, 48, 52, 54, 73, 77, 98, 99, 101, 106, 111, 128, 133, 134, 141, 150, 160, 178, 179, 184, 188–190, 192]
<i>peek_fraction</i>	The maximum of incremental popularity in all time intervals divided by the total popularity $P(t_o)$ at observation time. [52, 54]
<i>publication_time</i>	Publication time of an information item, e.g., year, month, week, day, hour, minute, etc. [7, 24, 52, 109, 131, 135, 154, 165, 185, 187, 192, 198, 216]
<i>stage</i>	Whether an item published at an early stage or late stage during observation time. [219]
$t_1$	$t_1$ measures speed how soon the first participant $u_1$ will join in the cascade. [13, 56, 111]
$t_k$	$t_k$ measures the how soon a cascade reaches size $k$ . [23, 38, 71, 103, 160, 173, 206, 213, 237]
<i>time_series_vector</i>	$k$ -dimensional time series vector in fixed $k$ number observation setting, or vector in fixed time observation setting. [38, 58, 59, 62, 74, 77, 81, 186]
$[(t_j - t_{j-1})]_j$	Time series between the $j$ -th participant $u_j$ and $(j - 1)$ -th participant $u_{j-1}$ , $j \in [1, k]$ , which is a $k$ -dimensional time vector. This can also be extended to $[(P(t_j) - P(t_{j-1}))]$ , where $[t_{j-1}, t_j]$ is a predefined time interval. [13, 38, 52, 59, 62, 81, 97, 99, 141, 155, 198, 206, 226, 228, 234]
<i>time_series_cluster</i>	Which trend cluster/shape a sequence of early time series belongs to. [52, 54, 62, 74, 97, 101, 102, 221, 229]



Table 11. Cascade graph,  $r$ -reachable graph, and global graph features.

Feature	Description
<i>authority/hub_score</i>	Authority or hub scores of nodes in graph $\mathcal{G}$ . [58, 59, 62, 133, 134, 157, 215, 226, 228]
<i>branching_factor</i>	Branching factor is the number of children nodes at each node in tree cascade graphs $\mathcal{G}_c$ . [49]
<i>centrality</i>	Centrality measured by a specific algorithm, e.g., centrality of eigenvector, closeness, and betweenness. [4, 72, 157, 226, 237]
<i>clustering_coefficient</i>	Clustering coefficient of graph $\mathcal{G}$ . [39, 56, 58, 59, 62, 81, 87, 157, 170]
<i>community</i>	Communities detected by using specific algorithms. Example features including number of infected communities, intra-communities, Gini impurity, etc. [29, 71, 119, 170, 189, 205, 206, 213]
<i>connect_component</i>	Connect component in the graph $\mathcal{G}$ , measured by number, component size, maximum component size, etc. [23, 58, 59, 62, 133, 134, 157, 164, 173, 191]
<i>degree</i>	Degrees (both in- and out-degrees in directed graph) of nodes in graph $\mathcal{G}$ . [1, 4, 11, 23, 29, 38, 45, 47, 49, 56, 62, 81, 85, 87, 97–99, 101–103, 119, 160, 170, 173, 189, 194, 198, 213, 222, 226, 228, 234, 237]
<i>density</i>	Density of graph, defined as the number of edges divided by all possible edges in $\mathcal{G}$ , i.e., $ \mathcal{E} /( \mathcal{V}  \times ( \mathcal{V}  - 1))$ . [23, 58, 59, 98, 101–103, 119, 133, 134, 157, 164, 173, 198, 222, 226]
<i>depth</i>	Depth of the path from node $u_0$ to node $u_j$ in graph $\mathcal{G}$ . [13, 14, 38, 39, 56, 58, 59, 72, 85, 97–99, 101, 103, 160, 170, 173, 206, 226, 234, 237]
<i>direct_connect</i>	Number of nodes in cascade graph $\mathcal{G}_c$ that are directly connected to the root node $u_0$ , i.e., $ \{u_j   \text{dist}(u_0, u_j) = 1, 1 \leq j \leq N\} $ . [38, 103, 173]
<i>edge_density</i>	Edge density is defined by the ratio of number of edges to the number of all possible edges. [14, 29, 62, 97, 99, 119]
<i>edges_</i> $\mathcal{G}_c$	Number of (weighted/unweighted) edges in cascade graph $\mathcal{G}_c$ . [23, 29, 38, 103, 164]
<i>edges_</i> $\mathcal{G}_c^1$	Number of (weighted/unweighted) edges in 1-reachable graph $\mathcal{G}_c^1$ . [29, 38, 62, 103, 119, 164, 173]
<i>indirect_connect</i>	Number of nodes in cascade graph $\mathcal{G}_c$ that are indirectly connected to the root node $u_0$ , i.e., $ \{u_j   \text{dist}(u_0, u_j) > 1, 1 \leq j \leq N\} $ . [38, 103]
<i>leaf_nodes_</i> $\mathcal{G}_c$	Number of leaf nodes in cascade graph $\mathcal{G}_c$ . [28, 29, 103, 119]
<i>nodes_</i> $\mathcal{G}_c$	Number of nodes in graph $\mathcal{G}_c$ , i.e., $ \mathcal{V}_c $ . [1, 13, 29, 59, 87, 101–103, 133, 134, 164, 170, 205, 206, 226]
<i>nodes_</i> $\mathcal{G}_c^1$	Number of nodes in 1-reachable graph $\mathcal{G}_c^1$ , i.e., $ \mathcal{V}_c^1 $ . [1, 29, 38, 58, 59, 71, 72, 119, 133, 134, 164, 205, 206, 213]
<i>nodes_</i> $\mathcal{G}_c^2$	Number of nodes in 2-reachable graph $\mathcal{G}_c^2$ , i.e., $ \mathcal{V}_c^2 $ . [62, 72, 103, 173, 206]
<i>pagerank</i>	PageRank. [4, 47, 48, 77, 81, 87, 105, 106, 157, 162, 198, 206, 222, 226, 228]
<i>reciprocity</i>	Reciprocity of directed graph $\mathcal{G}$ (cf. [63]), defined as the ratio of number of reciprocal links to the total number of links, i.e., $(\sum_{i,j}^{ \mathcal{E} } \mathbb{I}(A_{i,j} = A_{j,i}))/ \mathcal{E} $ . [58, 59, 62, 81]
<i>similarity_</i> $\mathcal{G}_c$	Similarity of two cascade graphs measured by a specific algorithm, e.g., graph edit distance or vertex/edge overlap. [198]
<i>structural_virality</i>	Structural virality is calculated by Wiener index (cf. [67]). [97, 160, 198]
$ \mathcal{V}_c \cap \mathcal{N}_g(u_j) $	Number of nodes who in cascade graph $\mathcal{G}_c$ are also neighboring nodes to $u_j$ in global graph $\mathcal{G}_g$ , e.g., how many followers of $u_j$ retweet $u_j$ 's tweet in Twitter, sometimes refer to the retweet ratio $ \mathcal{V}_c \cap \mathcal{N}_g(u_j) / \mathcal{N}_g(u_j) $ . [11, 38, 103, 106, 154, 167, 179, 237].
<i>subcascade_graph</i>	In tree cascade graph $\mathcal{G}_c$ , for each node in $\mathcal{V}_c$ , its subcascade graph is composed of all the descendants of this node. [29, 49, 170]
<i>triangles</i>	Number of triangles in graph $\mathcal{G}$ . [29, 103, 119, 133, 134, 163, 194]

Table 12. User/Item features.

Feature	Description
<i>activity<sub>u<sub>j</sub></sub></i>	Activity measures the frequency of user $u_j$ 's activities. [38, 69, 101, 103, 128, 165, 173, 210, 222, 228, 230]
<i>age<sub>u<sub>j</sub></sub>_account</i>	Age of user $u_j$ 's account from item publication date to the account creation date. [11, 24, 38, 62, 87, 91, 103, 105, 106, 128, 135, 141, 165, 173, 175, 178]
<i>age<sub>u<sub>j</sub></sub></i>	Age of user $u_j$ . [38, 49, 232]
<i>age<sub>I<sub>i</sub></sub></i>	Age of item $I_i$ . [24, 39, 48, 222]
<i>attractiveness<sub>I<sub>i</sub></sub></i>	Attractiveness of item $I_i$ is defined as the ratio of users who adopt this item after seeing this item. [122]
<i>favorites<sub>u<sub>j</sub></sub></i>	Number of favorite items user $u_j$ likes. [105, 106, 128, 154, 178, 192]
<i>followees<sub>u<sub>j</sub></sub></i>	Number of users who follow user $u_j$ , this is a special case of in-degree of user $u_j$ in global graph $\mathcal{G}_g$ (follower/followee graph). [7, 11, 30, 35, 38, 62, 84, 97, 128, 165, 178, 185, 190, 192, 198, 214, 230, 235, 242]
<i>followers<sub>u<sub>j</sub></sub></i>	Number of followers to user $u_j$ , this is a special case of out-degree of user $u_j$ in global graph $\mathcal{G}_g$ (follower/followee graph). [1, 7, 11, 20, 24, 30, 35, 38, 73, 77, 84, 86, 96, 97, 101–103, 105, 106, 128, 130, 135, 141, 154, 158, 162, 165, 167, 173, 175, 178, 190, 192, 198, 206, 214, 230, 234, 235, 237, 242]
<i>friends<sub>u<sub>j</sub></sub></i>	Number of friends of user $u_j$ , where <i>friend</i> is a reciprocal relationship between two users. [7, 24, 30, 38, 62, 73, 84, 91, 105, 106, 130, 135, 137, 141, 154, 178, 190, 192, 208]
<i>gender<sub>u<sub>j</sub></sub></i>	Gender of user $u_j$ . [1, 38, 49, 87, 103, 137, 185, 198, 232, 237]
<i>h-index</i>	Hirsch index of users' past publications. [47, 160, 162, 215]
<i>historical_items<sub>u<sub>j</sub></sub></i>	Number of historical items a user $u_j$ published. [11, 24, 35, 47, 81, 87, 90, 91, 101, 102, 106, 128, 135, 137, 141, 154, 165, 178, 185, 192, 198, 208, 215, 219, 222, 223, 228, 237]
<i>influence<sub>u<sub>j</sub></sub></i>	User $u_j$ 's influence measured by a specific criteria. [103, 162, 167, 201, 219, 222]
<i>interests<sub>u<sub>j</sub></sub></i>	User $u_j$ 's interests/preferences in topics or items. [1, 19, 222, 223]
<i>location<sub>u<sub>j</sub></sub></i>	The location (country/region) of user $u_j$ . [1, 23, 190, 198, 232]
<i>name<sub>u<sub>j</sub></sub></i>	Name of the user $u_j$ . [235]
<i>passivity<sub>u<sub>j</sub></sub></i>	Passivity/susceptibility measures how difficulty a user $u_j$ to be influenced. [102, 127, 164, 201]
<i>past_success</i>	Past success or historical popularity of users/items. [1, 7, 11, 24, 35, 47, 81, 87, 90, 91, 101, 102, 105, 106, 108, 130, 131, 135, 141, 162, 189, 208, 210, 215, 228]
<i>profile_views<sub>u<sub>j</sub></sub></i>	The number of profile views of user $u_j$ . [90]
<i>rating<sub>I<sub>i</sub></sub></i>	The rating of information item $I_i$ . [24, 39]
<i>referrer<sub>I<sub>i</sub></sub></i>	Referrer means a external source or link that pointed to the information item $I_i$ , e.g., a link refer to a video. [1, 24, 30, 52, 54]
<i>relevance<sub>u<sub>j</sub></sub>_I<sub>i</sub></i>	How relevance between user $u_j$ and item $I_i$ . [47, 222]
<i>similarity<sub>I<sub>i</sub></sub></i>	How many similar items or even clones/duplicates with regard to item $I_i$ , or how uniqueness $I_i$ is. [24, 38, 42, 109, 127, 200]
<i>type<sub>I<sub>i</sub></sub></i>	Which type the information item $I_i$ belongs to. [87, 154, 185, 219]
<i>type<sub>u<sub>j</sub></sub></i>	Which type the user $u_j$ belongs to, e.g., verified or premium account, personal or organization account. [35, 38, 91, 105, 106, 128, 130, 137, 154, 198, 208, 237]