Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 173 (2020) 201–209

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on *Smart Sustainable Intelligent Computing and Applications* under ICITETM2020

Predicting Information Cascade on Twitter Using Random Walk

Nidhi Singh[a], Anurag Singh[a], Rajesh Sharma[b]

[a]*Department of Computer Science & Engineering, National Institute of Technology Delhi, New Delhi-110040, India*
[b]*Institute of Computer Science, University of Tartu, Estonia*

Abstract

Online Social networking (OSN) platforms are the most widely used platforms for spreading information about current topics. One of the major advantages of these websites is the accessibility and reach to the users of these platforms. These platforms do not demarcate information cascading based on geography, race, caste, and creed to name a few. The prediction of a cascade is an important problem as it gives a deep understanding of opinion-shaping in audience. For eg. diffusion about a product, politics for, example, celebrity popularity. Many models have been proposed for solving this problem that uses content and user-based information, temporal and structural features of the network. Propagation is faster if the similarity between entities is high. In this paper, we propose a random walk based method that also exploits similarity measures (user and content) to predict the diffusion cascade. We evaluated our method using Twitter as a use case and demonstrated that our method outperforms the existing structural and content-based methods being proposed earlier.

Keywords: Information cascade; Social networks; random walk; random forest; link prediction

1. Introduction

The evolution of social media platforms like Twitter, Facebook, Whatsapp has changed the process of information cascade. Due to easy accessibility, especially because of smartphones, a large number of people have joined these social networking platforms. Also besides, these platforms have become a major source of information propagation or cascade. The significant news about disasters, massacre, epidemics, riots, political issues are often propagated through these platforms and thus, within a short time particular information reaches a large number of people.

* Corresponding author. Tel.: +91-727-661-5616 ;
E-mail address: nidhisingh.1809@gmail.com

1877-0509
10.1016/j.procs.2020.06.024

Information cascade can be defined as when one individual observes other individual's posts, tweets, etc. and starts sharing or reposting it. This process keeps on happening until individuals stop spreading or reposting it. In different social platforms the activity can correspond to sharing of views, posts, multimedia (like a photo, video) etc.

The propagation of news could be useful in shaping the opinions of users [1]. Analyzing the information diffusion could be useful for businesses in the launching of products [2]. It can also be used for understanding the stock markets. The government and political parties use this to spread ideology and gain support during election or launching of policies [3].

The concept of information cascade was first proposed by Bikhchandani et al.[4] in 1992 where they observed the behavior of people with respect to the fashion industry. Their major finding was based on an individuals observance of other people like celebrities, film stars and then following those patterns. Later, it diverged to other domains also where people propagated their culture, habits [5] etc. The researchers have analyzed cascading processes in social network platforms such as blogs [6], Facebook [7], Twitter [8], chain of emails [9], as well as in online marketing [10].

The process of the cascade has also been studied using epidemic models [11]. The cascade problem has also been defined as a regression problem [12], where the number of times a message is forwarded could depend on period time, the content of the message with respect to its temporal features or popularity of tweets based on the tweet as well as user features [13, 14]. The major emphasis in the earlier works has been on the features of the social network like a network of the users, the content of the information and temporal properties [15].

In the recent past, in addition to content-based features, researchers have started exploring User-Based, Structural Properties and Temporal features. Tsur et al. used the content-based features in the prediction of cascade [14]. Gazbo et al. used topic modeling for their prediction [15]. Cheng et al. used the user based feature of Facebook users for prediction [13]. The user's social status has also an impact on the propagation as celebrities, politicians, leaders have a mass following and it can impact the propagation [15]. Weng et al.proposed to consider network-based features (density of graph, proximity to the root node) in propagation [16]. Kleinberg et al. discussed temporal features [17]. Some other important temporal features like the number of views, shares per second are also very relevant features [16].

Diffusion can be studied using tree-like structures where initiator becomes the root nodes and those propagating it further becomes the child nodes [18]. Such trees are called diffusion trees. We will be using this model on twitter networks using the follower-followee relationship.

In this work, we propose a random walk based model which also exploits the features based on the content of the text being propagated and, users' similarity for predicting the cascade among users. The proposed algorithm gives higher accuracy than the existing approaches. The detailed methodology (See Section 3) and results (See Section 4) are presented in later sections.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 discusses our proposed approach. In Section 4, we present the results. Lastly, Section 5 concludes our work, along with our future directions for this work.

2. Related work

The existing literature can be categorized into the following four classes, based on the type of factors they have exploited for the prediction.

1. **Content Based Features:** Features based on the content of the text have often been used for predicting the propagation. The basic idea behind this process is that if the content consists of trending topics, it will eventually propagate. To identify the major topics in posts, algorithms for topic modelling such as LDA and text ranking have been exploited. This is particularly useful in case of tweets analysis, as one can easily find the major topics using hashtags. Along with topic modelling, sentiment analysis has also been exploited [15].

2. **User Based Features:** The idea behind these sets of features is that the users with a large number of connections will generate a larger cascade [13]. The connections can exist in the form of friends, followers, friends of friends. The user's social status has also an impact on the propagation as celebrities, politicians, leaders have a mass

following and it can impact the propagation [15]. The social activity of the user can also contribute. In general, the more active users have a higher impact than the less active users of social platforms.

3. **Structural Features of the network:** The network, which provides the channel for propagation is an important factor in propagation. Factors like densities of the graph and the proximity to root nodes have been considered works like [13, 16]. Here proximity is defined in terms of closeness to the root node. If its next neighbour, it will be affected much earlier than others.

4. **Temporal features:** Time is an important factor in making a topic popular and subsequently creating interest among the users in propagating the message. New topics are shared more rapidly compared to the topics which are relevantly older. With time, the diffusion slower downs [17, 13]. Some other important temporal features like the number of views, shares per second are also very relevant features [16].

Using these entire features, the diffusion trees are constructed from graphs. A diffusion tree is constructed on the basis of re-tweeting. In past, the combination of above-mentioned techniques has been widely used to predict the size of cascade on various social networking platforms. For comparing our work with existing work, cascade prediction on Facebook [13] is used.

Information cascade can also be framed as epidemic models since spreading is like infection [11] proposed a framework for disease spreading in social networks. Vespigani et al. [19] also proposed about the dynamics of the social network and epidemic spreading. It also discussed various factors which can be intrinsic or extrinsic. Intrinsic can affect the users initially at faster rates. Extrinsic are those which are responsible due to community or network where the node resides. The cascade problem can also be defined as a regression problem or binary classification problem. Bakshy et al. [12] suggested the diffusion problem as a regression model for time-evolving social networks. In regression studies the aim is to find the number of times a message will be forwarded and in classification, the goal is simply to find if a message will be forwarded or not [20].

3. Proposed Methodology

In this section, we first discuss the baseline method before presenting our proposed approach which is based on a random walk for predicting information cascade. We discuss our approach in particular with respect to Twitter settings, using which we have evaluated our model. We have used diffusion trees that are constructed from graphs. A diffusion tree is constructed on the basis of re-tweeting. Once a user has tweeted, the users that re-tweet after that are child nodes. Then, retweet on the child node's retweets. This continues until no more tweets are possible [13] (BL).

3.1. Baseline method

In Twitter settings, once a user has tweeted, the users that retweet after that are child nodes. This process continues until no more retweet happens. For comparing our work with existing work, cascade prediction in Facebook is used [13] as a baseline method. In this work, all the four features have been considered and then the prediction of cascade size is done using various classifiers.

It is important to note that the method being employed [13] is platform dependent as it is not possible to extract all the four features in present Twitter settings as the privacy policies of the Twitter prevents the access to all the four features extraction. Thus, in the absence of temporal and structural features, it is not easy to replicate the experiment. Therefore, in our baseline approach, we only consider the content and user-based features.

3.2. Proposed method

Our proposed model is a combination of 3 features: user similarity, content similarity and random walks. The random walk makes the model more robust by taking into account the arbitrary connection of nodes. Next, we provide a formulation of every feature is given.

Users' Similarity: In Twitter settings, for finding the user similarity, if two users have follower-followee relationship or have replied/retweeted to some original tweet then they are similar, otherwise not.

$$S_{i,j} = \begin{cases} 1, & \text{if user } u_i \text{ similar to the user } u_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Content Similarity: The similarity of tweets' is calculated using the cosine similarity. Based on two tweets' word collection, a vector is created. Later, binary vectors for each Tweet is found which is used in calculating Cosine Similarity. Every tweet can be considered as an individual document. It is given as:

$$CS_{\hat{t_1},\hat{t_2}} = \frac{\langle \hat{t_1} \cdot \hat{t_2} \rangle}{||\hat{t_1}||||\hat{t_2}||} \tag{2}$$

Here $\hat{t_1}$, $\hat{t_2}$ represents the tweet vectors.

Random Walks: We have also exploited random walks for predicting the cascade processes. The finite random walk consists of fixed numbers of steps. Consider a random walk in which starts from node n_i and eventually stops at node n_j. If p_i represents the probability of leaving node i and $p_{i,j}$ is the probability for reaching node j after leaving i. Then a random walk starting from the node i and of length, n can be represented as:

$$p_i = \{p_{i1}, p_{i2}, ...p_{in}\} \tag{3}$$

The probability to be in self-state is $p_{ii} = 1 - \alpha$ where α represents the prior that probability walker will certainly leave its current state. The prior probability is proportional to the out-degree of the node.

The matrix of similarity and probability is calculated as **S**, **CS** and **P** using Equation 1, 2, 3. The corresponding degree matrix is given as **D**. Once a steady state is reached, the probability that random walker will remain at node j is proportional to the similarity between two nodes and is given as:

$$\mathbf{P} = \mathbf{DS} \tag{4}$$

The final steady state of random walk can be calculated as:

$$\mathbf{R}(t+1) = \alpha\mathbf{PR}(t) + (1-\alpha)\mathbf{I} \tag{5}$$

$R(t)$, $R(t+1)$ are defined as state probability matrix at t and t+1 respectively.

Predicting cascade size: To predict the cascade size we used three metrics the user similarity **S**, the content similarity of tweets **CS** and, the probability of random walk $\mathbf{R_t}$. Next, we explain the pseudo-code of our approach. The first step is the network generation (line 1), in which all the users of Twitter in our datasets are considered. The edges between two users are created if they have a follower - followee relationship or have replied to the messages. Next, we find the users' similarity matrix as specified in Equation 1 (line 2). Subsequently, the next step is to find the similarity of tweets as given in Equation 2 (line 3). Then, find the transition probability matrix P (line 4). Lastly, steady-state probabilities are calculated using Equation 5.

Algorithm 1 Algorithm for Cascade Size

Input Edgelist file
Output State probability matrix

Generate a network represented as $G(V, E)$
Compute user similarity matrix , $S_{i,j}$ for every pair of vertex i, j using eqn 1.
Compute tweet similarity matrix , $CS_{i,j}$ for every pair of vertex i, j using eqn 2.
Calculate transition probability matrix P using eqn 4
for every node:
    compute steady state probability using eqn 5

Table 1: Results of classification using Baseline and Proposed approach

Classifier	Dataset	BL_Ac	RW_Ac	BL_Pr	RW_Pr	BL_Re	RW_Re	BL_F1	RW_F1	BL_AUC	RW_AUC
NB	GST	.7300	.8235	.8100	.8482	.7951	.9065	.8065	.8746	.6771	.8034
Random Forest	GST	.9000	**.9487**	.9412	**.9524**	.9302	**.9816**	.9412	**.9668**	.8744	**.9421**
SVM	GST	.8972	.9249	.9029	.9000	.9291	.9803	.9342	.9484	.8549	.9345
LR	GST	.8549	.9023	.8782	.9020	.9657	.9414	.8958	.9202	.8615	.8985
NB	ATROCITIES	.6347	.7532	.7273	.8868	.7380	.8331	.7326	.8350	.5796	.7261
Random Forest	ATROCITIES	.7087	**.7847**	.8125	**.7919**	.8125	**.9077**	.8125	**.8459**	.6682	**.8989**
SVM	ATROCITIES	.7223	.8731	.7873	.9000	.8733	.9073	.7767	.9087	.6179	.8759
LR	ATROCITIES	.7044	.8548	.7500	.8382	.8333	.9657	.7914	.8787	.6587	.8985
NB	OSCARS	.5384	.6978	.6957	.7692	.6724	.8448	.6835	.8081	.4145	.5779
Random Forest	OSCARS	.6218	**.695**	.7857	**.6622**	.6471	**.8991**	.7097	**.7626**	.6223	**.7392**
SVM	OSCARS	.6210	.7342	.7000	.7600	.6951	.8880	.6976	.8090	.5954	.6773
LR	OSCARS	.5856	.7429.	.6346	.7611	.7640	.8055	.7050	.8205	.5046	.6996

4. Results

In past, the combination of above-mentioned techniques has been widely used to predict the size of cascade on various social networking platforms. For comparing our work with existing work, cascade prediction on Facebook [13] is used. In that work, all the four features have been considered and then the prediction of cascade size was done using various classifiers. Hence this method will be considered as the baseline(BL) of our work.

4.1. Datasets

For evaluating the proposed model, we collected Twitter data (tweets) using Twitter's public API. Every tweet in dataset consists of tweet text, user id, reply id, re-tweet id information. The tweets are collected on three different topics. The first topic is about economic policy GST (Goods and Services Tax) implemented by the Indian government in July 2017; The second set of tweets are related to the Oscar awards in January 2018; The last set of tweets are about the recent riots in India in March 2018. The tweets with respect to the first and the last topic are mostly from India compared to the second topic, which has their tweeters worldwide. The dataset has been split using temporal split into training (80%) and testing (20%) sets. The class labels comprise of 1 if there is a connection and 0 if not. The training dataset consists of 108472 nodes and 562181 edges. The three testing sets comprise of: GST with 27118 nodes and 80251 edges; Atrocities with 30472 nodes and 82219 edges; Oscars with 32214 nodes and 92487 edges. We used four classifiers namely i) Nave Bayes (NB), ii) Support Vector Machines (SVM) iii) Logistic regression (LR) and, iv) Random Forest(RF). In Random forest, we have considered a number of trees as 5.

4.2. Metrics

The results are presented in Table 1. We have compared various classifiers based on accuracy, precision, recall, f1 and AUC score. Let, e_{tp} is the number of edges which are predicted by the classifier and also exist in the original graph, e_{tn} are the number of edges which are predicted by the classifier but does not exist in the original graph, e_{fn} are the number of edges which are not predicted by the classifier but exist in the original graph and e_{fp} are number of edges which are neither predicted by classifier nor in the original graph. Accuracy is the number of predictions that classifier predicted right to the total samples. Edges that were actually in the dataset as well as predicted by the classifier to the total number of edges in the dataset.

$$Ac = \frac{e_{tp} + e_{fp}}{e_{tp} + e_{tn} + e_{fp} + e_{fn}} \tag{6}$$

Precision is to measure the quality of our predictions only based on what our predictor claims to be positive. The edges that were predicted to the total edges that should exist in our scenario.

$$Pr = \frac{e_{tp}}{e_{tp} + e_{fp}} \tag{7}$$

The recall is to measure such quality with respect to the mistakes we did (what should have been predicted as positive but we flagged as negative). It can be defined as edges that our classifier missed to predict.

$$Re = \frac{e_{tp}}{e_{tp} + e_{fn}} \tag{8}$$

f1-score is defined as the harmonic mean of precision and recall.

$$f1 - score = \frac{2 * Pr * Re}{Pr + Re} \tag{9}$$

The area under the curve(AUC) is used to determine which model is best among the used one.

4.3. Results

It may be easily inferred from Table 2, the proposed method RW performed better for all the three datasets than the existing methods. The accuracy, precision, recall, f1-score and AUC of random walk (RW) is better for every dataset compared to the baseline method (BL). This can be accounted for using the random walk that was used in the proposed method. In earlier methods, if any of the structural, content, user and temporal properties do not show similarity for two users, then it is predicted that cascade cannot propagate. But using a random walk, we found out that probability for each pair of nodes to be connected. Using this RW probability as one of the features for the classifier, the more accurate prediction of the cascade is done. In our case, we just use 3 metrics: Content Similarity matrix CS , User Similarity Matrix S, and the random walk probability matrix R can be easily retrieved.

Our testing datasets are diverse. The theme of the three datasets is different from each other. So, the contents have no similarity between them. Although some users are common in the datasets, user similarity could be found for those cases only. Therefore, existing BL method performed poorly on the cross-domain datasets of Oscars and Atrocities. The proposed RW returns a good prediction value.

The error rate of the random forest classifier on both the BL and RW model is shown in Figure 1. The graph is plotted for different testing sets against the error rate. Only these two are chosen for plotting the graph as the accuracy of both these is high in their respective categories. Error Rate is defined as the number of misclassifications done by the classifier on the testing set. It is visible that the error rate of RF_BL is higher than RF_RW. The error rate for the RF_BL lies above 10 inferring high rates of misclassifications as shown in figure 1. Our model reduced the error rate by using the random approach. In earlier models, the probability of link formation depended on features of the graph only. If two nodes don't have common features, then their probability is zero for forming a link. But using the random

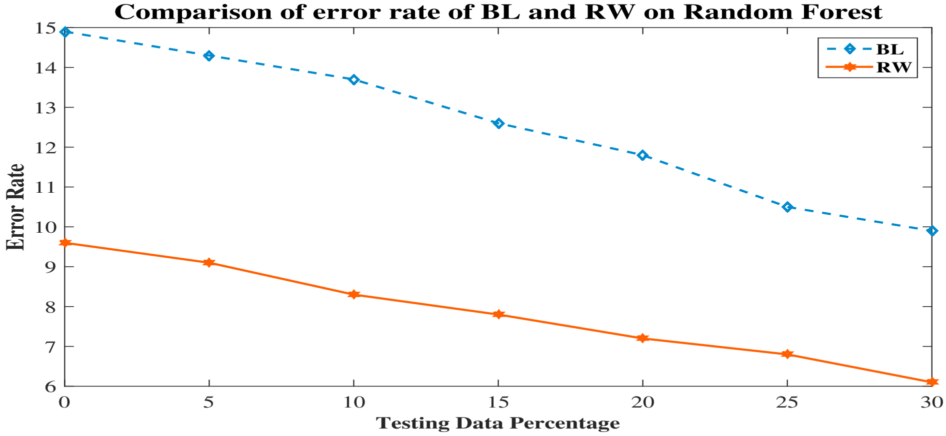

Fig. 1: Error rates of random forest on BL and RW

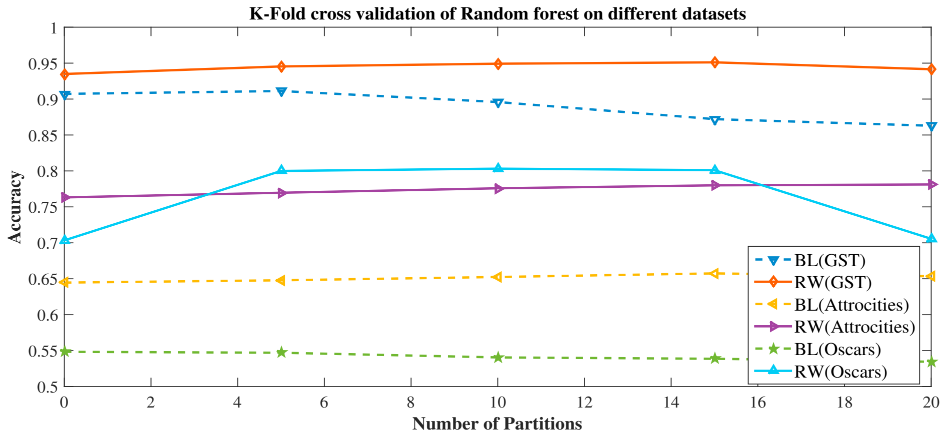

Fig. 2: Accuracy for RW & BL in K-fold validation

walk approach, the probability of link formation between every pair of nodes is found. The probability is non zero and hence the proposed model reduced the error rate.

4.4. k-fold Cross Validation

To validate this model, k-fold cross validation was done. In k-fold cross validation, the dataset is divided into k parts. Then the model is trained on $k-1$ parts and tested on k^{th} part. In this section, we present the results of comparison of different datasets on different methods.

In Figure 2, shows the variation of accuracy for all three data-set on both models. The folds depict the partition of the data. It can be inferred from the above figure, the accuracy increased until an optimal number of partitions for all the dataset. Then it started decreasing. Overall the accuracy of RW remained better than BL. The optimal number of partitions for which accuracy gave a maximum was around 15 folds. On the other hand, the reason for the decrease in accuracy can be accounted using overfitting. As number of partition increasse, the data points for training also increases. This results in over-training.Therefore, it classifies the training data perfectly but fails on testing data. Hence, resulting in lower accuracy.

Table 2: Results of K-fold Cross Validation

Partitions	GST		Atrocities		Oscars	
	BL_RF	RW_RF	BL_RF	RW_RF	BL_RF	RW_RF
1	0.9072	**0.9349**	0.6447	0.7632	0.5484	0.7034
5	0.9112	**0.9454**	0.6479	0.7697	0.5471	0.8000
10	0.8959	**0.9492**	0.6525	0.7759	0.5406	0.8032
15	0.8721	**0.9511**	0.6575	0.7800	0.5387	0.8011
20	0.8629	**0.9415**	0.6535	0.7812	0.5346	0.7056

5. Conclusion and Future work

In this paper, we discussed a novel method based on the random walk and similarity measures to predict the cascade on twitter. Using four different metrics and machine learning models and three real datasets, we have shown that the model performs better than the existing model. In the present work, we have used the basic random walk model. In future work, we would like to perform our tests using bigger datasets. Also, we would like to incorporate other ensemble methods such as boosting algorithms to improvise the cascade prediction. To enrich the model, we also plan to incorporate other similarity measures.

6. Acknowledgments

This work has been supported in part by the SoBigData project (under grant agreement no. 654024)

References

[1] Jeremiah W. Jenks. The guidance of public opinion. *American Journal of Sociology*, 1(2):158–169, 1895.
[2] W Glynn Mangold and David J Faulds. Social media: The new hybrid element of the promotion mix. *Business horizons*, 52(4):357–365, 2009.
[3] Andrea Pereira Van Bavel, Jay J. The partisan brain: An identity based model of political belief. *Trends in Cognitive Sciences*, 22:213, 18.
[4] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.
[5] Sheen S Levine and Robert Kurzban. Explaining clustering in social networks: Towards an evolutionary theory of cascading benefits. *Managerial and Decision Economics*, 27(2-3):173–187, 2006.
[6] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25:211, 03.
[7] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. *ICWSM*, 1(2):12, 2013.
[8] Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the Association for Information Science and Technology*, 64(7):1399–1410, 2013.
[9] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc of the 12th ACM SIGKDD international conference*, pages 44–54. ACM, 2006.
[10] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, page 5, 07.
[11] Alessandro Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.
[12] Eytan Bakshy, Brian Karrer, and Lada A Adamic. Social influence and the diffusion of user-created content. In *Proc. of the 10th ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.
[13] Lada Cheng, Justin Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, page 925. ACM, 2014.
[14] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM, 2012.
[15] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.
[16] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3:25, 13.
[17] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proc of the sixth international conference on Web search and data mining*, page 13. ACM, 2013.
[18] Jure Leskovec. Diffusion and cascading behavior in networks. In *NATO ASI Workshop on Mining Massive Data Sets for Security*, pages 169–185, 2007.
[19] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 26(4):521–529, 2002.
[20] Kateryna Lytvyniuk, Rajesh Sharma, and Anna Jurek-Loughrey. Predicting information diffusion in online social platforms: A twitter case study. In *International Conference on Complex Networks and their Applications*, pages 405–417. Springer, 2018.

[21] Manish Wankhede, Vijay Trivedi, and Vineet Richhariya. Location based analysis of twitter data using apache hive. *International Journal of Computer Applications*, 153(10), 2016.

[22] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440, 1998.

[23] Lorenzo Cazzoli, Rajesh Sharma, Michele Treccani, and Fabrizio Lillo. A large scale study to understand the relation between twitter and financial market. *IEEE Third European Network Intelligence Conference (ENIC)*, pages 98–105, 2016.

[24] XU Ming-ming. Media agenda and public opinions shaping [j]. *Journal of Ningbo University (Liberal Arts Edition)*, 3:022, 2005.

[25] Gregorios Katsios, Ning Sa, and Tomek Strzalkowski. Social convos: A new approach to modeling information diffusion in social media. In *International Conference on Applied Human Factors and Ergonomics*, pages 25–36. Springer, 2019.

[26] Gerardo Iñiguez, Zhongyuan Ruan, Kimmo Kaski, János Kertész, and Márton Karsai. Service adoption spreading in online social networks. In *Complex Spreading Phenomena in Social Systems*, pages 151–175. Springer, 2018.

[27] Cigdem Aslay, Laks VS Lakshmanan, Wei Lu, and Xiaokui Xiao. Influence maximization in online social networks. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 775–776. ACM, 2018.

[28] Eitan Muller and Renana Peres. The effect of social networks structure on innovation performance: A review and directions for research. *International Journal of Research in Marketing*, 36(1):3–19, 2019.

[29] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[30] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.