A Review of Critical Research Areas under Information Diffusion in Social Networks

Surbhi Kakar[1], Monica Mehrotra[2]
Department of Computer Science
Jamia Millia Islamia University, New Delhi, India

***Abstract*—An online social network is a network where people exchange their ideas or opinions. Exchange of ideas between users leads to spread of information at a larger scale in the social networks. This spread of information is also called information diffusion. This work is dedicated to identifying research areas under the umbrella of Information Diffusion. The objective of this work is to present an extensive review of such areas, identify the existing research gaps and explore future directions of work. The review also identifies the methodologies, features and aspects studied in the current literature and proposes the optimal feature set to improve performance. This review will enable researchers to quickly identify the research areas, the current gaps and steer them into the possible future directions associated with them.**

***Keywords*—Information diffusion; influence maximization; retweet prediction; influence models**

I. Introduction

A social network is a network comprising users and relationships between them. The users can be modeled as nodes whereas the relationships between them can be viewed as edges between the nodes. Users in a social network posts/tweet messages to be viewed by other users. These messages are an indicator of how they feel, their ideas and their opinions about specific topics. The posted messages might be re-posted/retweeted by another user in that network. Re-posting another user's idea indicates that the user has been influenced as he/she is adopting or agreeing to the same idea or opinion. Information Diffusion is a broad area and it has attracted several researchers from diverse streams to work in this domain. Existing reviews [4] [88] [89] have studied this research area majorly along two aspects, namely, Influence modeling and Influence maximization. Moreover, the dimension of learning influence probabilities [23] and influence maximization in dynamic networks [90] have not been reviewed by these studies. Therefore, our work focuses on exploiting these gaps and presenting a broader review under the umbrella of Information Diffusion. Following are some of the areas that our work distinguished in order to accomplish the given objective:

1) Modeling influence in a social network: This area will review several works published on modeling influence probabilities between users. Influence Probability is a measure of the probability with which a user influences another user in a network.

2) Maximizing influence in a dynamic social network: Influence maximization has been studied over *several* years now, by different researchers. However, less work has been proposed with respect to dynamic networks. As social networks are evolving continuously, their network state is constantly being updated. Our work, hence, apart from discussing state of art techniques in static graphs, also lays emphasis on current works being done under dynamic networks.

3) Retweet prediction: Our work also attempts to review information diffusion along a new aspect called retweet prediction. This area mainly studies the factors affecting the influence and virality of a tweet message t and predicting whether a specific tweet will be retweeted or not. Fig. 1 shows the areas under Information Diffusion pictorially.

Keeping in mind the above stated research areas, following research questions have been formulated:

RQ1. What methodologies, features and aspects do the current literature study?

RQ2. What are the gaps of the existing study?

RQ3. What future works can be proposed under these areas?

The rest of the paper is organized as follows:

Section II defines the methodology used, in order to collate the relevant research papers for our review. Section III summarizes the findings of our review under the identified research areas. In Section IV, the future scope for the existing works are proposed uncovering the gaps associated with them. The last section concludes and summarizes our work.

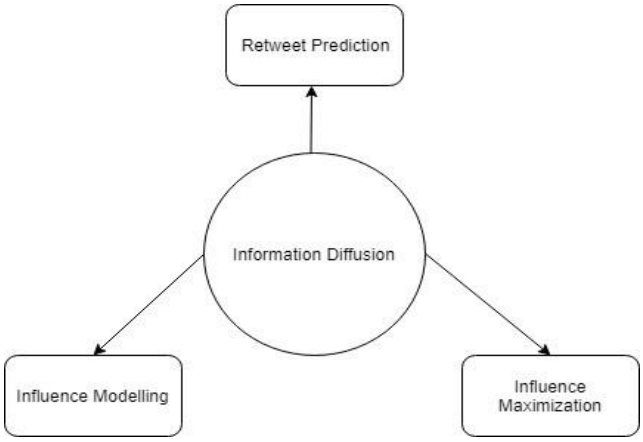

Fig. 1. Research Areas Under Information Diffusion

II. Methodology

A. Search Strategy

Once the research areas had been identified, search was performed for the relevant articles in the following databases:

- Scopus
- ScienceDirect
- ACM Digital Library
- Springer
- IEEE
- Google Scholar

The searched items included journals, conferences and miscellaneous items belonging to workshops, symposium and books.

The search string was formulated relevant to the research areas identified. The search strings used contained keywords like "influence", "model", "diffusion", "dynamic" etc. A sample search string used for our study is shown below:

"Influence" AND ("maximization" OR "maximizing" OR "maximize" OR "incremental" OR "increase" OR "optimization" OR "optimize" OR "optimizing") AND ("dynamic" OR "evolving" OR "evolve" OR "changing") AND "social" AND ("network" OR "networks").

The selection criteria used is similar to the one used in [1]. The first selection criteria applied to shortlist relevant articles for our review was filtering articles based on their abstract. The second selection criteria involved filtering articles based on introduction and conclusion. In this stage, only papers which addressed our research areas completely and were published majorly in top tier conferences and journals were shortlisted for full text review. Citation count was also considered as a tool for shortlisting papers. Majority of works included in the review had a citation count greater than 10, however, citation count criteria was not applied to recent works between the time period of 2017- 2020. Our full text review articles also included papers received through snowballing [2]. These articles were reviewed in order to get clear understanding of the papers addressing the core areas identified. Subsequently, a total of 90 articles were received and cited in our work. Fig. 2 depicts the statistics of the research items we reviewed in our work.

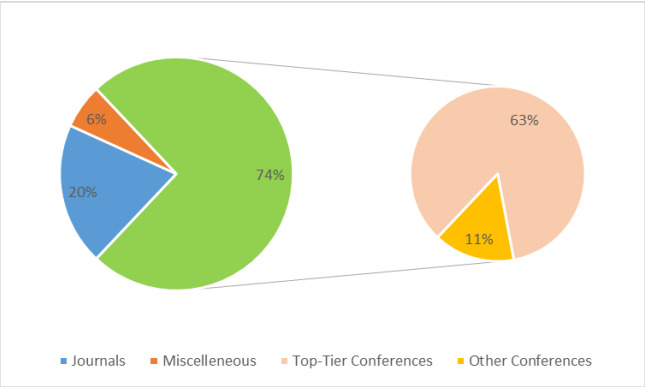

Fig. 2. Statistics of Research Items Reviewed.

B. Inclusion Criteria

The following inclusion criteria was considered while shortlisting articles for our study:

- The work published is peer-reviewed.
- The work is in English.
- It addresses the identified research areas.
- The work is validated through extensive experiments to support their hypothesis.
- It is published in top-tier conferences.

C. Data Extraction

After filtering the relevant articles for our review, following information was extracted out of each of them:

- Title of the paper
- Abstract
- Journal/conference/book in which it is published
- Author names
- Time period of the study

Research articles addressing separate areas identified in previous section were organized under distinct folders in the Mendeley Library. For each article, in addition to the information extracted, a brief about the methodology used, gaps uncovered and future scope for each study was prepared in a separate file. This enabled us to write our review more efficiently.

III. Findings

This section apart from summarizing the key findings uncovered while reviewing the past literature, will also explore the various methodologies, features, aspects associated with these areas, hence addressing RQ1 and RQ2.

Information diffusion depicts the flow of information in a network. It has several application areas like information recommendation, information prediction, viral marketing, outbreak detection, feed ranking in social networks, detecting popular topics, trust propagation [3] [20] [5] [6] [4] [7] [8] [9]. In- formation diffusion leads to flow of influence in a network. Influence refers to a user adopting another user's idea or behavior in a network. Online social networks are modeled as users depicting the nodes and the social ties between them as edges. Social influence refers to the impact a user has, on the behavior of another user in a network.

A. Current Research Areas under Information Diffusion

1) Modeling influence in a social network: This review section focuses majorly on works dedicated to learning influence probabilities between users. Though, we do provide a categorization of the present influence models used and studied, however, major part of the review under this section will include learning based models where influence probabilities are learned through different features.

Influence models predicts the probability between two users in a social network. This probability denotes the probability by which a user can influence another user in the network. Edge strength and node strength are used to measure the tie strength between users in a social network. Studies in [10] [11] used these metrics to measure influence, a user exerts on another user. Edge strength uses Jaccard coefficient metric to find out the strength of relationship between two users. The node strength, on the other hand, denotes the overall importance of a node in a social network, using centrality as a metric.

One of the earliest works in modeling influence diffusion were given by [10] [12]. These models work on the presumption that a user adopting a new behavior is a function of his/her number of activated neighbors (who have already adopted a new behavior). Both these models required influence probabilities to be already given as an input for calculating influence flowing through a network.

This work of modeling influence probabilities between users can be segmented into the following categories.

- Independent cascade model. Independent cascade model begins with a set of initially activated nodes. The activation of other nodes through already active nodes occurs in discrete time steps. At time t_1, a node u, will attempt to activate all its inactive neighbors with a probability p_{uv}, which is independent of their past actions. However, as discussed in [12], this node can only get a single chance to activate its neighbors which means it cannot activate its neighbors at the next time stamp, t_2.

- Linear Threshold Model. The Linear threshold model, on the other hand, also starts with a set of activated nodes. A weight w_{uv}, measures the probability with which a node u gets influenced by another node v (its activated neighbor). Each node which was activated at time stamp t continue to remain active at the next time stamp t+1. A node u with a specific threshold, θ_u, becomes activated, when the incoming weight of all its activated neighbors, v, exceeds the threshold of that node u [10] Mathematically, this can be denoted by:

$$P_{v \in \text{activated neighbour of } u} b_{uv} > \theta_u$$

Influence models have also been created as a generalization of Independent cascade model. Author in [13] designs a model based on topic information. Prior influence models like Linear threshold model and Independent cascade model take as an input, the probability with which a user will influence his/her neighbor. In addition to this probability, the authors also consider the probability that a user u reads a blog (denoted by r_{uv}) and the stickiness of the topic, S. If a user u reads a blog on a topic t, but the content is not sticky, then it will not be propagated from user u to v. Like the Independent cascade model, this model, too, has only once chance of activating a user v by another user u at a given timestamp.

However, a user, in real time, can try to activate another user multiple times. Also, the probabilities are not estimated rather than taken from a probability distribution which is far from realistic. Another assumption considered is that, if one user gets infected, he is always infected. Recent works [14] based on Independent Cascade model and Linear Threshold Model presents a reverse IC and LT model aiming to solve the Viral cost minimization problem. These models discard the assumption of the seed nodes being activated initially and hence show that the influence in such a case flows in the reverse direction as the focus now is on activating the seed node. However, the learning probabilities for their model is also drawn from a uniform distribution.

- Dynamic Models. Influence propagation, in real scenario, happens over time. As the network is continuously evolving, influence of a node changes from time to time. Therefore, to overcome the limitations of static influence models, [15], suggested dynamic models. These models included snapshot model and time ordinal model. The snapshot model captures an observation of a network state and time ordinal model was used to capture moment by moment influence of a network. However, storing snapshots required extensive memory and re-estimating probabilities in time ordinal model was time-consuming.

- Time based influence Models. These models considered the aspect of a node being influenced by another node within a given time. These models can further be classified into:

  - *Discrete time based models*: are the models which consider the influence diffusion in discrete time steps. The independent cascade model and Linear threshold model considered the influence diffusion in discrete time steps. Some of the works proposed under this branch [85] [17] [16] were similar to Independent cascade model and the Linear Threshold Model.

  - *Continuous time based models*: are the models which represent influence diffusion as continuous in nature. Author in [41] proposed continuous time models. They modeled influence probability between users as the conditional likelihood of transmission between a node i and j, given the activation time of i and j, where $t_i > t_j$, and a pairwise infection rate, $\alpha_{j,i}$. The probability that a user j can infect a user i can be represented by:

    $$p(t_i \mid t_j, \alpha_{j,i})$$

    Though, it could model influence continuously, this model is applicable to static networks. It also assumes that the probabilities are independent of each other.

- Learning based Models. Such models attempt to predict influence probabilities between users based on certain features. One of such works was proposed by [18] where they studied the independent cascade model based on user's past history. They used this model to predict information diffusion probabilities. This work considered estimating diffusion probabilities in multiple iterations for which Expectation maximization algorithm was used.

The drawback to this model is that it is not scalable to large datasets as the probabilities need to be re-estimated again for each of the iterations. Yet, another, limitation is that it assumes that a user can perform an action at most once.

One of the major works was contributed by [19] where they studied topic-level social influence. In their work, they majorly addressed the below questions:

- Studying influence propagation with respect to different topics.
- Quantifying the strength of these social influences Scaling their model to large scale real world networks.
- Scaling their model to large scale real world networks.

They used network structure information and topic distribution of all nodes to model social influence. Further, topic distribution for each node was evaluated by using the topic modeling approach.

A Topical factor graph was then proposed by them to build a unified probabilistic model which contained all the information. Also, to train their model, topical affinity propagation was employed by them. In order to scale their model to large data sets, a map reduce framework was adopted. However, the model cannot capture influence between users while building unified probabilistic model.

Author in [20] also studied social influence under topic aware perspective. However, they considered user authoritativeness and interests in a specific topic to model influence. They used Expectation maximization to learn parameters for their model.

Another benchmark work was given by [5], where they proposed the usage of an action log to estimate the probability with which a user may exert influence on another user and also the time by which a user may be influenced in that network. The action log kept track of the parents of a specific node. It also kept track of set of actions propagated from a node u and actions propagated back to u from another node v. These type of models were based on Bernoulli distribution. Each node here, had a fixed probability to influence the other node which could be measured by using maximum likelihood estimator. However, these models were independent of time as they could capture influence in the specific moment only.

Recent works suggest improved models for influence diffusion. Sentiment polarities have also been used as a feature to calculate sentimental influence between two users. Their work calculated sentiment polarity for each tweet for every user. They proposed an influence model based on positive influence probabilities and negative influence probabilities. The influence probabilities, however, were defined as a function of mentions, replies and retweets. It has also been shown that users with weaker influence probabilities tend to have a sentimental balance rather than users with higher influence probabilities who tend to tweet/retweet more [21]. An amalgamation of social influence as well as global influence was used to infer whether a user will be influenced or not given a piece of information on a network [6]. This work, apart from classifying users as active or inactive for a piece of information, also calculated diffusion probabilities with which a user may influence another user. These probabilities were a function of social influence as well as global influence of that user.

They calculated active and inactive payoffs for each user coming from their activated and inactivated neighbors in the network. If the active payoff for a user was larger than their inactive payoff, the user was marked as influenced for that piece of information. The social influence referred to the pairwise influence between users and was calculated as number of times a user retweets or accepts a certain message by his neighbors divided by the total number of messages posted/tweeted by that user. This influence was expressed as a non- negative vector of length k which varied with time. The global influence on the other hand referred to the overall influence of a user in a network. PageRank algorithm was used to calculate the influence of user in a network globally.

Author in [66] aims to predict the retweet probability for a message by a user. However, they also compute influence probabilities in their work in order to build their model. They claim that influence computation for a user can be expressed as a function of social and structural influence. Social influence depicts the pairwise influence between nodes and can be calculated as sum of random walk probabilities from all its active neighbors. Structural influence, on the other hand, can be viewed as a function of number of connected circles for a node in that network. Yet another work, proposed by [23], predicts probability with which a message diffuses in a network. According to them, these probabilities are based on features like network dynamics, message semantics, diffusion history, user preferences. A Bayesian belief network is used by them to model correlations between these features where nodes represent the features and the links represents the probabilistic dependence between them. They also used a Bayesian classifier to predict whether a link will be active or not. If the diffusion probabilities for a link was greater than 0.5, it was marked as active else as inactive.

However, this model does not work for a new user when no diffusion history is present.

Author in [67] used a concept of conforming weight and emotional conformance to measure influence between two nodes. Conforming weight used sentiments for their calculation, whereas emotional conformance denoted the degree to which a user conforms to his/her followees. Deep learning has also been used to model influence probabilities between users. In [24], author exploits the drawbacks of previous studies and claims to model social influence in social networks to further predict if a user will perform a certain action or not. Past studies have drawbacks that:

- They are based on an assumption that the probability of friends influencing a subject are independent of each other, and
- They do not consider the actions not performed by the subject (but performed by her/his friends) to learn the influence probabilities.

To overcome these drawbacks, the author used a deep neural network to model the interconnections between a subject and his active neighbors and to predict whether an action will be performed by a user or not. Each user along with his active neighbors are expressed as a one-hot vector and concatenate d into a single vector to be fed in the input layer of

the neural network. The output of the neural network is designed to be either a 1 or a 0. The output is 1, if given an action, the user performs the action else its 0.

The limitation to this model is that it does not clearly depict the strength of influence probabilities between users rather predict the behavior of a user. As it considers past history of subjects, the model is susceptible to failure when past history is not available. Another research proposed by [25] takes into account the heterogeneity among associations between nodes. They present MIF model to capture influence based on interactions, friendships, tags and topics discussed between nodes of a social network. Author in [26] also argues that influence probability in real world should not be uniform, hence propose, empirically motivated influence models where the probability can be modeled based on influence and susceptibility. A description of the features and the gaps in the existing study under influence models is summarized in Table I.

2) Maximizing influence in a dynamic network: Significant amount of work has been proposed under the area of Influence maximization. This problem can be defined as follows.

Given a social network G, and k number of nodes to be seeded, the problem of influence maximization is defined as finding k number of nodes which have their maximum expected influence spread throughout the network. The expected influence spread of a node is the number of nodes it can influence or activate throughout the network.

TABLE I. INFLUENCE MODELS

S.No	Reference	Characteristics	Gaps/Limitations
Models based on IC and LT Model			
1.	[10] [12] [13] [14]	Either fixed probability or derived from a prob- ability distribution	1. Assumes that an action is performed atmost once by a user. 2. Influence probabilities are required to be given as input rather than estimated. 3. Assumes the probability of one node influencing the other node is in- dependent of each other.
Dynamic Models			
1.	[15]	Snapshots and moment by moment influence measurement	1. Requires memory to store snapshots. 2. Slow as it requires time to re-estimate influence probabilities
Time Contrained Models			
Discrete Models			
1.	[85] [17] [16]	Time	1. Assumes that an action is performed atmost once by a user 2. The probabilities are not estimated rather to be given as input. 3. Probability of one node influencing another is independent of each other.
Continuous Time Models			
	[41]	time	1.Assumes probability of one user influencing other is independent of each other. 2. Can only be applied to static networks.
Learning Based models			
1.	[18] [5] [20] [6] [23] [24]	user/diffusion history	1.Not scalable to large networks 2. Assumes that a user- can perform an action atmost once 3. If the diffusion history is not present, prediction for a new user cannot be done
2.	[19] [20]	Topic based	Model cannot capture influence while building the unified probabilistic model
3.	[21] [67]	Sentiments	These models were independent of time
4.	[25] [26]	Associative	These models can further exploit senti- ments and emotions

Viral marketing is one of its application areas where k number of initial users are distributed free samples which act as an activator for other nodes in the network. These nodes influence other nodes by their word of mouth. The aim is to choose the k number of nodes efficiently so that maximum number of people can be activated or influenced at a given time. This problem was first investigated by [3]. They proposed a probabilistic model for selecting influential people in the network so that the influence spread is maximized. Author in [27] viewed the problem as a discrete optimization problem. As the problem is NP hard, they used a greedy algorithm which guaranteed that the spread can be approximated within 1-1/e of the optimal influence spread. They proposed Linear Threshold model and Independent cascade model as diffusion models. The input to their algorithm is the influence model to be used and an integer k, representing the number of nodes to be seeded. Their algorithm started with an empty seed set S and selected a node v (currently not present in the seed set) in an iterative manner which has the maximum marginal gain. This node having the maximum marginal gain was then added to the current seed set S. This process was repeated until k, number of seeds were selected. At the end, the final seed set was returned. The classic greedy algorithm presented by them can be summarized below in algorithm 1 where: σ (S) represented the influence spread of a seed set S.

Algorithm 1:

1. S= ϕ
2. i=1
3. while (i!=k)
4. u = argmax $_{v \in V}$ (σ (S ∪ {v})- σ (S))
5. S = S ∪ {u}
6. i++
7. end while
8. return S

Though their algorithm outperformed the existing classic degree and centrality based heuristics but suffered a major drawback of slow running time.

Influence Maximization, however has been studied with respect to several aspects as discussed below:

- Simulation Based Influence Maximization. This approach is based on Monte-Carlo simulations. In this approach, the influence spread over a seed set S , σ (S) was calculated over r rounds of simulations and finally the average of these rounds was returned as the estimated influence spread of the seed set. This provides a theoretical guarantee to accurately estimate influence spread in a network. The earlier approach proposed by [27] was not scalable to larger networks, so several other approaches were proposed for efficient computation of the influence spread. The next improvement was given by [28] in the form of CELF algorithm. Their algorithm exploited the concept of sub-modularity to reduce the number of calculations to be done on the nodes for computation of the influence spread. Their results showed that their algorithm was 700 times faster than the original greedy algorithm but yet it took hours to compute spread on few thousands of nodes. CELF++ was yet another improvement over CELF to reduce the unnecessary computations done in CELF algorithm [68].

Other techniques, [29] [30] [31] [32] also contributed to enhancing the scalability of their algorithms for influence maximization. Author in [29] computed influence spread for all nodes in the network by focusing on specific communities. Their algorithm first detected communities in a graph G, and then used dynamic programming for selecting the communities to identify influential nodes. The algorithm is designed to give provable guarantees and run on larger networks, however, this work focuses on finding top influential nodes and not specifically on Influence maximization. It is not necessary that the set of influential nodes identified will maximize the influence spread in a network as it may yield overlapping set of activated nodes by each influential node.

Another work of Influence maximization based on simulations was given by [30].

Author in [31] adopted pruned Monte Carlo simulations under the Independent cascade model. Their algorithm is based on the classic greedy algorithm, therefore, provides a theoretical guarantee of the influence spread of the nodes. Their algorithm first generated random graphs from the network G and then constructed Directed Acyclic Graph (DAG's) from them. The marginal influence for each node was then approximated by the average of the total weight of vertices reachable from a single vertex in each DAG. Nodes were then seeded using the greedy strategy and finally the DAG's were updated for the next iteration. In order to make their algorithm faster and efficient, they used pruned BFS and avoided gain re-computations techniques.

Author in [32] provided improvisations over CELF/CELF++ greedy based algorithm. They designed an algorithm called UBLF which estimated upper bounds of influence spread over all nodes, u ∈ V, where V is the set of nodes in the network. This could avoid the first few iterations that CELF/CELF++ considered in their algorithm. However, they present their model only under Linear Threshold and Independent cascade models and can be extended to other models as well. Yet another work was proposed by [25] where they considered the heterogeneous associations between nodes. They proposed a new influence model based on various factors like friendships, tags, interactions and topics but used the simulation approach to achieve influence maximization. Future directions of their work includes extending their influence model to incorporate sentiments and emotions and application of their IM solution to dynamic networks.

- Heuristic based Influence Maximization: These methods are more scalable to larger networks than works published under simulations but yields a poor quality of solutions as there exists no theoretical guarantee for the same.

Few algorithms under this category were suggested by [69] [33]. Author in [69] proposed an algorithm called PMIA, which considers that major influence for a node, u, flows in a local tree structure which is rooted at u. This work, avoided the need for simulations for computing the influence spread for a node in a network. Rather, the algorithm could compute the

influence exactly. It ignores all the paths whose propagation probabilities were less than a specific threshold, θ.

Author in [33], on the other hand suggested a similar approach for influence computation including the usage of pruning techniques to prune all paths whose probabilities were less than a specific threshold.

Both the algorithms were modeled under the Independent cascade model and can be extended to other models as well. Other authors, presented similar algorithms, but under the Linear threshold model.

In [70], the author claimed that the major influence in a network flows only in a small neighborhood and hence one needs to consider a small neighborhood for calculating influence and finding seed nodes for the given problem. The author, here, constructs local acyclic graphs for every node and calculates influence only in this neighborhood for each node instead of calculating it for the whole network. As the influence is calculated in LDAG's for each node, and not in the whole network, the influence computation is faster and hence can be applied to large networks. The influence maximization algorithm employing the greedy approach is then applied to these LDAG ' s and therefore, the updating of influence probabilities has to be done only for a few nodes and not all of them. The drawback to this approach was that, it required huge memory to store all the DAG's for influence computation.

In [71], the author, however used enumeration of simple paths for every node in the network. As the enumeration of all the simple paths in a network is also a NP hard problem, the algorithm restricted enumerating paths in a small neighborhood. They further adopted pruning strategies for paths with propagation probability smaller than a threshold, θ. The influence spread computation of a node was calculated as summation of the propagation probabilities falling on all the simple paths originating from that node. Influence maximization problem is then approximated using CELF algorithm. In order to further optimize, vertex cover optimization is employed to reduce the first iteration time of CELF.

Another algorithm, IRIE, [34], integrated both influence ranking and influence estimation methods under the Linear Threshold Model and Independent cascade model. This algorithm was proved to be scalable to larger networks. They used a global influence ranking method based on the belief propagation approach. Finally, the node with the highest rank was selected as the first seed node. For the selection of subsequent seed nodes, simple influence estimation was used.

Author in [72], used the PageRank algorithm for the problem of influence maximization. They computed the pagerank score for each node in the network and iteratively added nodes in the seed set, which yielded maximum marginal gain.

Recently, [35], designed an algorithm called, EASYIM, for influence maximization.

They claimed that previous models only considered positive influence from one node to other. However, their model evaluates probabilities in terms of opinions of nodes which consider negative influence as well. The author defines the interaction probability between two nodes u and v as the fraction of times a content is shared by u which is also adopted by v. The EASYIM algorithm defined by the author uses score assignment to perform effective and efficient seed selection. Every node u is assigned a score based upon the contribution it does to all the other nodes in the network. The contribution of a node, u, is evaluated by aggregating its contributions of all u-> v paths of length, less than l (a given parameter to control searching over all paths). The contribution of a path is defined as the product of all probabilities between edges on that node. In every iteration, the node with the maximum score is selected as the seed node. The process is repeated till we find k seed nodes where k is an input to the algorithm. Another contribution was proposed by [14] where they worked on a special case of Influence Maximization. They discarded the assumption of seed nodes being activated initially and presented a reverse IC and LT model to influence the seed nodes. A viral cost minimization problem was hence designed to achieve the goal of minimizing the cost to influence the seed set.

- Sketch based Influence Maximization. These methods address slow computations of influence spread as well as provides accurate guarantees. Such methods compute a number of sketches and evaluate influence spread based upon these sketches.

Author in [36] constructed sketches by extracting R snapshots of the entire network, G. These snapshots were generated by removing each edge, $< u,v >$, from the graph, G , with a probability, 1-p(u,v). Influence spread for each node, v_s, (where S is the seed set), was then calculated on these sketches. The node with the maximum marginal gain was then added to the seed set.

Author in [37] further introduced the SKIM algorithm which could scale to billions of nodes. They constructed reachability sets of a node across several propagation instances. A reachability sketch of a node comprised of nodes reachable form it. A combined reachability sketch of a node, captured its influence coverage across l instances. Their algorithm computed combined reachability sets till the time a node with maximum estimated influence was discovered. This node was then added to the seed set. The sketches were then updated where the node, just added to seed set and all the nodes, it can influence, were removed from the sketch. This process was repeatedly performed on the residual sketches till k number of nodes were seeded in the set.

The limitation to this approach was that it used extensive memory to store the sketches. Also, they assumed that the 1) propagation instances were given as input and 2) if a node gets activated, it will activate all the neighbors.

In [86], author proposed a Reverse influence sampling approach. Their algorithm samples a random number of nodes in the graph and builds reverse sketches to find out which nodes influence the sampled nodes. Other works further focused on reducing the number of sketches [38] [73] [39].

Author in [38] devised a two phase influence maximization algorithm (TIM) for maximizing influence in a network. It is

proved that this algorithm gives an approximation guarantee similar to greedy algorithm and takes an hour to run on a million node graph with a billion edges for specific setting of parameters.

This algorithm samples a random number of node s and generates its corresponding reverse reachability (RR) sets. The node which appears in majority of the RR sets is considered to be a seed node. For further iterations, the same process is repeated but the nodes activated in the previous round and the ones reachable through it are eliminated. The limitation of this approach is that it is susceptible to small number of seed nodes, which further increases the processing time. Also, as per [39], the number of samples generated can be greater than a specific threshold θ and these thresholds are not proven to be minimal. It also has an underlying assumption that once a node becomes a seed node, it will activate all its neighboring nodes.

Moreover removing such nodes over and over again results in lesser number of RR sets. In [73], author provided an extension to the TIM algorithm under the continuous diffusion model. They use the shortest path algorithm and reverse sampling techniques to reduce influence computation time on sketches.

Author in, [39] suggested another algorithm, Stop and Stare, for the above problem, aiming to address the shortcomings of the TIM algorithm. They devise two algorithms namely, SSA and D-SSA, both of which aimed at achieving minimum thresholds for the RIS samples. Their algorithm doubled the number of samples and then stopped to check the quality of the current solution. However, according to [40], there exists certain gaps in their study. Also, [26] exploited the drawbacks of IC and LT Model and proposed a model where the influence probability could be modeled as the inner product of influence and susceptibility. However, to achieve the goal of Influence Maximization, two phase IM approach was used which had its own drawbacks as discussed earlier.

- Time based Influence Maximization. This section of the review documents those work which focused on maximizing influence of a seed set within a given time. Earlier approaches did not consider the time constraint and the diffusion stopped once there were no more nodes to be influenced. However, time based influence maximization process stops when at most k number of individuals are influenced within a given deadline. Reference [85] studied maximizing influence in a network with a given deadline under the independent cascade model. They used heuristic algorithms including the dynamic programming procedure to compute exact influence. They claimed that their work achieves the same influence spread as that of the classic greedy algorithm approach and runs faster than the existing algorithms. Author in [17] proposed a new mode called CT-IC and used the simple path strategy to restrict computing influence spread. Their model was an extension to the Independent cascade model and incorporated time constraint and continuous influence flowing from one node to the other. This meant that a node had multiple chances of activating its neighbor, unlike the Independent cascade model. Other similar works were proposed by [16] [74]. However, [74] used the greedy strategy for the computation of influence spread. The above works were based on discrete time steps. However, [87] [42] [73] [43] were some of the works published under continuous time model. Recently, [44] propose approaches for approximating Influence maximization under continuous time model. They considered that the influence propagation decays over a time period and used the CELF algorithm for influence computation. However, their algorithm provided no theoretical guarantee for the same.

- Dynamic influence maximization. The techniques discussed above are suited to static networks only and does not take into account evolving nature of networks. The social networks, today are dynamic, where new nodes and relationships are added or deleted after specific intervals. Therefore, the aim is to now find a set of k -seeds in a dynamic network which keeps on evolving with time. As the network is dynamically evolving, the nodes which are taken as a seed at time t might differ at time t+1. The challenge in this area is to update this seed set after each timestamp such that the expected influence of these nodes results in maximum number of activated nodes with minimal time complexity.

Few techniques have been suggested by [45] [46] [75] [76] for influence maximization under dynamic networks.

Author in [45] uses a probing algorithm to probe only a portion of the nodes of the network at each timestamp and partially updates the network, thereby saving upon time and cost. Their goal is to minimize the error between the observed state of the network and the actual state of network. They further use degree discount heuristics to perform influence maximization. The influence model used is Independent cascade model. Influence maximization has also been worked upon in unknown social networks. Authors in [46] claim that existing works require complete topological information about the network before finding the set of seeds in order to maximize influence under them. They devise a probing algorithm to exploit only 1-10% of the topological information in order to maximize the influence spread for the seed nodes under the independent cascade model. However, their work does not include a comparison with other state of art techniques proposed by other researchers for dynamic influence maximization. Also, they use degree as a measure for maximizing influence.

Other authors propose an incremental strategy of seed selection at discrete time steps of network change. It claims that, given a network state and the topology change at time t, the network state at time t+1 can be constructed. Their algorithm uses live path strategy to compute influence spread. Influence spread computation is done only for the nodes affected by the topological change and hence is faster than the previous techniques [75]. To further improvise their seeding strategy, they suggest pruning strategies to further limit the search space for seed nodes at the subsequent time step. They use the Linear threshold model and require extensive memory. Recent authors [76] propose to maximize influence under a series of static snapshots of a network. They use interchange heuristic as a technique to update their seed set. At time t+1, a particular node v_s (already in seed set at time t) is replaced with the nodes in V-S where V is the set of nodes in the graph

and S is the seed set at time t . The node in V-S which gives the maximum replacement gain with v_s is chosen to replace v_s in the seed set at time t+1. To boost the efficiency of the algorithm, the upper bounds of all nodes in V-S are calculated. If any node u ' s replacement gain is larger than the upper bound of any other node v ' s gain, then node v ' s replacement gain need not be calculated. Thereby, decreasing the computations of replacement gain for all the nodes. The algorithm stops searching for node v_s if the largest replacement gain is less than a given threshold. The upper bounds for all nodes are updated as the network evolves. The drawback to this approach is that it require huge memory to store the snapshots of the network. Influence maximization has also been explored in areas specific to a topic or a location [47] [48]. However, we restrict going into details of these methods as our review attempts to touch several research trends in information diffusion instead of focusing on influence maximization in detail. Table II illustrates a summarized information about various aspects of influence maximization studied and the gaps associated with them. These studies attempt to optimize computational time to solve the problem of IM.

TABLE II. INFLUENCE MAXIMIZATION

S.No	Reference	Aspect Studied	Gaps/Limitations
1.	[27] [28][29] [68] [30] [31] [32] [25]	Simulations	Either not scalable or expensive
2.	[69] [70] [71] [34] [33] [72] [35] [14]	Heuristics	Either yielded a poor quality of solu- tions or provided no theoretical guarantee or required huge memory
3.	[36] [37] [86] [38] [73] [39] [26]	Sketch Based	These approaches either required extensive memory to store the sketches or were run under an as- sumption that if a node gets activated, it will activated all its neighbors. They were also susceptible to small number of seeds.
4.	[85] [17] [16] [74]	Discrete time	Considers diffusion in discrete steps. Also, they were only applicable to static networks..
5.	[87] [42] [73] [44] [43]	Continuous time	Either provided no theoretical guarantee or required huge consumption of memory
6.	[45] [46] [75] [76]	Dynamic network based	Either used IC or LT model for determining influence probabilites or required huge memory for storing snapshots.

3) Retweet prediction: Retweeting is an activity performed on a social network, Twitter, where, a user when influenced by the idea of another user, re-posts his/her message [49]. The message is known as a tweet and re-posting a friend's message is called as retweet. This section focuses on discussing the following research questions under this area:

a) Predicting whether a given tweet will be retweeted or not.

b) Factors affecting the retweetability of a tweet.

c) Predicting whether a specific user will retweet a given tweet or not.

Our rigorous review uncovered that the work under this area can be classified into two types of approaches:

Deductive Approach (Non-Predictive). This approach includes non-predictive models under retweet prediction. These models do not attempt to predict behavior rather deduce the factors affecting retweet behavior.

Reference [50] studied the various reasons and styles of retweeting, a user adopts in a social network. They used the content of the tweet to analyze why and how people retweet using a case study methodology. Another stream of work [77] emphasized that number of followers and followees have a big impact on retweetability of a tweet. They also proved that retweetability of a tweet correlates with the use of URL's and hashtags in the content of the tweet. They used a single layer perceptron model for this problem.

Few other authors [51] also worked on the same problem and concluded that features relating to user, tweet, sentiments and emotional divergence also correlates highly with the retweet frequency. Author in [78] also showed that sentiments and emotional divergence affect virality of a tweet using the sentistrength classifier. Author in [52] studied what makes people retweet in a network using user profile information, their topic of interest and content of the tweet. They used probabilistic methods to validate their study.

Inductive approach (Predictive approach). This approach covers models predicting retweet behavior of a tweet or a user. In [53], author suggested a factor graph model to predict users retweeting behavior. They addressed the problem of predicting i) whether users will retweet, a tweet message m, to their friends after reading it, given a set of tweets and a history of retweet behavior of users. ii) the range of the spread of a tweet, m , by a user, u. They also proved through extensive experiments that a user retweet behavior is influenced by factors like user information, tweet information and time of the tweet. Author in [65] attempted to predict the retweet probability of a tweet. They used content based features and logistic regression for their prediction. Further, [84] used conditional random fields to predict the retweeting behavior of each user within a network for a given tweet as a function of content and network based features. Author in [49] used social and tweet features to predict whether a tweet will be retweeted. They built models using human experiments as well as machine learning algorithms based on tweet ' s creation time.

Few authors [54] integrated structural features, content of the tweet, metadata features and temporal information to predict:

1) the retweetability of a tweet and 2) whether a message will be retweeted. The author used a logistic regression classifier for the same.

In spite of huge amount of work being done in this area, majority of the above works did not include a detailed analysis of feature extraction and selection. Other researchers in [22] attempted to predict who will retweet a message. They found that user's retweet history, status, active time and interest can act as promising features in predicting their retweet behavior. They proved using ranking methods that followees who retweeted or mentioned author ' s tweets frequently before are more likely to be retweeters.

Recent works [55] include addressing problem of retweet pre- diction on the request of a stranger. They study whether a user will retweet a given tweet if a stranger requests them to do so. Their model considers features like, user profile, social network, personality, activity, past retweeting behavior, and readiness of a user to retweet. In this work, the authors created twitter bots to send request messages for retweeting their tweet to people specific to a location or interested in a specific topic. They conducted extensive experiments under machine learning algorithms like random forest, AdaboostM1 for retweet prediction.

In [66], author worked on predicting a user's retweet behavior. They used Logistic regression and factor graphic model to predict the retweet probability of a user for a message. They considered personal attributes, topic propensity and instantaneity, in addition to influence locality features for modeling the prediction.

Other authors [79] used structural, textual and temporal information to determine whether a user will retweet a message or not. They used non-parametric statistical models in their work.

Majority of these works either did not include a detailed analysis of feature extraction or did not provide a comparison with the state of art techniques used for retweet prediction.

Author in [80] proposed models based on structural, user and tweet information to predict individual retweet behavior of users. They used logistic regression as a technique for predicting individual retweet behavior.

Few authors [56] suggested models based on one-class collaborative techniques. They claimed that user interest similarity and social influence can be used as promising features for predicting user's behavior.

Few researchers have also proposed neural network approaches for the above problem. As feature engineering is a laborious task, neural networks have been suggested by authors to achieve state of art performance. The authors in [81] propose an attention based neural network for approaching the problem of retweet prediction. They use word embedding to represent user, his attention interests, author and the tweet.

Recently, [57] constructed a user retweet behavior prediction model based on RBF (radical basis function) neural network. They also introduced another model called C- RBF (cloudbased RBF) using fuzzy which could incorporate the uncertainty in a user's behavior. Based on user profile information and historical behavior of a user, they analyzed number of potential users participating in a specific topic at different time periods using discrete time methods. The future scope of this work includes using deep learning techniques for the same problem. Another model proposed by [58] considers user preferences and the current hot topics they indulge in. They use a masked self-attentive model to achieve the goal of retweet prediction.

In [59], author build a topic specific model for predicting the retweet behavior. They used user level features, the content of the tweet, tweet/retweet history and emotions as features in their analysis. LDA was used to perform topic extraction.

However, all these methods often have a drawback of introducing noise while extracting the feature set and are incapable of capturing the context in a complete fashion.

Another way to approach the problem of retweet prediction was modelled as a recommend system based on matrix factorization. One of the works based on matrix factorization [82] use social information of a user and the message semantics to predict retweet behavior of a user based on his social network. Other works based on matrix factorization includes [60] [83] [61]. However, matrix factorization methods are unable to capture contextual information completely. Recently, [62], suggested a new model for retweet prediction based on matrix tensor factorization. They propose to capture contextual information using user similarity, message similarity and pairwise influence between users. A further direction for this work can be considering user similarity based on the type of the user (occasional or frequent), emotions and beliefs a user normally associates with.

This research area also includes other problems predicting the frequency at which retweets occur [63], counting retweet times of a tweet [64]. However, these have purposely not been included in the review to restrict the scope of our paper. Table III presents a description of the features and the methodology adopted by the existing studies under retweet prediction. These studies attempt to optimize the accuracy to solve the problem of retweet prediction.

TABLE III. RETWEET PREDICTION MODELS

S.No	Reference	Features	Methodology
Non-Predictive Models			
1.	[50] [77] [51] [52]	user level/tweet fea- tures	[50]Emperical Methods [77] Single layer perceptron model [51] Naive Bayes model [52] Probabilistic Methods
2.	[51] [78]	Sentiments/emo- tions	[51] Naive bayes [78] Senti-strength classifier
3.	[65] [84] [49] [55] [80] [82] [59] [58] [83]	Structural/user/tweet features	[65]Logistic re- gression [84] Conditional random fields [49]PSA algorithm [55]AdaboostM1 [80]Logistic regression [82] Probabilistic Matrix Factorization [59] Conditional Probability methods [58] Self attentive model [83] Matrix Factorization
4.	[53] [54] [22] [79] [57]	Temporal infor- mation with other features	[53]Factor graph model [54] Logistic regression [22] Ranking methods [79] Statistical models [57] RBF Neural Net- work
5.	[66] [56] [62] [60] [61]	Social influence along with other features	[66]Logistic regression and Factor graph model [56]One-class Collaborative Filtering [62]Matrix Tensor Factorization [60]Matrix Factorization [61] Matrix Factorization

IV. FUTURE SCOPE

This section aims at addressing RQ3. The research areas reviewed in the paper mainly addresses influence modeling, maximizing influence in dynamic networks and retweet prediction.

Early influence models either required influence probability between users to be given as input or it was derived from a fixed probability distribution. Influence models based on time are also generally an extension of Independent cascade model or Linear Threshold Model. However, these models also have the certain assumptions/drawbacks as listed below:

- The influence probability are independent of each other.
- If one person is activated, all their neighbors will be activated.
- Requires updating of influence probabilities with time. Re-estimating the probabilities temporally for the whole network is time-consuming.

Recent works show, considering stronger features like sentiments for modeling influence leads to better results. There- fore, our work proposes using user based features (including emotions, value systems and beliefs), topic based features to capture user interests, and structural features as optimal feature set for improving the accuracy of such models.

Another work that can be put forward is studying these features and their impact on influence temporally. While reviewing, we also found that not much work has been done on learning activation threshold for Linear Threshold models. This is another area of future scope for the researchers.

Another area that was under review was that of, maximizing influence in dynamic social networks. To the best of our knowledge, a major gap in this area is that all the approaches used to solve this problem either use Independent cascade model, Linear Threshold Model. These methods have their own drawbacks as stated in the previous sections. The proposed solutions in this area are summarized below:

1) Using the proposed influence models described above for maximizing influence in a network.

2) Integrating probing techniques and parallel processing to improve the performance in dynamic networks.

The last area reviewed in the paper addresses retweet prediction. After performing a rigorous review in this area, we uncovered that the optimal feature set can include structural features, user based features including personality, value systems, user interests, content based features, and influence features as the optimal feature set.

V. CONCLUSION

This paper conducts a review on key research areas identified under information diffusion in social networks. A review is presented under the areas of influence modeling, influence maximization in dynamic networks and retweet prediction.

An appropriate search strategy was adopted by the authors in order to review relevant research papers. In total, 90 papers were marked as relevant and were analyzed.

Key findings under each area were extracted, revealing the corresponding gaps and future scope under them. With a

deeper analysis of the current literature, we can conclude that previous works can be extended and improvised by studying the impact of features like emotions, values and beliefs on each of them. Also, as the use of deep learning approaches can learn optimal feature selection, such techniques can be used to obtain state of art performance. It was also uncovered that efficient seed selection in dynamic networks can be done by integrating probing techniques and parallel processing. This may further reduce the running time of the algorithm and yield a better seed set in dynamic networks.

To summarize, this work can contribute to the researchers who are doing their initial review on information diffusion in social networks. It will enable them to identify the key trends under this area and understand the gaps in the existing studies. These gaps can be exploited to provide further contributions in this area.

REFERENCES

[1] F. Weidt and R. Silva," Systematic Literature Review in Computer Science-A Practical Guide",Relatórios Técnicos do DCC/UFJF 1, 2016.

[2] S. Jalali and C. Wohlin, "Systematic literature studies: database searches vs. backward snowballing," ACM-IEEE international symposium on empirical software engineering and measurement, pp. 29–38, 2012.

[3] P. Domingos and M. Richardson, "Mining the network value of customers", Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 01´, pp. 57–66, 2001.

[4] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks, a survey." ACM SIGMOD Record, vol. 42, no. 1, pp. 17–28, 2013. [Online]. Available: 10.1145/2503792.2503797 ;https://dx.doi.org/10.1145/2503792.2503797

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks", Proceedings of the third ACM international conference on Web search and data mining, pp. 241–250,2010.

[6] D. Li, Z. Xu, Y. Luo, S. Li, A. Gupta, K. Sycara, S. Luo, L. Hu, and H. Chen, "Modeling information diffusion over social networks for temporal dynamic prediction," IEEE Transactions on Knowledge and Data Engineering, 2017.

[7] C. Kang, C. Molinaro, S. Kraus, Y. Shavitt, and V. S. Subrahmanian, "Diffusion centrality in social networks," Advances in Social Networks Analysis and Mining (ASONAM), pp. 558–564, 2012.

[8] M. Taherian, M. Amini, and R. Jalili, "Trust inference in web-based social networks using resistive networks," Third International Conference on Internet and Web Applications and Services, pp. 233–238, 2008.

[9] J. J. Samper, P. A. Castillo, L. Araujo, and J. J. Merelo, "Nectarss, an rss feed ranking system that implicitly learns user preferences", arXiv preprint cs/0610019, 2006.

[10] M. S. Granovetter, "The Strength of weak ties," Social Networks, pp. 347-367,1977.

[11] R. A. Hanneman and M. Riddle,"Introduction to social network methods", Introduction to social network methods, 2005.

[12] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," Marketing letters, vol. 12, pp. 211–223, 2001.

[13] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins, "Information diffusion through blogspace," ACM SIGKDD Explorations Newsletter, vol. 6, no. 2, pp. 43–52, 2004. [Online]. Available:10.1145/ 1046456.1046462;https://dx.doi.org/10.1145/1046456.1046462

[14] A. Talukder and C. S. Hong, "A heuristic mixed model for viral marketing cost minimization in social networks," International Conference on Information Networking (ICOIN), IEEE, pp. 141–146, 2019.

[15] D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, X. Lan, and S. Suri, "Sequential Influence Models in Social Networks," ICWSM, pp. 26–26, 2010.

[16] B. Liu, G. Cong, D. Xu, and Y. Zeng, "Time constrained influence maximization in social networks," IEEE 12th international conference on data mining, pp. 439–448, 2012.

[17] J. Kim, W. Lee. Yu. H, "CT-IC: Continuously activated and time-restricted independent cascade model for viral marketing", Knowledge based Systems, pp. 55-68, 2014.

[18] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model", International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 67–75, 2008.

[19] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 807– 816, 2009.

[20] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," Knowledge and Information Systems, vol. 37, no. 3, pp. 555–584, 2013. [Online]. Available: 10.1007/s10115-013-0646-6;https://dx.doi.org/10.1007/s10115-013-0646-6

[21] Y. Wu and F. Ren, "Learning sentimental influence in twitter," International Conference on Future Computer Sciences and Application, ICFCSA, pp. 119–122, 2011.

[22] Z. Luo, M. Osborne, J. Tang, and T. Wang, "Who will retweet me?: finding retweeters in twitter," Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 869–872, 2013.

[23] D. Varshney, S. Kumar, and V. Gupta, "Predicting information diffusion probabilities in social networks: A Bayesian networks based approach," Knowledge-Based Systems, pp. 66–76, 2017.

[24] L. Luceri, T. Braun, and S. Giordano, "Social Influence (Deep) Learning for Human Behavior Prediction," International Workshop on Complex Networks. Springer, pp. 261–269, 2018.

[25] X. Deng, F. Long, B. Li, D. Cao, and Y. Pan, "An Influence Model based on Heterogeneous Online Social network for Influence Maximization," IEEE Transactions on Network Science and Engineering, pp. 1–1, 2019. [Online]. Available: 10.1109/tnse. 2019.2920371;https://dx.doi.org/10.1109/tnse.2019.2920371

[26] S. Aral and P. S. Dhillon, "Social influence maximization under empirical influence models," Nature Human Behaviour, vol. 2, no. 6, pp. 375–382, 2018. [Online]. Available: 10.1038/s41562-018-0346-z;https://dx.doi.org/10.1038/s41562-018-0346-z

[27] D. Kempe, J. Kleinberg, and Éva Tardos, "Maximizing the spread of influence through a social network," Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137–146, 2003.

[28] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriesen, and N. Glance, "Cost-effective outbreak detection in networks," Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420–429, 2007.

[29] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks",Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1039–1039.2010.

[30] Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, and K. Xie, "Simulated Annealing Based Influence Maximization in Social Networks,"Twenty-fifth AAAI conference on artificial intelligence, pp. 127– 132, 2011.

[31] N. Ohsaka, T. Akiba, Y. Yoshida, and K. I. Kawarabayashi, "Fast and Accurate Influence Maximization on Large Networks with Pruned Monte-Carlo Simulations, Twenty-Eighth AAAI Conference on Artificial Intelligence, pp. 138–144. URL: internalpdf://fast_accurate_ Im_smulation.pdf. 2014.

[32] C. Zhou, P. Zhang, W. Zang, and L. Guo, "On the Upper Bounds of Spread for Greedy Algorithms in Social Network Influence Maximization," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 10, pp. 2770–2783, 2015. [Online]. Available:10.1109/tkde.2015.2419659;https://dx.doi.org/10.1109/tkde. 2015.2419659

[33] J. Kim, S. K. Kim, and H. Yu, "Scalable and parallelizable processing of

influence maximization for large scale social networks?", IEEE 29th international conference on data engineering (ICDE), pp. 266–277, 2013.

[34] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 918–923, 2012.

[35] S. Galhotra, A. Arora, and S. Roy," Holistic influence maximization: Combining scalability and efficiency with opinion-aware models", Proceedings of the 2016 International Conference on Management of Data, pp. 743–758, 2016.

[36] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, "Staticgreedy: solving the scalability-accuracy dilemma in influence maximization," Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 509–518, 2013.

[37] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 629– 638, 2014.

[38] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 75–86, 2014.

[39] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," Proceedings of the 2016 International Conference on Management of Data, pp. 695–710, 2016.

[40] K. Huang, S. Wang, G. Bevilacqua, X. Xiao, and L. V. S. Lakshmanan, "Revisiting the stop-and-stare algorithms for influence maximization," Proceedings of the VLDB Endowment, vol. 10, no. 9, pp. 913–924, 2017. [Online]. Available: 10.14778/3099622.3099623;https://dx.doi.org/10.14778/3099622.3099623

[41] M. G Rodriguez, D. Balduzzi., & B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks". arXiv preprint arXiv:1105.0697, 2011

[42] N. Du, L. Song, M. G. Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," Advances in neural information processing systems, pp. 3147–3155, 2013.

[43] M. Gomez-Rodriguez, L. Song, N. Du, H. Zha, and B. Schölkopf, "Influence Estimation and Maximization in Continuous-Time Diffusion Networks," ACM Transactions on Information Systems, vol. 34, no. 2, pp. 1–33, 2016. [Online]. Available: 10.1145/2824253;https://dx.doi.org/10.1145/2824253

[44] M. Xie, Q. Yang, Q. Wang, G. Cong, and G. D. Melo, "DynaDiffuse: A Dynamic Diffusion Model for Continuous Time Constrained Influence Maximization," Twenty-Ninth AAAI Conference on Artificial Intelligence., AAAI, pp. 346–352, 2015.

[45] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, "Influence maximization in dynamic social networks," Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 1313–1318, 2013.

[46] S. Mihara, S. Tsugawa, and H. Ohsaki, "Influence Maximization Problem for Unknown Social Networks," Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1539–1546, 2015.

[47] J. Guo, P. Zhang, C. Zhou, Y. Cao, and L. Guo, "Personalized influence maximization on social networks," Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 199–208, 2013.

[48] G. Li, S. Chen, J. Feng, K. L. Tan, and W. Li, "Efficient location-aware influence maximization," Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 87–98, 2014.

[49] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter," Fifth International AAAI Conference on Weblogs and Social Media, pp. 586–589, 2011.

[50] S. Golder, Tweet, and R. Tweet, Conversational Aspects of Retweeting on Twitter, 43rd Hawaii International Conference, pp. 1–10, 2010, URL: pdf://analyze_content_tweet.pdf

[51] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," Proceedings of the 22nd International Conference on World Wide Web, pp. 657–664, 2013.

[52] S. A. Macskassy and M. Michelson, "Why do people retweet? anti-homophily wins the day!.," Fifth International AAAI Conference on Weblogs and Social Media, pp. 209–216, 2011.

[53] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding retweeting behaviors in social networks," Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1633–1633, 2010.

[54] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," Proceedings of the 20th international conference companion on World wide web, pp. 57–58, 2011.

[55] K. Lee, J. Mahmud, J. Chen, M. Zhou, and J. Nichols, "Who will retweet this? automatically identifying and engaging strangers on twitter to spread information, Proceedings of the 19th international conference on Intelligent User Interfaces, pp. 247–256, 2014.

[56] B. Jiang, J. Liang, Y. Sha, R. Li, W. Liu, H. Ma, and L. Wang, "Retweeting behavior prediction based on one-class collaborative filtering in social networks," Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 977–980, 2016.

[57] Y. Liu, J. Zhao, and Y. Xiao, "C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network," Neurocomputing, vol. 275, pp. 733–746, 2018.

[58] R. Ma, X. Hu, Q. Zhang, X. Huang, and Y. G. Jiang, "Hot topic-aware retweet prediction with masked self-attentive model", Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 525–534, 2019.

[59] S. N. Firdaus, C. Ding, and A. Sadeghian, "Topic specific emotion detection for retweet prediction," International Journal of Machine Learning and Cybernetics, vol. 10, no. 8, pp. 2071– 2083, 2019. [Online]. Available: 10.1007/s13042-018-0798-5;https://dx.doi.org/10.1007/s13042-018-0798-5

[60] M. Wang, W. Zuo, and Y. Wang, "A multidimensional nonnegative matrix factorization model for retweeting behavior prediction, Mathematical Problems in Engineering" 2015.

[61] B. Jiang, Z. Lu, N. Li, J. Wu, and Z. Jiang, "Retweet prediction using social-aware probabilistic matrix factorization," International Conference on Computational Science, pp. 316–327, 2018.

[62] B. Jiang, F. Yi, J. Wu, and Z. Lu, "Retweet prediction using context-aware coupled matrix-tensor factorization," International Conference on Knowledge Science, Engineering and Management, pp. 185–196, 2019.

[63] R. Kobayashi and R. Lambiotte, "TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics," Tenth In- ternational AAAI Conference on Web and Social Media, pp. 191–200, 2016.

[64] E. F. Can, H. Oktay, and R. Manmatha, "Predicting retweet count using visual cues," Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 1481–1484, 2013.

[65] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," Proceedings of the 3rd international web science conference, pp. 8–8, 2011.

[66] J. Zhang,, J. Tang, J. Li., Y. Liu, C. Xing,,"Who Influenced You? Predicting Retweet via Social Influence Locality" ACM Transactions on Knowledge Discovery from Data,pp.1–26,2015.

[67] Q. Wang, Y. Jin, S. Cheng, T. Yang, "ConformRank: A conformity-based rank for finding top-k influential users", Physica A: Statistical Mechanics and its Applications, pp. 39–48, 2017

[68] A.Goyal, W. Lu, L.V. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks", Proceedings of the 20th international conference companion on World wide web, pp. 47–48, 2011

[69] W.Chen, C. Wang , Y. Wang , "Scalable influence maximization for prevalent viral marketing in large scale social networks", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1029–1038, 2010

[70] W. Chen, Y. Yuan , L. Zhang , "Scalable influence maximization in social networks under the linear threshold model", IEEE international conference on data mining, pp. 88–97, 2010

[71] A. Goyal, W. Lu , L.V. Lakshmanan , “Simpath: An efficient algorithm for influence maximization under the linear threshold model”, IEEE 11th international conference on data mining, pp. 211–220, 2011

[72] Q. Liu, B. Xiang , E. Chen , H. Xiong , F. Tang, J.X. Yu, “Influence maximization over large- scale social networks: A bounded linear approach”, Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 171–180, 2014

[73] Y. Tang, Y. Shi, X. Xiao , “Influence maximization in near-linear time: A martingale approach”,Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1539–1554, 2015

[74] B. Liu, G. Cong, Y. Zeng , D. Xu, Y.M. Chee, “Influence spreading path and its application to the time constrained social influence maximization problem and beyond”, IEEE Transactions on Knowledge and Data Engineering, pp. 1904–1917, 2014

[75] Y. Wang, J. Zhu, Q. Ming, “Incremental influence maximization for dynamic social networks”, International Conference of Pioneering Computer Scientists, Engineers and Educators, Springer, 2017.

[76] G. Song, Y. Li, X. Chen, X. He, J. Tang, “Influential Node Tracking on Dynamic Social Network: An Interchange Greedy Approach”, IEEE Transactions on Knowledge and Data Engineering 29, pp. 359–372, 2017

[77] B. Suh, L. Hong, P. Pirolli , E.H. Chi ,”Want to be retweeted? large scale analytics on factors impacting retweet in twitter network”, IEEE Second International Conference on Social Computing, 2010

[78] R. Pfitzner , A. Garas, F. Schweitzer, “Emotional Divergence Influences Information Spreading in Twitter”, Sixth international AAAI conference on weblogs and social media, pp. 2–5, 2012

[79] Q. Zhang , Y. Gong, Y. Guo, X. Huang, “Retweet behavior prediction using hierarchical dirichlet process”,Twenty-Ninth AAAI Conference on Artificial Intelligence., AAAI. pp. 403–409, 2015

[80] X. Tang , Q. Miao , Y. Quan , J. Tang, K. Deng ,“Predicting individual retweet behavior by user similarity: A multi-task learning approach”, URL: https://dx.doi.org/10.1016/j.knosys.2015.09.008,doi:10.1016/j.knosys.2015.09.008, 2015

[81] Q. Zhang , Y. Gong, J. Wu, H. Huang, X. Huang ,“Retweet Prediction with Attention-based Deep Neural Network”, Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 75–84, 2016

[82] K. Zhang, X. Yun, , J. Liang, X.Y. Zhang, C. Li, B. Tian, “Retweeting behavior prediction using probabilistic matrix factorization”, IEEE Symposium on Computers and Communication (ISCC), (pp. 1185-1192), 2016

[83] C. Wang , Q. Li, L. Wang , D.D. Zeng,“Incorporating message embedding into co-factor matrix factorization for retweeting prediction”, International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1265–1272, 2017

[84] H.K. Peng , J. Zhu, D. Piao, R. Yan, Y. Zhang, 2011. “Retweet modeling using conditional random fields”, 11th IEEE International Conference on Data Mining Workshops, pp. 336–343, 2011

[85] W. Chen, W. Lu, , N. Zhang, “Time-critical influence maximization in social networks with time-delayed diffusion process”, Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI. pp. 1–5, 2012

[86] C. Borgs, M. Brautbar, J. Chayes, B. Lucier, “Maximizing social influence in nearly optimal time”, Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms, pp. 946–957, 2014

[87] M. G. Rodriguez and B. Schölkopf, “Influence Maximization in Continuous Time Diffusion Networks, Proceedings of the 29th International Conference on Machine Learning, ICML, 2012.

[88] M. Li , X. Wang, K. Gao, & S. Zhang, “A survey on information diffusion in online social networks: Models and methods”. Information , 8 (4), 118, 2017

[89] S.S. Singh, K. Singh, A. Kumar, H.K. Shakya, B. Biswas “A Survey on Information Diffusion Models in Social Networks”, International Conference on Advanced Informatics for Computing Research (pp. 426-439), 2019, Springer, Singapore.

[90] G. Tong, W. Wu, S. Tang and D. Du, "Adaptive Influence Maximization in Dynamic Social Networks," IEEE/ACM Transactions on Networking, vol. 25, no. 1, pp. 112-125, Feb. 2017.