Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Finding influential users in microblogs: state-of-the-art methods and open research challenges

Umar Ishfaq, Hikmat Ullah Khan, Shahid Iqbal & Mohammed Alghobiri

To cite this article: Umar Ishfaq, Hikmat Ullah Khan, Shahid Iqbal & Mohammed Alghobiri (2021): Finding influential users in microblogs: state-of-the-art methods and open research challenges, Behaviour & Information Technology, DOI: 10.1080/0144929X.2021.1915384

To link to this article: https://doi.org/10.1080/0144929X.2021.1915384

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2021.1915384

Finding influential users in microblogs: state-of-the-art methods and open research challenges

Umar Ishfaq[a], Hikmat Ullah Khan[a], Shahid Iqbal[a] and Mohammed Alghobiri[b]

[a]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Rawalpindi, Pakistan; [b]College of Computer Science, Department of Information Systems, King Khalid University, Abha, Saudi Arabia

ABSTRACT

Social networks are online platforms that people use for interaction, information sharing and propagation of new ideas. Finding influential users in online social networks is a significant research problem due to its vast research applications in information diffusion, marketing and advertising. The relevant literature presents several models proposed for identifying influential users in social networks. In this survey, we present a review of the most relevant studies on influential users mining in microblog networks. First, we propose a new taxonomy by classifying the influence finding algorithms into five main categories based on their underlying framework and baseline methods. Second, each study is analysed according to the proposed framework, experimental datasets, validation approaches and evaluation results. Finally, the survey concludes with discussion on applications from the relevant literature, exploring open research challenges and presenting possible future research directions. The findings of this survey indicate that influential users mining in microblogs has many applications in marketing, advertising and information diffusion. In addition, this survey can be used as a guideline, particularly by young researchers, for establishing a baseline before initiating a research or identifying attractive as well as relevant research insights.

ARTICLE HISTORY
Received 27 May 2020
Accepted 2 April 2021

KEYWORDS
Social Network; Microblog; Social Influence; Node Ranking; Twitter

1. Introduction

Online social networks (OSNs) are dynamic platforms for millions of users worldwide who share common interest, activities or real-life connections. Online social interaction exponentially accelerates information dissemination among users in the network. Ranging from discussion groups and social communities to tagging systems, mobile social networks, games, recommendation engines and virtual worlds, OSNs have transformed the way a user navigates and communicates on the internet. Therefore, user interaction in OSNs generates large volumes of data, thereby, creating opportunities to perform an in-depth analysis of human behaviour from different aspects (Lehmann et al. 2012). This analysis can help understand the online social connections between users from all over the world as well as resolve several social and societal issues.

A social network can be viewed as a graph where users are represented as vertices and relationships among users as edges. The significance of users in OSNs can be computed using different techniques from the graph theory (Girvan and Newman 2002; Mislove et al. 2007; Jin et al. 2013) such as centrality measures, PageRank (Page et al. 1999) variants or influence propagation models (Kempe, Kleinberg, and Tardos 2003). In addition, research identified that hybrid techniques which are based on a combination of one or more of the above-mentioned approaches significantly enhance the process of mining user significance in social networks (Zhao, Li, and Jin 2016b; Drakopoulos et al. 2017; Jianqiang, Xiaolin, and Feng 2017). Resultantly, information propagation in OSNs can be significantly enhanced for devising better marketing strategies (Girvan and Newman 2002; Mislove et al. 2007; Jin et al. 2013).

Social influence in OSNs refers to a user's ability to change the opinion or behaviour of other users in the network (Szell, Lambiotte, and Thurner 2010). Social influence spreads depending on the connection or relation among the network users. The connection strength between any two users depends on their overlapping neighbours (Granovetter 1973). Normally, influential individuals have considerable connection strength, however, this criteria does not always hold true (Cha et al. 2010). Relevant literature on information dissemination reflects that only a few of the influential users qualify the criteria to form and

CONTACT Hikmat Ullah Khan hikmat.ullah@ciitwah.edu.pk Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt 470040, Pakistan

influence the opinion of a large number of members of an online social community (Katz, Lazarsfeld, and Roper 1955). Recently, the domain of influential user mining has gained significant importance. Identifying influential users has several applications (Lü et al. 2016a), particularly, in accelerating information propagation for marketing various types of products and services (Leskovec, Adamic, and Huberman 2007) or preventing the spread of disinformation (e.g. online rumours, negative behaviour, viruses, etc.) (Zhao et al. 2012; Ma et al. 2015; Abawajy, Ninggal, and Herawan 2016). In addition, examining influence patterns can help understand rapid adoptions and variations in specific trends that can greatly benefit the advertisers for implementing effective marketing campaigns. However, it has been argued that many of the easily influenced individuals are the main cause of the rapid dissemination of information in OSNs (Watts and Dodds 2007). Therefore, quantifying influence and identifying true influentials is significantly important from both analysis and design perspectives.

In this survey, we strive to present a comprehensive analysis of diverse algorithms for finding influential users in microblog networks. Reviewed algorithms have been classified into five main categories based on their underlying techniques, baseline methods and novelties. In addition, several techniques have been studied for evaluating the performance of influence finding algorithms. Furthermore, the survey presents some of the potential applications and major research issues most relevant to the domain of influential users in OSNs. To the best of our knowledge, the proposed taxonomy of the models for influential user mining and open research challenges in microblog networks is the first of its kind.

The rest of the paper is organised as following: Section 2 presents the taxonomy and analysis of relevant algorithms on identifying influential users; Section 3 presents most relevant real-world applications; section 4 highlights the current research issues and section 5 presents the conclusion based on the discussion in Section 2, Section 3 and Section 4, respectively.

2. Taxonomy of existing studies on finding influential users

This survey presents a novel taxonomy of the most relevant existing literature on finding influential users in microblog networks. Research has identified that straightforward validation process for the effectiveness of influential user mining algorithms in OSNs cannot be achieved as full diffusion details are unavailable (Khan et al. 2017). One of the key reasons includes privacy concerns and technical issues arranged by social network users. The proposed taxonomy includes five broad categories as shown in Figure 1. In each category, respective models are chronologically organised and investigated given as following:

2.1. Using PageRank-based models for influential user analysis

PageRank-based techniques utilise the classical PageRank or its variants for identifying the *top-k* influential nodes in social networks. In this study, PageRank-based models are further divided into two different categories given as following:

2.1.1. Models based on temporal features

Temporal PageRank algorithms capture the time-bound user influence based on the notion that user influence in social networks is highly dynamic which implies that user influence decays with time. Therefore, it is important to consider the temporal features while analysing OSNs and ranking the top-k influential users (Riquelme and González-Cantergiani 2016). In the existing literature, several algorithms have been proposed which employ the classical PageRank and its variants as well as time-bound features of social networks. Following is a brief discussion on the most relevant and recent temporal PageRank-based models:

2.1.1.1. Fading memories and raw reputation ranking scheme (FadeRank). In social networks, user reputation is an important indicator that can be exploited for promoting authoritative content and marginalising the spammers. However, reputation must be periodically updated using the entire history of user activity. In big social networks like Twitter, these updates cause serious performance issues. Research has been proposed which addresses this issue of updating the reputation of users in the most recent time window by integrating with the summary of historical activities in social networks (Bartoletti, Lande, and Massa 2016). The authors aggregate user's historical data in constant time and space by employing the fading memory technique (Srivatsa, Xiong, and Liu 2005). User reputation is computed by considering user's recent (raw) behaviour, user behaviour in aggregated history and gradient of behaviour change.

One of the novelties of the study is that the average execution time for processing data at each update remains constant as compared to baseline techniques. In addition, FadeRank integrates the fading memories with user's raw reputation which significantly reduces the whitewashing attacks. Fading memory is a

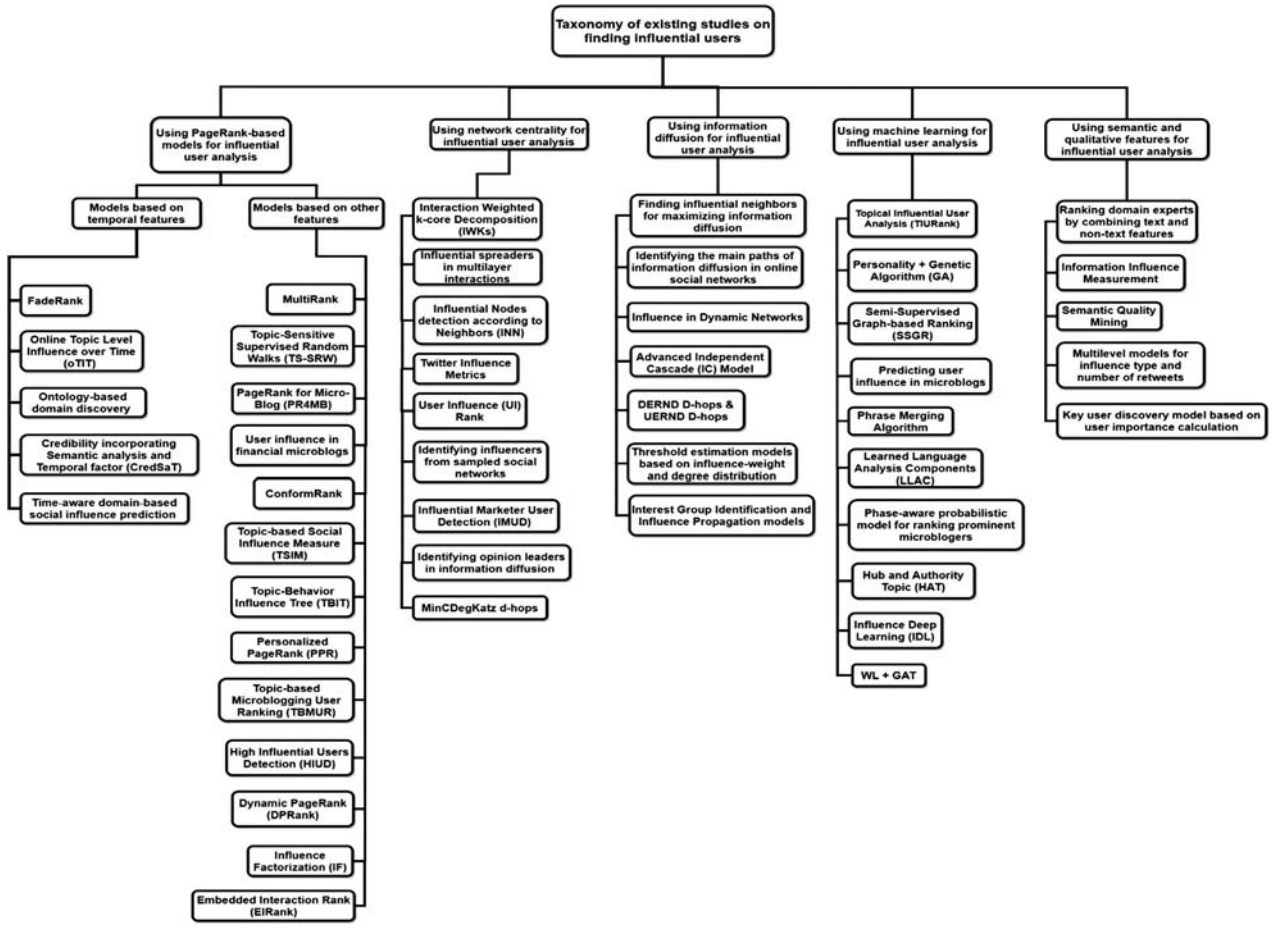

Figure 1. Taxonomy of existing literature on finding influential users.

performance optimisation approach which allows to compute the bounded digest of complete past historical activities in constant time and space. In this context, FadeRank stores only the most recent raw reputation values whereas the rest i.e. older values, is aggregated (faded) with an accuracy which decreases proportionally to their age.

2.1.1.2. Online Topic Level Influence over Time (oTIT). Similarly, research has identified that incorporating varying tendencies and trends of online influence play an important role in measuring user influence in dynamic social networks (Su et al. 2018). In this context, oTIT (Su et al. 2018) is a probabilistic generative model which combines multiple features of a social network such as text and links to identify and analyse the topic-level influence of microbloggers. On the other hand, the study utilises classical Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003) to discover the topics of interest from users' aggregated posts for identifying topic-level influence of users. Resultantly, the links are formed from the identified topic repository as words. Next, topic-oriented co-followers of a user are identified which indicates the topical influencers.

Unlike prior models, oTIT captures the time stamp when a link is generated, thereby, yielding a smooth topics distribution over links and the time. The topic distribution learning process is based on Gibbs sampling (Griffiths and Steyvers 2004), therefore, varying topical influence trends of a user can easily be drawn. Similarly, using Bernoulli distribution, the study captures the reasons of following a user online. However, to capture reasons of following a user other than topic-based interests or similarities, the topic independent links are grouped into another cluster. Consequently, the study identifies topical as well as non-topical influential users. The temporal influence of microbloggers f on a topic k until time T is given as:

$$\text{Influence}\ (f)@(k, T) = \gamma + \sum_{t=1}^{T} n_k^{(f,t)} \times e^{-(T-t)/\lambda} \quad (1)$$

$$\lambda > 0$$

where $n_k^{(f,t)}$ indicates the number of occurrences of a user f in topic k, λ represents the decay parameter for

the exponential decay and γ is the parameters of Dirichlet (Beta) priors on Multinomial (Bernoulli) distributions. Finally, the TIT is integrated with influence decay method, called oTIT (online TIT), which tracks the influential users in the data streams to effectively capture the temporal influence in a dynamic environment given as

$$\text{Influence } (f)@(k,\ T) = \gamma_{k,f}^{(s+1)} \tag{2}$$

Here, the key idea behind oTIT is to adapt the TIT model over data the stream $s \in [1,\ S]$ and use counts in the current stream to fine-tune the hyper-parameters for the successive stream i.e. $s + 1$.

The results prove that oTIT outperforms existing LDA-based approaches and identifies the current influence by considering the dynamic nature as well as dynamic scenarios of the microblogging platforms. Moreover, oTIT is effective in terms of time, space and reliability on real-world microblog networks.

2.1.1.3. Ontology-based domain discovery. One of the advantages of OSNs is the presence of millions of users and their online interactions for many hours. This opens new opportunities for marketers and advertisers to understand their loyal and potential customers through a phenomenon called 'segmentation'. Segmentation is vital for effective marketing and it requires commercial organisations to classify customers based on their interests, purchasing habits, geographic locations, financial status and the level of brand interactions. Therefore, it is necessary for such organisations to analyse a customers' social content and classify them into suitable groups that will ultimately result in delivering the right message to the right group.

Recently, Abu-Salih, Wongthongtham, and Chan (2018) presented a consolidated framework which determines the domain of textual content of user tweets and estimates user interest in 'politics'. Subsequently, the discovered domain of user interest is utilised in addressing the domain-based user trustworthiness. In the first phase, a time-aware semantic analysis is applied on users' historical content using five well-known machine learning classifiers. However, the study is limited to only on/off domain classification and groups users into politics and non-politics categories. The historical content represents the recent and former tweets which are analysed through the implemented classifiers. In the second phase, the results of the previous analysis are utilised as the primary input for predicting the domain of users' future tweets.

One of the advantages of the study includes an effective approach for social business intelligence using the notion of social trust, semantic web mining and machine learning applications. The study employs ground truth dataset and a comparison to benchmark the performance of implemented classifiers. In addition, the classification task is performed both at user and tweet levels which provides the necessary groundwork for better understanding and estimating of user interest in several domains of knowledge.

2.1.1.4. Credibility incorporating semantic analysis and temporal factor (CredSaT). Similarly, CredSaT (Abu-Salih et al. 2019) is based on online-segmentation of microblog users. CredSaT is a fine-grained framework which semantically analyses the temporal dataset stored in a distributed environment. The dataset is temporally sequenced into user data and metadata where each sequence represents a specific time period. Semantic analysis enriches the textual content of temporally sequenced data chunks using the existing Ontologies and Linked data. Subsequently, each message is linked to a particular domain for providing the semantics of the textual data. Resultantly, useful knowledge is inferred which paves the way for further analysis.

Later, credibility analysis measures trustworthiness of a user in each data chuck, thereby providing an overall credibility value in a particular domain. A time-aware, domain-based user credibility ranking approach measures initial credibility value of user tweets based on user's historical data. Resultantly, a ranked list of users is obtained with a corresponding credibility score for each particular domain.

One of the advantages of the study includes a novel metric consisting of new and current features as well as time-aware data chunks for determining domain-based trustworthy users as compared to baseline methods. Furthermore, the study provides a higher degree of tweet data analysis by analysing the taxonomies of user tweets and website content of the associated URLs, rather than analysing user's timeline as one block. Finally, taxonomy analysis of user's tweets, and content analysis of the associated URLs also help discover anomalous users such as spammers.

2.1.1.5. Time-aware domain-based social influence prediction. On the other hand, big social data (BSD) is a relatively new phenomenon which has greatly influenced the research on web mining and social network analysis (SNA). Particularly, 'social trust' has captivated the attention of computer scientists and formal organisations, as well as information consumers and information processors. Consequently, it has become

essential to implement an effective framework for temporally measuring a user's credibility in all categories of BSD. The study of (Abu-Salih et al. 2020) aims to determine top-*k* domain-specific influential users by incorporating semantic analysis and various machine learning techniques. Semantic analysis enriches the textual content of user tweets for linking each tweet with a particular domain.

Subsequently, a domain-based user credibility ranking approach is applied which consists of an advanced set of key attributes. The credibility ranking analyses the collection of a user's tweets for measuring user's credibility score in a time-sliced window and determines an overall user influence in dissimilar domains. Finally, machine learning (ML) techniques such as Naïve Bayes (NB), Logistic regression (LR), Decision trees (DTs), Deep learning (DL), Generalised linear model (GLM), Random forest (RF) and Gradient boosted decision trees (GBDTs) are implemented for predicting influential domain-based users. ML-based techniques provide an additional layer of validation.

One of the advantages of the study includes a benchmark comparative analysis for selecting an optimal technique from several ML-based implementations to evaluate domain-specific Twitter trustworthiness. The experimental results prove that GLM achieves the best performance as compared to other implemented techniques. In addition, proposed user credibility ranking framework demonstrates higher effectiveness in predicting domain-based influential users. Table 1 summarises the PageRank-based temporal models discussed above including model features, baseline methods, novelties, experimental datasets, validation techniques and important results.

2.1.2. Models based on other features

Similarly, influential users have also been identified by integrating PageRank-like algorithms with features such as user activity, user quality, user credibility, user behaviour, sentiment analysis of textual content, various types of random walks, embedded interactions between users and many other features of microblog networks. Following is a brief discussion on most relevant and recent models organised in chronological order:

2.1.2.1. Multirank. Most of the previous studies identify topical influential users by partially utilising follower-relation. However, Zhaoyun et al. (2013) proposed a multi-relational model called *MultiRank* which measures users' influence based on random walks in microblogs. MultiRank exploits four types of relationships between users: (i) retweet relationship (ii) reply relationship (iii) reintroduce or copy relationship and (iv) read relationship. The authors employed a novel approach for catering to the uncertainty in reintroduce and read operations. First, the study inferred the probability of the tweets being reintroduced by cosine similarity between the two documents and time interval distribution. Intuitively, if the time gap between the two tweets is too long, the probability of tweet reintroduce is lower.

Similarly, the study inferred the probability of reading by a user's daily tweeting rules which shows a user's behaviour. If a user's behaviour is similar to that of another, then the user has a higher probability of reading another user's tweets. Subsequently, the model couples four random walks on the multi-relation microblog network for identifying influential users. The combined random walks are performed with respect to a random surfer model where the surfers are capable of being influenced by four behaviours of their friends.

One of the novelties the study includes a combined random walk on a multi-relational network which is guaranteed to converge. In addition, the study considers both transition probabilities for inter and intra-network respectively.

2.1.2.2. Topic-sensitive supervised random walks (TS-SRW). Similarly, TS-SRW (Katsimpras, Vogiatzis, and Paliouras 2015) identifies the topical influential users using supervised random walks. However, TS-SRW was originally developed for addressing the link prediction problems in social networks (Backstrom and Leskovec 2011). Since PageRank has been widely used and proved to be an effective graph ranking technique, therefore, the authors applied PageRank for iteratively computing PageRank-like scores for each node in the graph network based on the topic-specific influence that is derived from the textual content of a node.

Expressly, the model is divided into three phases: (1) topic extraction (2) parameter learning and (3) PageRank-based random walk. A user's topics of interest are extracted using LDA (Blei, Ng, and Jordan 2003). Subsequently, the topic similarity between the users is computed. Subsequently, in parameter learning, the model learns a set of parameters of an objective function. The objective function is designed to assign a transition probability to each edge in the graph network. Finally, these parameters are used by the random walk to rank the graph nodes accordingly. Since higher topic similarity and edge weights result in higher transition probability leading to higher influential users.

One of the advantages of the study is a comprehensive set of features and effective parameter learning

Table 1. Models based on Temporal features for influential user analysis.

Year	Model name	Model Features	Baseline	Novelties	Dataset	Evaluation criteria	Results
2016	FadeRank (Cappelletti and Sastry 2012 Bartoletti, Lande, and Massa 2016;)	Tweet, Retweet & Follower Relationship	PageRank (Page et al. 1999) & TURank (Yamaguchi et al. 2010)	Time for processing data at each update is reduced to constant. Optimisation through fading memory which mitigates whitewashing attacks	**Twitter Datasets** (D1 + D2 + D3) with 11 months duration each **No. of users** – ∼11 K + ∼12 K + ∼12 K **No. of Tweets** – ∼14M + ∼12 M + ∼11 M No. of retweets – ∼5 M + ∼4M + ∼3 M **No. of Follower relationship** – 15 M + 15 M + 11 M	Kendall's correlation & Precision	FadeRank outperforms the baseline algorithms as FadeRank identifies the influential users with higher precision and in constant time. In addition, FadeRank is more resilient to whitewashing attacks as compared to its forgetful versions
2018	oTIT (Su et al. 2018)	User, Link & Time	PageRank, Link-LDA (Erosheva, Fienberg, and Lafferty 2004) & FLDA (Bi et al. 2014)	Capturing temporal influence (time)	Sina Weibo[2] Dataset 1 Users = 0.4 MWords = 207 M Links = 46 MTime-tagged links = 7 M Dataset 2 Users = 1.1 M Words=415 M, Links = 98 M, Time-tagged links = 12 M From December 1st, 2015 to January 5th, 2016	Hit Count at *k*, mean average precision at k, efficiency & expert judgement	TIT shows effectiveness and efficiency as compared to baseline methods and provides new insights regarding influence dynamics
2018	Ontology-based domain discovery (Abu-Salih, Wongthongtham, and Chan 2018)	User-count, Tweet-count, Tweet content & Temporal factor	TwitterRank (Weng et al. 2010)	Time-aware semantic analysis of historical content of a user's tweets using well-known ML techniques and predicting the domain of future tweets of a user	**Twitter** (between November 2014 and April 2015) No. of users – 33,298 No. of tweets – 1,636,320 No. of replies – 793,640	Accuracy, Precision, Recall, F-score, Logarithmic loss & t-test	The proposed framework effectively discovers the domain of Twitter content at user as well as at tweet level
2019	CredSaT (Abu-Salih et al. 2019)	User-count, Tweet-count, Retweet-count, Reply-count, Likes-count, followers-friends relation, Sentiment analysis of replies & Temporal factor	TwitterRank (Weng et al. 2010), High In-degree & High Domain-based Key Attributes	Incorporating a list of fine-grained key attributes for creating feature-based user ranking	**Twitter** No. of users – 33,298 No. of tweets – 1,636,320 No. of replies – 793,640 (Between November 2014 and April 2015)	Precision, Recall, F-score and (Normalised discounted cumulative gain) nDCG	CredSaT shows better performance over baseline methods in identifying domain-based influential users
2020	Time-aware domain-based social influence prediction (Abu-Salih et al. 2020)	User-count, Tweet-count, Retweet-count, Reply-count, Likes-count, followers-friends relation, Sentiment analysis of replies & Temporal factor	TwitterRank (Weng et al. 2010), High In-degree & High Domain-based Key Attributes	An overarching time-aware credibility framework with advanced set of sentiment and semantic attributes and a benchmark comparison for determining optimal technique from the applied ML algorithms	**Twitter** No. of users – 33,298 No. of tweets – 1,636,320 No. of replies – 793,640 (Between November 2014 and April 2015)	Classification error, Accuracy, Precision, Recall, F-score & ROC comparisons	The proposed approach is able to identify the domain-based influential users. Of all ML-based algorithms, GLM achieves best performance

approach for computing the edge weight that ultimately results in accurately identifying influential users. In addition, the study presented a detailed comparative analysis with state-of-the-art existing methodologies on more than one dataset.

2.1.2.3. PageRank for MicroBlog (PR4MB). In OSNs, users play different roles which require higher attention to user characteristics for identifying influential users with higher accuracy. However, traditional PageRank algorithm considers only the link structure of social networks and assumes the same importance for all users (Tunkelang 2009b). Mao and Zhang (2016) proposed an improved PageRank-based algorithm for mining top influentials in OSNs. PR4MB is a dynamic mining technique that employs features such as activity, quality and credibility of users. User activity reflects the blog quantities of a user and is given as:

$$\text{Activity}\ (u) = \frac{n_u}{N} \tag{3}$$

where n_u represents the number of blogs published by user u and N is the total number of users writing on this microblog. In addition, blog quality can enhance a user's effectiveness (Lü et al. 2015). Therefore, the study computes blog quality with respect to subject of observation i.e. users paying attention to social issues are evaluated at a higher level as compared to other users. The quality of a user u is computed as

$$\text{Quality}\ (u) = \frac{m_{u'}}{M_i} \tag{4}$$

where $m_{u'}$ represents number of blogs of user u on a concerned topic T and M_i indicates the total number of blogs published by the user in the social network being analysed. Similarly, credibility is an important factor which can reflect a user's influence. Microblogging platforms provide certification mechanism for verifying the credibility of influential users. User credibility can be computed as

$$\text{Credibility}\ (u) = \begin{cases} 1/O(u), & \text{when user } u \text{ is certified} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $O(u)$ represents the outbound users of the user u in the network. Next, the weightiness of a user is computed given as

$$w(u) = \text{Activity}(u) + \text{Quality}(u) + \text{Credibility}(u) \tag{6}$$

where Activity(u), Quality(u) and Credibility(u) represents the activity, quality and credibility of the user u in the network. Finally, the study improves the PageRank by weightiness setting of a user. Given a directed graph $G = \langle V, A \rangle$, its weighted google matrix is $Q = q_{u,v}$ where $q_{u,v}$ can be computed as

$$q_{u,v} = \begin{cases} w(u)/O(v) & \text{when} \langle v, u \rangle \epsilon\, A \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $O(v)$ denotes the number of outbound nodes of the user v. In fact, $q_{u,v}$ represents the probability that user v jumps to user u.

One of the advantages of the study includes a comprehensive set of features for mining influential users. In addition, the study shows higher accuracy and efficiency as compared to traditional PageRank (Page et al. 1999) which utilises only the link structure.

2.1.2.4. User influence in financial microblogs. On the other hand, studying the effects of user influence measures on financial social networks is an important research domain. Research has been proposed which considers the sentiment analysis and centrality measures for identifying the top influentials in financial microblogs (Cortez, Oliveira, and Ferreira 2016). The study employs user labelled dataset in which the text messages are classified as optimistic or pessimistic. These labelled text messages are further used in quantifying a user's sentiment. Since active interactions present a more realistic and dynamic picture, therefore, the research is based on active actions such as retweet, share or replies for creating a directed graph network between microbloggers. These directed graphs are later used for computing the four topological influence measures namely in-degree, betweenness, pagerank and number of posts resulting in several top lists of users.

Next, the study utilises a robust rolling window evaluation method (Moro, Cortez, and Rita 2014) where the entire dataset is split into several training and test datasets ordered in time. The study is based on the assumption that for each day, a user might issue several optimistic or pessimistic sentiment opinions. In addition, the study considers unseen test data that relates with the next six or four months. The sentiment opinions of all users based on their text messages are aggregated under an overall optimistic index. Finally, a particular top list of users is considered relevant if their associated optimistic index correlates with the optimistic index of other users.

Finally, the study is evaluated in a two-step process. First, the study consider Percentage of Quality Users (PQU) as classified by the StockTwits. Second, the correlation analysis between the top list and other users based on a user labelled dataset and assessed in terms of three message selections: ALL { general sentiment }and filtered by two technological giants Apple –

AAPL and Google – GOOG. The result shows that some top lists have achieved more than 80% PQU values and more than 0.6 Spearman correlation score. Similarly, best PQU results were obtained by the posts measure and best correlation results were achieved by the PageRank measure respectively.

One of the advantages of the study is that financial related messages are filtered and weighted with respect to user influence measures. The filtering process removes noise which leads to better sentiment indexes resulting in improved forecasts of stock market behaviour. In addition, the study employs a robust evaluation scheme including correlation with a StockTwits labelled dataset used as the gold standard.

2.1.2.5. Conformrank. Similarly, ConformRank (Wang et al. 2017) an improved PageRank-like model that is based on sentiment analysis. The model employs features such as tweet, tweet sentiment, retweet graph, in-neighbours, out-neighbours, leaf node and root node in addition to emotional conformity and conformity weight. Here, in-neighbours and out-neighbours indicate the follower–followee relationship in a graph-based OSN. However, the retweet network in ConformRank is an unconnected graph due to direction of the signed edges which is a big problem. Therefore, for each topic, influence is computed from different retweet networks.

On the other hand, root nodes propagate influence and act as tweet-spreaders. However, leaf nodes block retweet spread and stop influence propagation. Next, the study computes the influence of two types of network nodes: leaf nodes and non-leaf nodes. The influence of leaf nodes is simply zero. However, non-leaf nodes are further divided into non-leaf nodes without-neighbours as leaf nodes and other non-leaf nodes. The influence of the former depends on the emotional conformity of out-neighbours. Similarly, the influence of other non-leaf nodes balances influence and the emotional conformity of their out-neighbours.

One of the advantages of the study is an enhanced sentiment analysis based on creating a standard dictionary with additional inclusion of informal words and expressions. The experimental results reveal that publishing positive posts greatly enhances the social influence of a user. However, having more degrees does not guarantee large influence in the network.

2.1.2.6. Topic-based social influence measure (TSIM). As mentioned earlier that social networks are dynamic platforms for online interactions between millions of users from all over the globe. An important characteristic of dynamic social platforms is that influence of users is not identical (Katz, Lazarsfeld, and Roper 1955). The structure of social networks and shared text reveals a substantial amount of information about users, their interests and topic-based social influence. A relatively recent work (Hamzehei et al. 2017) highlighted many useful insights from the existing literature on ranking users in OSNs. First, the study emphasises the need of an authentic mechanism for discriminating the influentials from non-influential users not just retrieving the *top-k* influential users. Second, machine learning based approaches such as SVM are found to be too expensive for experimenting on very large social network data. Thirdly, being a subjective phenomenon, social influence is difficult to evaluate.

In addition, the study (Hamzehei et al. 2017) proposed a topic-based influence method (TSIM) which incorporates features such as network structure, user-generated contents, history of user's activities and the engagement of network users in a user's activities. TSIM represents users with respect to their topical interests and social influence around each observed topic. Subsequently, TSIM integrates social connections between users, topic content information and user-topic relationships into the same principled model. From the structure of the social network, the influence-related attributes are identified such as number of friends and network centrality of a user. Similarly, a user's topical influence is measured from the content of the broadcasted message. The broadcasted message may be retweeted, receive replies or contain user mentions, respectively.

One of the novelties of the study is that it considers topical interests and influence of an individual which facilitates in predicting a user's influence on a new topic. In addition, TSIM is evaluated using a unique real-world social network.

2.1.2.7. Topic-behaviour influence tree (TBIT). Similarly, TBIT (Wu et al. 2017a, 2017b) is a graph-based approach which aims to maximise user influence spread by designing a minimum path for information propagation. The proposed model examines user influence in an OSN from two different aspects. First, topics of interest are identified through social interaction and second, the identified topics are utilised for analysing user behaviour. The topical influence of a user is determined through the network relationship which is based on similarity of hashtags and massage content. Similarly, behavioural relationship between users is studied through social interaction features such as retweet, reply and mention. The behavioural relationship describes the size of influence and the propagation time between users for a topic.

Subsequently, the effects of direct influence and multiple repetitions or homophily-driven influence of the topic are studied. By excluding homophily-driven influence, the authors significantly reduce the nodes to be analysed. Subsequently, both types of user relationships i.e. topical and behavioural are merged to generate the edges of the topic-behaviour network which follows a Bayesian structure. Finally, user influence tree is generated from the topic-behaviour network using a heuristic search scheme where the influence trickles down such that the influence of a node depends on all of its parent nodes.

One of the novelties of the study includes a Bayesian network tree of textual and network features which shows increased accuracy by minimising propagation path and maximising number of influenced users. In addition, the study effectively identifies the influential users by excluding homophily-driven influence which results in a significant reduction in the nodes to be analysed.

2.1.2.8. Personalised PageRank (PPR). OSNs are effective word-of-mouth platforms for spreading knowledge, particularly marketing campaigns. Twitter is the most popular and most targeted social networking platform; therefore, marketers can greatly benefit by identifying the influential users in OSNs. The study of (Alp and Öğüdücü 2018) integrates various network features with user activities and ranks users in three different phases namely topic modeling, user modelling and influence analysis.

The topic modelling starts with the removal of unnecessary syntactical noise from the tweets using Morphology-aware model (Yildiz et al. 2017) specifically designed for the Turkish language context. Secondly, the study improves the shortcomings of LDA mechanism for shorter texts using pooling techniques (Alp and Öğüdücü 2015). In addition, the study proposed smoothed LDA which identifies topics from the filtered tweets. Thirdly, domain experts analyse and validate only the coherent topics. Finally, the tweets are classified into topics and suitable labels are assigned to the topical networks of Twitter users based on tweet content.

Similarly, in user modelling, the study computes the feature scores such as focus rate, activeness, authenticity and a novel 'speed of getting reaction' for each user. Focus rate stipulates that users who often post on fewer topics are likely to be highly influential. Similarly, activeness further examines three different perspectives of a user on Twitter. Authenticity quantifies the originality of the tweets of a user and speed of getting reaction represents the response of a user on the tweets of another user in the form of retweets. Next, these features are incorporated in PageRank for identifying top influential users.

The experimental evaluation of PPR is based on Spark framework using map-reduce for effective processing of high-cost graphics. In addition, the study proposed a novel feature which captures top influentials resulting in high levels of information diffusion. Furthermore, the study proved the effectiveness of the proposed approach over baseline methods using 'spread score' and a survey from the volunteers.

2.1.2.9. Topic-based microblogging user ranking (TBMUR). Similarly, research has identified that social trust plays an important role in determining community's sentiment polarity about a particular topic (Sherchan, Nepal, and Paris 2013). By incorporating social trust in sentiment analysis, influential users can be effectively identified in OSNs (Eliacik and Erdogan 2018). Based on social trust, TMBUR (Eliacik and Erdogan 2018) is a novel sentiment mining algorithms which employ textual information as well as the network features of microblog networks for analysing the association between influential users and sentiment polarity of topic-specific social communities. TBMUR integrates novel features such as User Trust, Influence Period and Similarity for identifying the topical influencers in different phases including the formation of topical social communities, influential user metrics computation and sentiment analysis for text polarity.

A topical social community is represented as an undirected graph of similar interest users. A friendship relation exists between any two microbloggers if they follow each other. The condition for a new user to join the topical social community is that the user must have friendship relation with no less than two distinct important users of the social community. TBMUR measures the impact of influential users as well as the trust of microbloggers in topical social communities. The TBMUR ranks are calculated using the following equation:

$$\mathrm{TBMUR}(u_i) = (1-d)\; nor\_ut_{u_i} + d \sum_{u_j \in f(u_i)} nor\_ip_{u_i u_j}\; nor\_S_{u_i u_j} \frac{\mathrm{TBMUR}(u_j)}{f(u_j)} \tag{8}$$

where $nor\_ut_{u_i}$ represents the ratio of user trust of a single user (u_i) to the sum of all the topical community users and $nor\_ip_{u_i u_j}$ and $nor\_S_{u_i u_j}$ are normalised influence period and normalised similarity score, respectively. Similarly, user trust (ut), influence period (ip)

and similarity (s) are the newly proposed metrics and d represents a constant term known as 'damping factor'. After exhaustive experimental evaluations, the researchers (Page et al. 1999) have assigned d, a fixed value i.e. 0.85. The proposed metrics are normalised such that values always converge in case of random walks. On the other hand, the incoming and outgoing links in PageRank algorithm are transformed into friendship link and are specified as f in equation (8) for *TBMUR* influential user rank calculation.

Subsequently, each of the three novel metrics is computed for user influence computation. User trust is further computed through degree of friendship, degree of expertise and degree of activity. In addition, the friendship duration is considered when computing the degree of friendship. Furthermore, the study considers the influential user to be a person who frequently posts valuable information about specific topics of interest. Resultantly, the expertise level is quantified in terms of number of shares, likes, mentions and resends received by a user's posts in a social network. Similarly, user's level of activity shows the frequency of issuing topic-relevant shares of a user which gives a clue on a user's involvement on the topic. Finally, sentiment analysis is based on Liu's classical text polarity mining approach (CSPC) (Liu 2012).

The study has several novelties such as it extends the classical sentiment polarity analysis by introducing a novel PageRank-based approach. In addition, the study considers features beyond textual contents such as mood and behaviour and presents a correlation analysis of financial social community with stock exchange. Finally, TBMUR shows increased accuracy over the baseline models by considering the level of user influence and a novel topical user influence.

2.1.2.10. High influential users detection (HIUD). Recently, it has been identified that social behaviour can also be incorporated in sentiment analysis for identifying high influential users in OSNs (Zhao et al. 2019a, 2019b). First, specific features are selected for effectively describing a user's behaviour. The selected features are adjusted using the principle of homogeneity in sociology. Next, Word2vec (Mikolov et al. 2013) has been utilised for computing the similarities between users and their followers. The values of computed similarities determine the weights of each feature which is consistent with the sociological laws.

Subsequently, a hybrid K-means clustering algorithm generates user clusters whereas number of clusters are obtained using the canopy (Mao 2012) approach. In addition, the study includes a user impact computation function which integrates multiple features such as user activity, number of fans and fan activity. Consequently, candidate clusters which meet the features of high influential user are selected. Finally, the proposed model is evaluated against the official ranking of microblogs and PageRank (Page et al. 1999).

The advantages of the study include a hybrid clustering approach using multiple features of user behaviour and network topology. In addition, the study shows better performance as compared to the baseline methods using two real-world microblog datasets. Finally, the proposed HIUD effectively visualises and characterises the geographical distribution of *top-15* influential users.

2.1.2.11. Dynamic PageRank (DPRank). Although OSNs provide a major platform for information sharing, accommodate sharing of live streaming data among millions of connected users worldwide. However, social information on OSNs is mostly analysed locally rather than considering network dynamics over time. Resultantly, users who publish low-quality posts degrade the accuracy in capturing highly influential users and topical tweets. DPRank (Shi et al. 2020) is a novel social sensing model which evaluates the dynamic topical influence of users during social information evolution. First, an automatic filtering scheme based on HITS algorithm (Shi et al. 2017) is applied for creating a high-quality influential user training data set. Next, a new post influence detection model, called dynamic network structure (DNS) (Chen and Wang 2015), comprehensively measures the real dynamic influence of user tweets. Subsequently, the study exploits user influence and influence of various topics for computing the topical influence contribution of users publishing high-quality tweets in the network. Finally, DPRank measures the topical influence contribution of each user who publishes, retweets or replies to posts under different topics.

In addition, DPRank computes the effective distance and degree of influence for ranking users of dynamic social network. Effective distance between two users represents their corresponding influences on each other and it is computed using features such as number of posts, retweets or replies. Next, degree of influence is utilised for creating a new dynamic network of users based on the quality of users and their corresponding influences on members of the community of different interests. Finally, these newly created dynamic networks are exploited for extracting a dynamic ranking of users of real importance under various topics.

One of the advantages of the study is that it effectively identifies the real post influence on different topics during event diffusion and evolution. In addition, the

influential spreaders identified with DPRank can be regarded as the most proper set when compared with the baseline which shows significant improvement and validity of the proposed approach.

2.1.2.12. Influence factorisation (IF). As discussed above in PPR that influential users can be effectively ranked in three phases by integrating various network features with user activities (Alp and Öğüdücü 2018). The extended version of PPR, called Influence Factorisation (*IF*) (Alp and Öğüdücü 2019) integrates the same network features with user activities for identifying topical influencers. However, *IF* uses latent factor prediction of user features using Matrix factorisation strategy for identifying the influential users. The factorisation strategy is further divided into User – User Influence Factorisation (UUIF) and User – Topic Influence Factorisation (U-TIF). UUIF is a matrix consisting of normalised retweet rates of network users. Similarly, U-TIF consists of computed scores of user features such as activeness, focus rate, authenticity, follower count and newly incorporated hybrid features. Next, a separate matrix is generated for each user-specific feature. Next, the study uses Alternating Least Squares (ALS) algorithm which is an efficient matrix factorisation methodology for finding the influential users (Koren, Bell, and Volinsky 2009). The study is implemented in a distributed environment using Spark MLlib for speeding up the ALS computations.

One of the novelties of the study is incorporating hybrid features in addition to network and user features which improves the efficiency in identifying the influential users. In addition, *IF* is implemented in a

parallel distributed environment which makes it highly scalable and suitable for large-scale datasets. Furthermore, *IF* shows linear time complexity as compared to baseline methods in finding influential users and *IF* can be easily extended.

2.1.2.13. Embedded interaction rank (EIRank). In addition to marketing and advertising, OSNs are widely used to recommend users, particularly influential users, to others (Riquelme and González-Cantergiani 2016). In a recent study, user's embedded interaction information (during and after the recommending process) has been utilised for identifying highly influential users (Bo et al. 2020). The proposed model, called EIRank (Bo et al. 2020) integrates embedded interaction information into a PageRank variant which leads to a better influence measure. First, different types of interaction graphs are embedded into embedded spaces. Resultantly, various types of interaction information can be effectively integrated and collectively represented. Subsequently, a novel closeness measure is defined in the embedded spaces for quantifying the closeness of users in terms of their interaction. Finally, EIRank incorporates the closeness measure for ranking users based on both their network connections and interactions.

The study shows several advantages such as the proposed closeness measure is defined based on the extent of many types of interaction between users. In addition, the study presents a comprehensive experimental analysis by using various sampling strategies and different sampling sets. The results show that EIRank computes social influence better by way of a user recommendation task using a real-world Twitter dataset and several evaluation parameters. Table 2 summarises the PageRank-based algorithms which utilise other (non-temporal) features.

2.2. Using centrality-based models for influential user analysis

Centrality-based models identify the top influential users by utilising single or combination of centrality measures such as degree, betweenness, closeness or other variants. Most relevant centrality-based models for finding top influential users are discussed in the following:

2.2.1. Interaction weighted k-core decomposition (IWKs)

Degree centrality is a widely adopted influence measures, however, it is of little relevance since a node having many less influential neighbours has much lower influence as compared to a node having few high influential neighbours (Al-Garadi, Varathan, and Ravana 2017). On the other hand, global measures such as betweenness and closeness centrality can effectively identify influential users. However, global measures are not applicable in large social networks due to high computational costs. Similarly, *k-core* based algorithms are efficient in finding the influential nodes in online social networks (OSNs) (Chen et al. 2012). However, existing *k-core* decomposition techniques consider network links equally while identifying the influential spreaders in unweighted networks.

Interaction weighted *k*-core decomposition (IWKs) (Al-Garadi, Varathan, and Ravana 2017) is a newly proposed link-weighting scheme that is based on social interaction among users. The social links between the users are weighted based on the number of interactions consisting of retweet and mention attributes using a

Table 2. Models based on other features for influential user analysis.

Year	Model name	Model features	Baseline	Novelties	Dataset	Evaluation criteria	Results
2013	MultiRank (Zhaoyun et al. 2013)	retweet, reply, reintroduce (copy) & read	TwitterRank (Weng et al. 2010)	Combining multi-relations of the features such as retweet, reply, reintroduce and read for measuring the influence of a user	**Twitter** Users > 260 K, Tweets > 2.7M (Chinese language Twitterers)	Sensitivity, f-measure & expert judgement	MultiRank outperforms the baseline TwitterRank which is proven on a multi-relational influence network
2015	TS-SRW (Katsimpras, Vogiatzis, and Paliouras 2015)	Tweets, followers, friends, favourites & lists	PageRank (Page et al. 1999), TwitterRank (Weng et al. 2010), Velocity and No. of followers	Proposed model and machine learning-based evaluation technique are the novelties	**Twitter** (i) politics, (ii) World Cup 2014 (iii) Scala and (iv) Snow 2014 Data Challenge test	Kendall Tau Correlation	TS-SRW is correlated with baseline methods and produces better user rankings in different scenarios
2016	PR4MB (Mao and Zhang 2016)	User Activity, User Quality, User Credibility, User Weightiness & Weighted Google Matrix	PageRank (Page et al. 1999)	Proposed features and dynamic mining algorithm	**Twitter** Last 5 years data are crawled using Twitter API and social network is generated for selected 3049 users	Execution/Running Time	PR4MB shows higher accuracy and objectivity as compared to baseline
2016	User Influence in Financial Microblogs (Cortez, Oliveira, and Ferreira 2016)	In-degree, betweenness, number of edges and quality of edges based on retweet network and number of tweets	PageRank (Page et al. 1999), betweenness (White and Borgatti 1994) and in-degree	Applying user influence measures in financial social media is the novelty	**Stock Twits**[3] Users > 19 K Tweets = 2 M From June 2010 to March 2013	Percentage of Quality Users (PQU) & Spearman's rank correlation	With respect to PQU, the posts measure performs best. PageRank achieves the best results according to correlation analysis
2017	ConformRank (Wang et al. 2017)	Tweet, Tweet Sentiment, Retweet, Follower, Followee, Conforming weight & Emotional conformity	RetweetRank (Page et al. 1999), icnRank (Chen et al. 2011) & Degrees are the baselines	Emotional weights and Emotional Conformity	**Sina Weibo** Users > 219 K, Tweets > 2.7 M, Unidirectional relations > 3.2 M	Precision, recall, F1-measure & Kendall Tau Correlation	ConformRank shows better performance than baseline. Positive posts can significantly enhance a user''s social influence
2017	TSIM (Hamzehei et al. 2017)	Follower scale, Topic activity, Topic-based Attractiveness & PageRank centrality	PageRank (Page et al. 1999)	Proposed features & unique dataset	**Twitter** Users > 301 K, Tweets > 101 K	Expert opinion & user citations on google scholar	Topic-based influential users are effectively identified as compared to classical PageRank
2017	TBIT (Wu et al. 2017a, 2017b)	Message content, Hashtags, Retweets Replies & Mentions	PageRank (Page et al. 1999), ProfileRank (Silva et al. 2013), TwitterRank (Weng et al. 2010) and MIIB (Khan, Daud, and Malik 2015)	Constructing topic-behaviour heterogeneous networks and generating Influence Trees based on topic-behaviour networks	**Twitter** Tweets > 1M, Follower/ followee relationships > 376 K (From May 28 to June 7, 2016) & TUAW[4] dataset	Kendall coefficient, stability of users' influence, length of influence tree & OSim (Haveliwala 2003)	TBIT demonstrates increased accuracy and efficiency as compared to the baseline methods
2018	PPR (Alp and Öğüdücü 2018)	Focus rate, Activeness, Authenticity & Speed of getting reaction	PageRank (Page et al. 1999) and TwitterRank (Weng et al. 2010)	Newly proposed feature 'Speed of getting reaction' and evaluation using 'Retweet rate'	**Twitter** Users = 186 K, Tweets = 38 M, Links = 16 M, (From 4, 2015 to January 12, 2016)	Spread score and human-based survey	PPR effectively identifies the influential users than baseline and Spread score is an effective evaluation metric
2018	TBMUR (Eliacik and Erdogan 2018)	User Trust, Influence Period, Similarity & Sentiment Analysis	PageRank (Page et al. 1999), InterRank (Sung, Moon, and Lee 2013) and XiangRank (Li et al. 2013)	Proposed features are the novelty	**Twitter** User timelines = 529, Preprocessed tweets > 1.2 M from April 14, 2014 to March 09, 2015	Accuracy, precision, recall, f-measure & correlation analysis	TBMUR performs an effective sentiment analysis on topical social communities than baseline
2019	HIUD (Zhao et al.)	Followers, Attentions, Reposts, Comments & Thumbs-up	PageRank (Page et al. 1999) & FBR (number of fans and blogs)	Proposed features & hybrid clustering approach	**Sina Weibo** Users – 63,641, Microblogs – 84,168. From May 03, 2014 to May 11, 2014 **Tencent Weibo** User – 21,051, Microblogs – 96,542, Following-followee relationships – 54,325.	The Pearson correlation analysis	HIUD shows better performance in identifying high influential users than baseline methods

				From April 01, 2012 to April 09, 2012			
2019	Dynamic PageRank (DPRank) (Shi et al. 2020)	User-count, Tweet-count & Retweet-count	HITS+TSLDA (Shi et al. 2018) and PageRank+HEE (Shi et al. 2017)	An improved PageRank algorithm that extracts the dynamic topic-based influence of social network users during event diffusion and evolution	**Twitter** (Twitter 2019) No. of users – 36,052 No. of posts – 1,500,000	Correlation coefficient analysis (Inman and Bradley 1989), Accuracy, Precision & Recall	DPRank shows higher efficiency and effectiveness in discovering highly influential users and influential posts as compared to baseline methods
2019	Influence Factorisation (IF) (Alp and Öğüdücü 2019)	Activeness, Focus rate, authenticity, follower count and hybrid features	PageRank (Page et al. 1999) [5], TwitterRank (Weng et al. 2010) and PPR (Alp and Öğüdücü 2018)	Predicting future influencers as well as current ones & Implementation of ALS algorithm on Spark MLlib	**Twitter** Users = 186 K, Tweets = 38 M, Links = 16 M From 4, 2015 to January 12, 2016	Node activation & Spread score	IF efficiently identifies the current as well as future influentials as compared to the baseline methods
2020	ElRank (Bo et al. 2020)	Favourite, Mention, Reply, Retweet, & following relationship	PageRank, Node2vec (Grover and Leskovec 2016) & TwitterRank (Weng et al. 2010)	A social network influence ranking method based on the embedded interaction networks	**Twitter** (between May and August 2019) No. of users – 14,709 No. of tweets – 9,232,343	Percentage of friendship relationships	ElRank outperforms the baseline methods across a range of experimental settings in measuring online influence for user recommendation

single network (i.e. the social network connection) between the same users. The weighting scheme is based on the underlying observation that interaction among the users plays a significant role in quantifying influence spread in OSNs. A strong interaction among the users results in a high probability of influence spread in the network and vice versa (Backstrom et al. 2011). The study improves the original *k-core* ranking by defining the weight of the edges (in a directed graph) given as

$$W_{i,j} = t_{(i,j)} \tag{9}$$

Here, t represents interaction weight between the nodes in the network. It is computed using two interaction factors which play a significant role in OSNs given as

$$t = p_{(i,j)} + e_{(i,j)} \tag{10}$$

where $p_{(i)}$ represents a node's strength to spread online content in the network. It also measures the ability of a node to generate online content using the features like retweet or share in OSNs given as

$$p_{(i)} = \sum_{j}^{N} n_{ij} \tag{11}$$

where n_{ij} denotes the total number of the content propagated between the nodes i and j. Similarly, $e_{(i)}$ represents the engaging strength of node i and it quantifies a user's engagement in conversations and is computed as

$$e_{(i)} = \sum_{j}^{N} m_{ij} \tag{12}$$

where m_{ij} represents a total number of engagements between the nodes i and j. Afterwards, each node is assigned a weighted degree according to the following rule:

$$k_i^w = \lambda \boldsymbol{k_i} + (1 - \lambda) \sum_{j \in R} W_{ij} \tag{13}$$

where R denotes the set of neighbours of a node i and the value of λ in this study is set to 0.5. Lambda (λ) is a tuneable parameter and it computes the link-interaction weights and the degree equally. After the preparation phase, *k-core* decomposition applies the pruning routine of the original method.

One of the advantages of the proposed (IWKs) model is that it considers the degree as well as interaction weights between the nodes. The model assigns network nodes to a large number of layers as compared to the original *k-core* decomposition. Moreover, the model

distinguishes between the influential spreaders better than the baseline methods.

2.2.2. Influential spreaders in multilayer interactions

Existing studies validate the proposed models through simulating the diffusion dynamics rather than following the real information flow in microblog networks (Pei et al. 2015). Results obtained, even from the well-performing algorithms, depend on the structure of a particular algorithm which causes inconsistency. For instance, PageRank and degree centrality fail in ranking the influence of users on a wide range of social networks having real spreading dynamics. Research identifies that top influentials can be consistently discovered in the *k-core* across various dissimilar social platforms (Pei et al. 2015). Al-Garadi et al. (2016) present a topological representation of an OSN by considering the multilayer interactions and the overlapping links as weights. The influential spreaders are identified from multilayer networks using the most prominent algorithms such as PageRank (Page et al. 1999), degree centrality (Freeman 1977) and *k*-core (Dorogovtsev, Goltsev, and Mendes 2006) respectively. Each layer in multilayer interactions represents a different type of relationship (i.e. following, retweet, mention) between the nodes that are aggregated into a single network.

Next, the study investigates the effects of different multilayer topological representations on the accuracy of the algorithms. First, in the absence of complete global network structure, sum of the nearest neighbour's degree is highly reliable local proxy for a user's influence. Secondly, the accuracy of identifying the influential spreaders depends not only on improving the ranking algorithms but also on developing a network topology that represents the diffusion of information in an efficient manner. Thirdly, the ranking algorithms show better performance on the proposed network topological representation as compared to single-layer representations. Finally, the study identifies that there is not a single influential spreader identification algorithm that always performs efficiently under any topological network representation.

The study includes several novelties such as a rich multilayer aggregated network which plays a significant role in understanding the information diffusion in a real OSN and selecting an efficient algorithm for identifying the influential spreaders. The study compares the ranking algorithms using diverse evaluation parameters such as recognition rate and imprecision functions. In addition, the findings in the study provide practical guidelines for designing optimised strategies regarding viral information diffusion.

2.2.3. Influential nodes detection according to Neighbours

Considering the multilayer interactions among network users can effectively identify the top influentials. Influential nodes detection according to Neighbours (INN) (Sheikhahmadi, Nematbakhsh, and Zareie 2017) is a novel approach based on the interactive relationships among the nodes in the network. The model ranks the nodes based on the role of each node's neighbours and the strength of relationship among the nodes. The role indicates how rapidly neighbours respond to the published messages of a node in the network. Subsequently, the network is partitioned into several social communities and the amount of interaction among the users is used as the weight of relations within the resulting subgraph for each community. In addition, for each node in a community, details about a node's activities and its neighbour's reactions are recoded in a local table for detecting the influence-weight between them.

One of the challenging features of a social network is that it is highly dynamic. Therefore, after a specific time some user reactions do not worth. INN considers the time between the creation of a post and the user reactions received on it. This time interval is termed as 'delay' and it is used for classifying a node's neighbours in the network. To do so, the model considers the standard deviation to the average of reaction delay which helps in categorising the identified influenced individuals as 'early influenced', 'mediocre influenced' and 'late influenced', respectively. Resultantly, influential users are identified with higher accuracy despite eliminating less important nodes. Similarly, some nodes whose neighbours effectively contribute to the spreading process are selected as the initial seed. The proposed model adopts a semi-local method which considers level one and level two neighbours for selecting the seed set according to the number of community nodes in different communities. Finally, spreading influence of the nodes is computed for each community with respect to level one and level two neighbours.

One of the advantages of the proposed model is an efficient graph partitioning strategy for fast extraction of the community structure in large social networks. INN shows higher accuracy despite excluding the less important nodes and significantly reduced time complexity despite considering level one and level two neighbours during the seed selection process. In addition, INN significantly reduces the time cost of weight determining for communities by eliminating the nodes which have passive neighbours i.e. neighbours which are just recipients and do not play a role in spreading the published messages.

2.2.4. Twitter influence metrics

Network functionality in OSNs is an important aspect for influence ranking. However, functional properties are network-dependent since each social network offers different interaction possibilities. Drakopoulos et al. (2017) proposed an analytical framework which extends the performance of an influence metric using probabilistic tools from multiple fields such as information theory, psychometrics and data mining. The proposed framework consists of influence metrics which can be classified into two different categories i.e. first order and higher order. First-order metrics utilises only the account-specific information of users. On the other hand, higher-order metrics derive the social influence of a user's account as a function of the influence of follower accounts. Resultantly, the digital influence of a user is effectively captured. The computation of higher-order metrics is based on Katz centrality (Katz 1953) and TunkRank (Daniel Tunkelang 2015). The former exploits the network structure whereas the latter employs the functional aspects.

One of the novelties of the study is that it integrates various structural and functional properties including a newly proposed 'atomic influential' metric which results in capturing a major part of Twitter activities. In addition, a linear probabilistic evaluation proves the effectiveness of the proposed models over the baseline methods.

2.2.5. User influence (UI) rank

In microblogs networks, user influence depends on the ability to generate information as well as to persuade others to disseminate that information. For maximising the influence of a user, it is important to be in a central position in information dissemination network besides creating valuable tweets. In addition, tweet topic differences play a significant role in determining social influence. User influence (UI) Rank (Jianqiang, Xiaolin, and Feng 2017) is a newly proposed model that utilises the concept of random walks for measuring user's tweet contribution and characteristics of information propagation. User's tweet contribution is measured using *retweets* and *comments*. Similarly, a user's position in the follower–followee network represents his/her relative importance. Users can spread information quickly depending upon their network centrality and the influence of their followers.

Next, UIRank employs diverse centrality measures such as degree, betweeness and closeness centrality for identifying true influential users. Users with higher degree centrality have a higher probability for their tweets to be read by many users and spread within the network. With greater betweenness centrality, users can spread information quicker through fewer fans and followers. On the other hand, greater closeness centrality indicates a user's strong control over information dissemination. Using the follower graph, the model iteratively computes the influence score for each user.

One of the advantages of the study is a comprehensive comparative analysis with baseline methods. UIRank outperforms the baseline using evaluation measures such as recall, precision and F1-score. The experimental results indicate that retweet-count alone reveals little about a user's influence level. Users with the highest rate of retweet-count not necessarily be an influencer. In addition, as social influence highly depends on and varies with respect to topics, therefore, UIRank considers tweet topic differences while identifying the influential users.

2.2.6. Identifying influencers from sampled social networks

In SNA, node sampling is a promising research domain for identifying influential users (Bakshy et al. 2011). One of the reasons include enormous efforts and time to obtain the complete structure due to false data, missing data or node sampling from the social networks. In the existing literature, several methodologies have been proposed for estimating the structures of social networks (Gjoka et al. 2010; Ai et al. 2013; Ahmed et al. 2014a, 2014b, Wu et al., 2016). However, the study of Tsugawa and Kimura (Tsugawa and Kimura 2018) examines the effects of several node sampling strategies to OSNs on the effectiveness of popular influence measures. The experimental evaluation is conducted using three biased sampling strategies and random sampling. The findings of the study reveal that influential users can be identified from biased sampling strategies such as sample edge count (Maiya and Berger-Wolf 2011), breadth-first search and depth-first search (DFS) with an accuracy slightly smaller as compared to the accuracy obtained from that of the complete social networks. This may be because in social networks, influential users tend to have higher degrees (Weng et al. 2010), therefore, the negative effects of biased sampling strategies are expected to be lower as compared to the effects of random sampling. In addition, sampled social networks eliminate nodes which act as noise. Resultantly, in some case, influential users can be identified with higher accuracy as compared to using the entire network.

The advantages of the study include examining the effects of various node sampling techniques on the effectiveness of popular influence measures as compared to existing studies which (Kim and Jeong 2007; Pei et al. 2015; Salamanos, Voudigari, and Yannakoudakis 2017) investigate the effects of node sampling only on

the stability of influence measures. In addition, the study demonstrates the possible advantages and necessary sample sizes for identifying the influential users in OSNs.

2.2.7. Influential Marketer User Detection

OSNs are widely used by various types of commercial organisations for maximising the spread of influence. Maximisation of influence can be significantly enhanced through the identification of top-*k* influential users. In most of the existing studies on influence maximisation, topological locations of the nodes in the network have been considered as a measure for determining their influentiality (Zareie, Sheikhahmadi, and Jalili 2019a). However, Zareie, Sheikhahmadi, and Jalili (2019a) posit that users' interests should be considered while identifying the influential users for viral marketing. Similarly, the seed set should be selected while keeping in view of their interest in a particular piece of information, otherwise, seed nodes will not be willing to spread information or send recommendations to their friends (Chevalier and Mayzlin 2006).

Subsequently, users' influence is specified in the proposed 'Influential Marketer User Detection (IMUD)' on the basis of interests of their friends in the marketing message. Next, users' interest vector in different time stamps is extracted and compared with the feature vector of the message. Finally, using a novel algorithm, user influence is measured based on interest scores of their friends and top-*k* most influential users are selected.

The novelties of the study include users' network position, contextual information and interest in spreading message. In addition, seed set are selected such that both seed nodes and their neighbours are willing to spread information to their friends resulting in maximised influence spread.

2.2.8. Identifying opinion leaders in information diffusion

Similarly, the study of Rehman et al. (2020) identifies the opinion leaders in online discussion forums for efficient marketing campaigns. Opinion leaders play an important role in spreading information and keeping the isolates interested in online discussion network. Opinion leaders are identified using in-degree and out-degree centralities and ranked within the network based on their betweenness centrality scores.

Subsequently, the study examines the flow of information and connection patterns among the identified opinion leaders. The experiments are conducted using a publicly available Higgs Twitter dataset (Leskovec and Krevl 2011) as Twitter has quick flow of information which results in huge impact on opinion formation of the general public (Rehman et al. 2020).

The findings of the study enlist five key users along with their interaction patterns. Although, conversation starter is an opinion leader, however, conversation starters are generally unable to control the opinion of online users participating in the discussion. On the other hand, the remaining four types of opinion leaders play a far better role in rapid information distribution within the network. For instance, influencers have higher credibility among the masses and influencers are mentioned or retweeted by network builders. On the other hand, information bridge regularly seeks information and acts as the source of information provider for other users in the network.

One of the advantages of the study includes an efficient graph visualisation for the complete reply network and a comprehensive community evolution analysis. In addition, the study identifies various types of opinion leaders and examines their connection patterns for rapid information dissemination.

2.2.9. MaxCDegKatz d-hops 'and' MinCDegKatz d-hops

On the other hand, influential users play a pivotal role in increasing the adoption of the promoted information and behaviour within a social network. Therefore, Alshahrani et al. (2020) proposed two efficient algorithms for maximising influence through identifying top-k influential users. The algorithms integrate local and global centrality measures on a graph-based network with preselected weight edges which should surpass a predefined threshold value. In addition, each seed set is separated by a certain number of hops which differ depending on the graph radius. Resultantly, the algorithms prevent the selection of the influential users which have influenced the same nodes in the network. The study includes extensive experiments on large-scale graph datasets for both directed and undirected networks under Independent Cascade (IC) and Linear Threshold (LT) models (Shakarian et al. 2015).

The findings of the study indicate that weighted integration of centrality-based heuristic algorithms show better performance in terms of run-time complexity and influence coverage as compared to various existing methods. In addition, the results of the study show consistency as the experiments are performed on several large-scale graph datasets. The centrality-based models discussed above for influential user analysis are summarised in Table 3.

Table 3. Centrality-based models for influential user analysis.

Year	Model name	Model features	Baseline	Novelty	Dataset	Evaluation criteria	Results
2016	IWK_S (Al-Garadi, Varathan, and Ravana 2017)	Tweets, Retweets & Mentions	Degree centrality (Freeman 1977), PageRank (Page et al. 1999) and K-shell (Batagelj and Zaversnik 2003)	Improved K-shell algorithm with link-weighting technique & computing recognition rate of imprecision functions	**Twitter** Dataset – 1 (De Domenico et al. 2013) Tweets > 985 K, Users > 456 K, Directed edges > 14 M (From 1st July 2012 to 7th July 2012) Dataset – 2 (Weng, Menczer, and Ahn 2013) Tweets > 121 M, Unique users > 14 M	SIR model & Kendall's Tau correlation	The proposed model outperforms degree and betweenness centrality and performs as good as closeness centrality with significantly reduced computational cost.
2016	Influential spreaders in multilayer interactions of OSNs (Al-Garadi et al. 2016)	Following, retweet & mention	Degree centrality (Freeman 1977), PageRank (Page et al. 1999) and k-core (Dorogovtsev, Goltsev, and Mendes 2006)	Comparative analysis of ranking algorithms on a multilayer aggregated network	**Twitter** Dataset 1 (Higgs Twitter) Nodes > 456 K Edges > 14 M From 1st to 7th July 2012 Dataset 2 (Weng, Menczer, and Ahn 2013) Nodes > 14 M, Edges > 595 K	Imprecision function & recognition rate	The proposed model produces reliable results than baseline and discovers that accuracy in finding influential users depends on developing an appropriate network topology
2017	INN (Influential Nodes detection according to Neighbours) (Sheikhahmadi, Nematbakhsh, and Zareie 2017)	User Activity (Post, Likes, Comments, Sharing, action log) & Timestamp of user activity	Cluster rank (Chen et al. 2013), Coreness + centrality (Bae and Kim 2014), IMSN (Sheikhahmadi and Nematbakhsh 2017), H-index (Mcdonald 2005) & Semi-local degree centrality (Ren and Linyuan 2014)	Nodes Classification based on level two neighbours spreading capability	**Higg twitter**[5] Nodes > 456 K Edges > 15 M From 1st to 7th July 2012	SIR model & spreading velocity	INN show higher accuracy and correctness in detecting the spreading influence of nodes as compared to baseline methods
2017	Twitter Influence Metrics (Drakopoulos et al. 2017)	First-order metrics & Higher order metrics	Katz Centrality (Katz 1953) & TunkRank (Tunkelang 2009a)	Newly proposed feature 'Atomic influential metric' & higher order metric computation from first-order metrics	**Twitter** From November to December 2016 (accounts with educational Hashtags only)	Normalised correlation coefficient, Kullback–Leibler divergence, Tversky index & Pearson skewness coefficient	The proposed Twitter influence metric outperforms the existing digital influence rankings and captures a major part of Twitter activities
2017	(UI Rank) User Influence Rank (Jianqiang, Xiaolin, and Feng 2017)	User Tweet contribution (Tweet) & Information spread (Retweet and Comments)	TunkRank (Huang and Xiong 2013), the FansRank (Kwak et al. 2010), RetweetInfluence (Cha et al. 2010) and out-degree centrality	Proposed features and comparison with the baseline algorithms	**Sina Weibo** Users > 516 K, Tweets=217M, Follower/followee relations > 25M From September 1, 2015 to November 1, 2015	Precision, Recall and F1-measure	The proposed model outperforms the baseline using classical machine learning based evaluation metrics
2018	Identifying influencers from sampled social networks (Tsugawa and Kimura 2018)	Tweet, Retweet, Mention & follower relationship	Sample edge count (SEC) (Maiya and Berger-Wolf 2011), breadth-first search (BFS) & depth-first search (DFS)	Examining the effects of node sampling from social networks and collaboration networks on the effectiveness of popular influence measures at identifying influencers	**Twitter-follow** No. of nodes – 50 K No. of links ~ 33 K, No. of information cascades ~ 21 K Twitter-mention No. of nodes – 390 K, No. of links – 539 K, No. of information cascades – 100 K	Overlap & Normalised influence	Using biased sampling strategies, the results obtained from a sampled social network is comparable to the results obtained from the complete social network. However, in collaboration networks, network sampling

(*Continued*)

Table 3. Continued.

Year	Model name	Model features	Baseline	Novelty	Dataset	Evaluation criteria	Results
							does not yield promising results.
2019	Influential Marketer User Detection (IMUD) (Zareie, Sheikhahmadi, and Jalili 2019a)	Tweet & Comment	Degree centrality, Degree discount centrality (Chen, Wang, and Yang 2009), Degree punishment (Wang et al. 2016), Initial multi-spreader nodes selection (Sheikhahmadi and Nematbakhsh 2017), Graph coloring distance (Guo et al. 2016) & Multi-criteria influence maximisation (Zareie, Sheikhahmadi, and Khamforoosh 2018)	A method for measuring user influentiality based on user interest. User interest is computed using user's network position as well as contextual information. An algorithm for selecting top-k most influential users	**Twitter** No. of nodes – ∼465 K No. of edges – ∼834 K 1st order degree of nodes – 3.59 2nd order degree of nodes – 816.84	Normalised average interest value, Average degree, Accuracy, Complexity analysis	IMUD shows higher effectiveness in terms of complexity, accuracy and normalised average interest
2020	Identifying opinion leaders in information diffusion (Rehman et al. 2020)	Retweet, Mention & Reply	Betweenness centrality (White and Borgatti 1994)	Efficient graph visualisation for the complete reply network, comprehensive community evolution analysis & identifying various types of opinion leaders and examining their connection patterns	**Higgs Twitter** Retweet Network No. of nodes – 256,491 No. of edges – 328,132 Mention Network No. of nodes – 116,408 No. of edges – 150,818 Reply Network No. of nodes – 38,918 No. of edges – 32,523 From 1st to 7th July 2012	Out-degree, In-degree, Network Modularity & Connection pattern	As compared to conversation starter, other opinion leaders play a far active role in information dissemination. However, a dataset consisting of a longer duration is better suitable for higher reliability.
2020	MinCDegKatz d-hops (Alshahrani et al. 2020)	Retweet, Mention & Reply	Degree centrality, PageRank (Page et al. 1999), Degree Discount Heuristic (Chen, Wang, and Yang 2009), BCT Algorithm (Nguyen, Dinh, and Thai 2016a), TIM + Algorithm (Tang, Xiao, and Shi 2014), IMM algorithm (Tang, Shi, and Xiao 2015), SSA (Nguyen, Thai, and Dinh 2016b) & DSSA (Nguyen, Thai, and Dinh 2016b)	Improved time complexity with good influence spread capabilities. Integrating centralisation metrices and separating seed set selection such that influential users with influence the same nodes are excluded.	**Munmun_twitter_social (Twitter)**[6] No. of nodes – ∼465 K No. of edges – >834 K Type of graph – Directed Diameter – 8	Mean, Standard deviation and Variance under Independent cascade (IC) model and linear threshold (LT) model (Shakarian et al. 2015) & Running time complexity	The proposed algorithms show better performance as compared to several existing techniques under IC & LT diffusion models. In addition, the proposed algorithms show low run-time complexity.

2.3. Using information diffusion-based models for influential user analysis

Information diffusion-based analysis utilise independent cascade (IC) model, linear threshold (LT) model or a combination of both for identifying the influential users. Following is a discussion on relevant diffusion models for identifying influential users:

2.3.1. Finding influential neighbours for maximising information diffusion

Traditional influence maximisation approach ignores the network topology; therefore, it is not as preferable for viral marketing as information diffusion-based methods Kim, Beznosov, and Yoneki (2015a). The study of (Kim, Beznosov, and Yoneki 2015a) proposed a decentralised version of influence maximisation by influencing k neighbours as compared to arbitrary users in the network. The study introduced novel features including user propagation weight, content interestingness and decay factor for providing a more general and pragmatic diffusion model. User propagation weight represents average propagation rate of users to their neighbours. Similarly, content interestingness of information reflects the number of users who freely share information with their neighbours. Finally, decay factor is associated with the freshness of information which might drop exponentially with the passage of time.

In addition, the study investigates four neighbour selection schemes using intensive simulation and presents a guideline for choosing a selection method under specific conditions. Overall, hybrid selection scheme outperforms other selection schemes. Hybrid scheme maximises information diffusion using the characteristics of a user's posts without requiring information about a node's degree. Resultantly, neighbours with high diffusion rates are selected as compared to users having a high number of neighbours.

One of the advantages of the study includes an intensive experimental evaluation using real and simulation-based synthetic propagation rates. The results show that synthetic propagation rates are not as suitable for understanding the information diffusion as real-world propagation rates. In addition, the study compares the performance of simple as well as sophisticated neighbour selection schemes for maximising influence spread.

2.3.2. Identifying the main paths of information diffusion in OSNs

On the other hand, growing research on relationship strength in OSNs shows that there is higher probability of socially active links and the existence of main paths which can play a significant role in information dissemination. The relevant literature consists of studies which are mainly focused on identifying the spread pathways of definite events. However, Zhu et al. (2016) proposed a quantitative method for analysing and evaluating network paths from a global perspective. The study consists of two main steps: finding main paths of information diffusion and finding influential users. Main paths are identified based on intrinsic characteristics and the objective laws of a post's spreading process. Furthermore, weight of the network link or tie strength is measured using historical forwarding frequency between each pair of network nodes. Tie strength tends to determine which follower in the network is more likely to repost the message after a recent message post. Therefore, tie strength has a significant impact on the flow direction of messages in the network. Finally, a PageRank variant (Li et al. 2014) is applied for identifying the influential nodes.

The study includes several advantages such as identifying the influential users and main spreading paths as main paths which connect the influential users. Identification of main spreading paths significantly improves information diffusion. In addition, the proposed approach avoids man-made subjectivity to some extent as the weights of network links are measured using real interaction information. Similarly, Dijkstra's algorithm (Dijkstra 1959) optimisation reduces the time complexity while computing the significance of spreading paths between two influential nodes. Finally, the positive effect of the weak links is not ignored as the links with smaller weights can become part of the main paths while maintaining the path importance.

2.3.3. Influence in dynamic networks

One of the challenging task in SNA is differentiating influence maximisation from finding influential users as both domains have entirely different implications (Yang et al. 2017b). For instance, influence maximisation is one of the core techniques in viral marketing (Domingos and Richardson 2001). However, influence maximisation is not as effective as finding influential users in cold-start scenarios where appropriate users are recommended to newcomers. Yang et al. (2017b) proposed an effective polling-based method for tracking the influential users in a dynamic social network. The network dynamics are modelled as a stream of edge-weight updates. Next, an efficient incremental approach updates random reversely reachable (RR) sets against the network changes. In polling, a node is randomly picked, and its live-edge influencers are selected through the process of Monte Carlo simulation. After M polls,

identified set of influencers is known as random RR set, where M is a random varibale. The M number of RR sets form a hyper-graph H with users as a set of nodes as V and RR sets as hyper-edges.

In dynamic networks, it is highly important to closely monitor the influence of online users for identifying the true influencers. Therefore, the random RR sets are maintained over streams of edge-weight updates to output unbiased influence spread under both linear threshold (LT) model and independent cascade (IC) model (Kempe, Kleinberg, and Tardos 2003). In addition, an inverted index is maintained in LT to access any random RR set passing through a node. The graph of the entire retweet network is stored and maintained such that it allows access to each node along with its neighbours. Similarly, the nodes and live edges among the nodes for random RR sets in IC are also recorded. The study classifies the live-edges into breadth-first search (BFS) edges and cross edges. An edge is classified as BFS edge if a live-edge, say (v_i, v_j), makes the node v_i propagate for the first time while generating a RR set by reversely propagating from the starting point following breadth-first search order; otherwise, the edge is classified as cross edge.

Furthermore, all live-edges pointing to a node in an RR set are stored in an adjacency list. For rapid node retrieval, the nodes in an RR set are treated as strings and are kept in a prefix tree manner. The main difference in data structure for storing RR sets is that in LT, the propagation probabilities are stored on live edges whereas in IC, propagation probabilities are stored in graph data structure which improves space complexity. The proposed model takes streams of edge-updates as input and top-k users with influence score crossing a certain threshold are identified as the influential twitterers.

One of the distinguishing features of the study is an efficient and effective algorithm which identifies the influential users and maximises the influence spread. Results prove that the cost of incremental updates and maintenance is significantly reduced as compared to re-generating M number of RR sets. In addition, the proposed algorithm determines an appropriate sample size for influence maximisation and finding influential users resulting in strong quality guarantees.

2.3.4. Advanced independent cascade (IC) model

In microblog networks, a user's high authority can significantly increase the probability of information dissemination. Yang et al. (2017a) proposed a novel advance independent cascade (IC) model which examines the information diffusion in microblogs and identifies the topical influencers. The study identifies that there are some important and familiar attributes based on user activity for spreading information i.e. the attention received from other users. User activity shows the positivity for sharing information in a social network given as:

$$A_i = \frac{\sum_{j=1}^{m} f\Delta t_j}{m} \tag{14}$$

where A_i denotes the activity of a user i, Δt_j indicates the period given as $\Delta t_j = (j = 1,\ 2,\ \ 3,\ \ \ldots..m)$ and $f\Delta t_j$ represents a number of shared information in time Δt_j. Similarly, the study quantifies the spread attributes U_i given as:

$$U_i = \alpha A_i + \beta\gamma_1 + \lambda(\gamma_2 + \gamma_3) \tag{15}$$

Here, γ_1, γ_2 and γ_3 represents the average attribute values of forwarding, comment and the love for message posts, respectively. Similarly, α, β and γ represents the weights of the attributes. The study utilises analytic hierarchy process (AHP), which is a widely used complex decision-marking technique for computing the weights.

Next, LDA (Blei, Ng, and Jordan 2003) processes user information for topic relevance R_i given as:

$$R_i = \frac{1}{n}\sum_{j \in M_i} \overline{m_{jc}} \tag{16}$$

where m_{jc} is word segmentation of a message j which stores the word frequency m_j and word segmentation. After matching high-frequency words w_n with m_{jc}, topic relevance $\overline{m_{jc}}$ is given in Equation (16). In short, R_i represents the relevance of a user on certain topic domain. Subsequently, the propagation model, that is an improvement in the classical independent cascade (IC) model, is set up with directed edges of different weights given as

$$M = \eta_{ij} = \begin{cases} w(v_i,\ v_j) & (v_i,\ v_j)\ \epsilon\ E \\ 0 & \text{others} \end{cases} \tag{17}$$

Here, M is the adjacency matrix, v_i and v_j denote the nodes with directed edge and η_{ij} represents the probability of information spread from v_i to v_j. The process of information propagation starts with some random seed node as input to the source. The identification of these seed nodes is crucial for information propagation. Resultantly, the nodes are ranked according to relevant importance given as

$$P_i = \lg(\deg_i^{\text{out}}) \times R_i \times U_i \tag{18}$$

where $\deg_i^{\text{out}}$ represents number of attentions paying neighbours of a user i and P_i is the original but not the true importance of a user i due to the influence of its neighbours. As the influence from a node to its

neighbours is not equal due to different probabilities of information spread, therefore, the study exploits the arithmetic of PageRank (Sarma et al. 2013) for ranking seed nodes. The importance of a node is computed using the following equation:

$$PR(v_i) = \frac{1-d}{n} + d\sum_{j=1}^{m} \frac{w(v_i, v_j) \times PR(v_j)}{\deg_i^{\text{out}}} \quad (19)$$

Here, n is the size of the graph with a directed edge from v_i to v_j and d refers to the damping factor. It has a fixed value assigned after experimental evaluation which equals 0.85 (Page et al. 1999). On the other hand, $PR(v_i)$ indicates the importance of the node v_i and it converges after several iterations. Similarly, $PR(v_j)$ represents the original importance of the node v_j. These node ranks are later used for identifying the seed nodes which play an important role in information propagation. Propagation starts with the activation process where if a node is activated at step t, then the activated node will activate its neighbours independently with the probability at step $t+1$ in a cyclic manner. The activation probability is similar to weights of directed edges in a graph $G(V, E)$. The proposed model makes the propagation process more aligned and analogous to real information propagation by setting up a threshold. A user will be activated successfully only if the activation probability exceeds the threshold.

One of the advantages of the study is an effective analysis of topic relevance, spread attributes and user authority in a microblog network. In addition, the propagation process is improved by introducing a threshold and setting up a simulation process to investigate the time of information spread and the number of affected nodes.

2.3.5. DERND D-hops and UERND D-hops

Similarly, Alshahrani et al. (2018b) proposed two novel algorithms for maximising the influence spread in both directed and undirected graph networks under the IC and LT models. The algorithms are based on radius-neighbourhood degree approach for the selection of top-K influential users. The study improves the seed set selection of previous versions (Alshahrani et al. 2018a) by designing a selection threshold for each type of graph network. Resultantly, the choice of low influence nodes i.e. the nodes with smaller power to sway other nodes in adopting or promoting a specific behaviour, is prevented. The selection threshold is based on the structural properties which vary with the type of the graph network. Subsequently, the study fixes the neighbourhood hops at radius minus 1 that helps in selecting the nodes which influence a different range of users.

The novelties of the study include two new algorithms for directed and undirected graph networks with significantly reduced time complexity and enhanced influence coverage by fixing the limitation of consecutive see set choice. In addition, the study improves the efficiency of the seed selection process in both types of graph networks by using a predefined multi-hops distance which allows the selection in a comparatively large region for selecting the most suitable nodes as the seed set.

2.3.6. Threshold estimation models based on influence-weight and degree distribution

On the other hand, most of the influence maximisation techniques employ either IC or LT models for node activation in mining influential users. One of the major issues with IC includes a single chance for active neighbours to active another network node with a particular probability. Similarly, in LT model, a node activation requires the aggravated influence of all activated neighbours to cross a certain threshold value. In this context, the study of (Talukder et al. 2019) presents a comprehensive survey of various threshold values as threshold plays a significant role in maximising node influence. The authors argue that existing threshold values are computed arbitrarily which requires formal techniques to optimise these values. In addition, the survey includes four novel threshold estimation models using influence-weight and degree distribution.

Later, the proposed models are evaluated against four state-of-the-art solutions using real-world social networks. The findings of the survey indicate that the estimated threshold values lie within the range of mostly used values. Therefore, the estimated values of the survey can be used in any LT-based influence maximisation algorithm. In addition, the survey shows that threshold values are application dependent and vary with applications of influence maximisation, influence-weight and the degrees nodes in the network. Moreover, the threshold values are classified into three main categories namely unique threshold, random individual threshold and miscellaneous threshold.

The study includes several novelties such as the models provide a more specific and narrow range of threshold values as compared to broad ranges offered by many existing models. In addition, the proposed estimation algorithms are faster, scalable and independent of the influence-weight estimation models. Therefore, the algorithms are applicable to both single-attribute and multiple-attribute influence-weight estimation models. Although the estimation algorithms are LT-based, however, they can be applied with standard IC model and Epidemic diffusion model (Table 4).

Table 4. Information diffusion-based models for influential user analysis.

Year	Model name	Model features	Baseline	Novelty	Dataset	Evaluation criteria	Results
2015	Finding influential neighbours to maximise information diffusion (Kim, Beznosov, and Yoneki 2015b)	User propagation weight, content interestingness & decay factor	Influential Neighbour Selection (INS) (Kim and Yoneki 2012)	Proposed features and experimental evaluation using real as well as synthetic propagation rates	**Twitter** Users > 220 K, Tweets = 1.1 M, Retweets = 290 K, Mentions > 16 K, Hashtags = 515 K, Distinct URLs = 25 K, Follower/followee relationships = 79 M From May 05, 2010 to May 12, 2010 UK elections	Ratio of activated nodes	Hybrid scheme for neighbour selection outperforms other baseline schemes for spreading influence.
2016	Identifying the main paths of information diffusion in OSNs (Zhu et al. 2016)	Post, Follower & Repost	Tie strength & PageRank (Page et al. 1999)	Identifying main spreading paths which connects highly influential users. Computing the importance of main paths using tie strength and Dijkstra's optimisation (Dijkstra 1959) for reducing time complexity and higher information diffusion.	**Sina Weibo** No. of nodes – 1488 No. of links – 16,324 Average Degree – 10.97 Network Diameter – 13 Average path length – 4.249 Average clustering coefficient – 0.114	In-degree, Number of first-order followers, Number of second-order followers, Active forwarding paths & Forwarding amount	Main-path extraction results in community-like effect where the influential users in the same group tend to interact with each other frequently. Main-path extraction tends to provide a clear understanding of intrinsic relations of OSNs and help adopt suitable strategies for promoting or restraining information diffusion
2017	Influence in Dynamic Networks (Yang et al. 2017b)	Tweets & Retweets	Linear Threshold (LT) Model & Independent Cascade (IC) Model (Kempe, Kleinberg, and Tardos 2003)	A polling-based method for tracking influential nodes & an efficient incremental approach for updating random reverse reachable (RR) sets against network changes	**Twitter** Nodes > 41 M, Edges > 1.4B, Average degree = 35.3 including datasets from Wiki-Vote, Flixster, soc-Pokec & flickr-growth	Recall, Maximum Error & Running Time	The proposed model outperforms the baseline in reducing the cost of incremental updates and maintenance as well as guarantees strong quality in finding influential users
2017	Advance IC Model (Yang et al. 2017a)	User activity, forwarding, love (approval to the message) & comments	betweenness centrality & IC Model (Ai et al. 2013; Pei et al. 2015)	Improved version of IC model	**SinaWeibo** Nodes > 38 K, Connections (directed edges) > 57 K From November 2015 to January 2016	Expert judgement & Number of infected nodes	The proposed advance IC model outperforms the baseline in terms of influence propagation.
2018	DERND D-hops & UERND D-hops (Sun et al. 2018 Alshahrani et al. 2018b;)	No. of users & Follower relationship	Degree centrality, PageRank (Page et al. 1999), Degree Discount Heuristic (Chen, Wang, and Yang 2009), BCT Algorithm (Nguyen, Dinh, and Thai 2016a), TIM + Algorithm (Tang, Xiao, and Shi 2014), IMM algorithm (Tang, Shi, and Xiao 2015)	Newly proposed DERND D-hops & UERND D-hops which improve the efficiency of seed selection process and resolve the limitation of consecutive see set choice for better influence coverage	**Munmun-twitter-social**[7] No. of nodes – 465,017 No. of edges – 834,797 Graph type – Directed Network Diameter – 8	Influence coverage & Time complexity	The proposed algorithms outperform the baseline methods in terms of influence coverage and time complexity
2019	Threshold estimation models based on influence-weight and degree distribution (Talukder et al. 2019)	Follower–followee relationship	OIM algorithm (Lei et al. 2015), T-SKIM model (Cohen et al. 2014), SIMPATH (Goyal, Lu, and Lakshmanan 2011)& LDAG models (Chen, Yuan, and Zhang 2010)	Two newly proposed algorithms which provide a more specific and narrow range of threshold values. In addition, the algorithms are applicable to both single-attribute and multiple-attribute influence-weight estimation models	**Ego-Twitter** (Leskovec and Mcauley 2012) No. of nodes – 81,306 No. of edges – 1,768,149	Threshold Values, Running Time & Threshold Fitting Distribution	The proposed estimation algorithms are faster, scalable and independent of the influence-weight estimation models.

2020	Interest Group Identification and Influence Propagation models (Abd Al-Azim et al. 2020)	No. of users & Follower relationship	The Improved K-shell approach (Liu et al. 2015), Diversity-Strength Ranking (DSR) and Extended DSR (Zareie, Sheikhahmadi, and Jalili 2019b), Degree/Out-degree & Community Scale-Sensitive Maxdegree model (Hao et al. 2012)	The proposed algorithms with high interactivity and ranks distinction as well as ultimate observer nodes improve the quality of ranking nodes	**Ego-Twitter** (Leskovec and Mcauley 2012) No. of nodes – 81,306 No. of edges – 1,768,149	Mean Absolute Deviation, Silhouette Coefficient metric, Ranking Success Factor & Monotonicity relation	The proposed IGI and IP models outperform the baseline methods in identifying the interest groups and ranking nodes in terms of their influence propagation.

2.3.7. Interest group identification (IGI) and influence propagation (IP) models

Like influential nodes, interest groups also have practical applications in viral marketing, monitoring the opinion of people, social psychology analysis and the discovery of communities in OSNs. Interest groups represent varying clusters of social network users with dynamic structures emanating from the interest of users in the propagated contents. The study of (Abd Al-Azim et al. 2020) captures the interest groups by clustering the social network nodes with respect to specific disseminated content and rank node influence in each interest group based on their role in disseminating that content. In addition, the study introduces 'ultimate observers' which are identified for adjusting the ranks of influential nodes within the interest groups. Moreover, a comprehensive evaluation is presented through extensive experiments on several benchmark datasets which shows that the proposed *IGI* and *IP* models outperform the baseline methods in identifying the interest groups and ranking nodes in terms of their influence propagation.

One of the novelties of the study is that the proposed algorithms facilitate the node ranking by ensuring high interactivity between members of the interest groups and ranks distinction for each node. In addition, the ultimate observer nodes significantly improve the quality of node ranks. Finally, the study examines the effectiveness of the ranked nodes on content dissemination.

2.4. Using machine learning-based models for influential user analysis

Learning-based models use the machine learning based classification or regression techniques for identifying the top influential users. Following is a discussion on the most relevant machine learning-based models for identifying influential users:

2.4.1. Topical influential user analysis (TIURank)

Topic-specific influence analysis is an important technique in microblogs. Existing studies neglect the relationship strength and interaction frequency between users of a social network. Liu et al. (2014) proposed a latent variable model (TIURank) based on Poisson regression for estimating the relationship strength between pair of users. The study uses a heterogeneous graph of users, their retweet relationships and interaction frequency. The key assumption of the work is derived from the theory of homophily (Mcpherson, Smith-Lovin, and Cook 2001) which says that users tend to form ties with other users having similar interests. Interaction frequency is utilised for estimating

relationship strength between Twitter users. Next, each user is ranked according to a specific topic using the retweet network. The experimental results on five topical datasets show that TIURank shows higher precision and relevance on identifying topic-specific influential users in Twitter.

One of the advantages of the study is incorporating a comprehensive set of features including retweet relationship and interaction frequency. The study presents more realistic relationship strength as compared to using binary interaction in existing studies. In addition, the study evaluates the performance of *TIURank* using diverse evaluation metrics.

2.4.2. Personality + genetic algorithm (GA)

Personality characteristics are one of the effective and relatively new aspects of information diffusion in social networks. Golkar Amnieh and Kaedi (2015) proposed an optimisation model using personality characteristics and network structure for identifying the influential users which ultimately results in maximising influence spread. According to the big-five theory (Mccrae and John 1992), personality is divided into five different traits.

However, this study considers only extroversion and openness as these two personality traits greatly affect a person's decision-making in considering a new product or a new message. Extroversion is defined as sociability or positive effect and is computed as

$$\mathrm{we}(i) = \frac{\mathrm{degree}(i)}{\text{egocentric network density } (i)} \tag{20}$$

where the degree represents the number of connections between the node i and other nodes and density indicates the ratio between the number of the existing connections to the maximum possible connections in a graph. Density can be categorised into sociocentric and egocentric density where sociocentric density is focused on relationship patterns and egocentric density relies on the connections of a specific node in a social network. On the other hand, openness is defined as the feature which distinguishes a realistic person from a highly imaginative person and is negatively associated with egocentric density. One possible explanation of this phenomenon is that people with high level of openness tend to be friends with users who are surrounded by low-density network. The study computes the openness as

$$\mathrm{wo}(i) = 1.0 - \frac{\text{egocentric network density } (i)}{\max(\text{egocentric network density})} \tag{21}$$

Next, the study applies a real coded genetic algorithm as it is one of the most efficient algorithms in optimisation problems. Resultantly, acceptance function $f_i(t)$ is required which shows the extent to which a node i accepts the product at time interval t while considering the acceptance situation of the neighbouring nodes. In bass model (Goldenberg, Libai, and Muller 2001; Rand and Rust 2011), acceptance function depends on two parameters: external influence and social influence. In this study, acceptance function is computed as

$$f_i(t) = p + q\left(\frac{n_a(t)}{n}\right) \tag{22}$$

where p and q are external and social influence on acceptance, respectively. Similarly, n represents the total number of neighbours of i and $n_a(t)$ indicates the number of neighbours who are influenced and accepted the product at time interval t. The value of the coefficient p and q is directly dependent on type of the product. Therefore, products having a large external influence (i.e. p) are accepted naturally by customers. Subsequently, the study applies the net present value (NPV) (Goldenberg et al. 2007) which measures both the value of acceptance and percentage of acceptance of a product given as

$$\mathrm{NPV}(G,\ S, f(t)) = \sum_{t=0}^{\infty} a(t)\rho\lambda^t \tag{23}$$

Here, $a(t)$ represents a number of acceptors at time interval t, ρ is the profit and λ denotes the rate of discount at the time of acceptance of the product. Furthermore, the study uses the characteristics of social network graphs such as degree (Hakimi 1962), average path length (Fronczak, Fronczak, and Hołyst 2004), clustering coefficient (Watts and Strogatz 1998), personal preferences and two-steps. Two-steps represents the number of accessible nodes by passing two edges from the node i. Finally, the nodes in the graph obtain their influence scores using the linear combination of the above characteristics and the top nodes with higher scores are selected as the most influential for seeding. Therefore,

$$\begin{aligned} w_{\mathrm{comb}}(i) =& \alpha_d wd(i) + \alpha_a wa(i) + \alpha_c wc(i) \\ &+ \alpha_r wr(i) + \alpha_t wt(i) + \alpha_o wo(i) + \alpha_e we(i) \end{aligned} \tag{24}$$

where $wd(i)$, $wa(i)$, $wc(i)$, $wr(i)$, $wt(i)$, $wo(i)$ and $we(i)$ represent the characteristics of degree, average path length, clustering coefficients, personal preferences, two-steps, openness and extroversion, respectively. Similarly, α represents the weight with respect to each characteristic. These weights are initially unknown;

however, optimal values are determined using genetic algorithm with real coded values.

The study includes several advantages such as a rich set of characteristics from graph structure as well as from big-five theory of personality traits. In addition, character weights are adjusted using a real coded genetic algorithm for determining the optimal impact in finding influential users for maximising information diffusion.

2.4.3. Semi-supervised graph-based ranking (SSGR)

Recently, research has been proposed which aims to identify experts using Twitter lists and relations among the users (Weng et al. 2010; Pal and Counts 2011; Ghosh et al. 2012). However, these studies only partially employ such relations. However, semi-supervised probabilistic (SSGR) model (Wei et al. 2016) is a novel algorithm which jointly employs three types of relations for identifying topic-specific experts. The probability of a user being an expert on a given topic has been estimated using local relevance and global authority. Local relevance measures the similarity between the published tweets of a user against a given query. On the other hand, global authority shows the global expertise of a user on a specific topic. SSGR is based on a normalised Laplacian regularisation term by using relations such as follower-relation, user-list relation and list-list relation for offline computing the topical authority of a user. A Twitter list is created by topical experts consisting of their followers. Therefore, the metadata such as the title of a list represents the crowdsourced topical annotations of the followers in that list (Ghosh et al. 2012). Subsequently, the models employ a Gaussian-based approach for computing local relevance online between the users and the given arbitrary topical queries. Finally, the topical experts are ranked based on the computed scores of global authority and local relevance.

The study includes various novelties such as jointly exploiting different types of relations in follower graphs and Twitter lists. In addition, the authors introduce a novel 'wisdom of Twitter crowds' measure that lies within Twitter lists and it acts as the supervised information for extracting topical experts. Consequently, the model computes a loss term that guarantees that the global authority of a user will lie within the wisdom of Twitter crowds.

2.4.4. Predicting user influence in microblogs

Research reveals that influential users can be identified with higher precision by integrating user and content features into regression models (Zhou et al. 2016). User features include several properties relating to the author of the original post such as number of followers, number of friends, number of favourites, etc. Similarly, content features include various statistics about the original post such as number of hashtags, URLs, words in microblog, etc. The content features are further divided into productivity and popularity. The study consistently identifies the individual level influence by aggregating all posts of every user in the network. Next, the study fitted three popular regression models which take the user and content features as predictors and output top influential users. Finally, the study identified that individual influence of most microbloggers changes over time that is in line with the findings of Akritidis, Katsaros, and Bozanis (2009).

One of the novelties of the study is that it emphasises the ex-ante prediction of the influentials over ex-post explanation as compared to existing studies which mostly identify influential users only in retrospect. In addition, the study considers the temporal aspects of user influence and identifies that user influence is dependent on their own properties and published messages.

2.4.5. Phrase merging algorithm

Similarly, Zhou et al. (2017) proposed an effective algorithm that combines multiples statistical features with topic activities for finding topical influential users in microblogs. Topic features represent probability distribution over several topics of interest for each user. On the other hand, the study considers 19 statistical features, of which, 13 are related to user properties and 6 are associated with the post contents. The mean of all messages of a user over a period of time is taken as user's statistical features.

In social networks, mostly people tend to reply or forward messages they deem tempting. Therefore, first, the study computes the influence score of messages using reposts and replies and later, the influence score of users using h-index [14]. Subsequently, the topics of interest for each user are identified using a novel phrase merging algorithm which transforms the segments of a document into high-quality 'bag of words' for topic extraction. The adjacent words are merged based on the frequency of co-occurrence. The significance score of the frequency and co-occurrence of two adjacent words is given as

$$\mathrm{sig}(p_i,\ p_j) = \frac{g(p_i,\ p_j)}{\min\left(g(p_i),\ g(p_j)\right) - g(p_i \& p_j) + 1} \quad (25)$$

Here, g refers to actual number of words in all documents and p_i and p_j represent the successive co-occurrence of the two words, respectively. The merging of the words depends on the significance score. High

significance score indicates that the two different words are strongly associated and should, therefore, be merged. At each iteration, the adjacent words having an occurrence rate greater than the threshold are merged into a phrase. Each of the newly merged phrase is taken as a single unit for the next iteration. The algorithm ends either when there is nothing to merge or the significance score does not meet the threshold. Next, LDA (Blei, Ng, and Jordan 2003) takes these newly merged phrases as input and results in an $m \times n$ matrix, where m represents the number of users and n indicates the number of topics. Each row in the matrix represents the probability distribution over all identified n topics of interest for each user. Finally, the influence score of each user is predicted with higher precision, particularly after adding the topic information, using three popular regression models. The experiments are performed in a distributed environment using Spark framework for processing huge data. The results show that adding topic information improves all three regression models, however, Gradient Boosting performs better than others.

The novelties of the study include an improved LDA which effectively extracts topic features. In addition, the study employs diverse user and topic features for predicting the influence score of users with higher precision.

2.4.6. Learned language analysis components (LLAC)

Interestingly, social science can be employed for deriving rich attributes of users and message contents which improves the identification of influential users (Rosenthal and Mckeown 2017). Inspired by the work of Schultz et al. (2007), the proposed model, called learned language analysis components (LLAC), utilises a supervised influential user finding technique which detects situational influence across five online genres based on a sophisticated learned analysis of various dialects, discourse and author characteristics. These learned components capture the characteristics of online conversions in microblogs by examining the language and participation of potential influencers.

Subsequently, first set of features include claim and argumentation which are collectively known as persuasion. The process to persuade someone starts with making a claim and then following closely by argumentation. Argumentation represents the justification to a claim. Experiments in social science reveal that argumentation is a sign of being influential (Schultz et al. 2007).

The proposed model automatically labels the sentences of the discussion as claim or argumentation. The claim system is a supervised system which uses features such as committed belief (Prabhakaran, Rambow, and Diab 2010), opinion (Rosenthal, Mckeown, and Agarwal 2014), Parts-of-Speech (POS) and n-grams to test whether a sentence is a claim.

On the other hand, the argumentation system utilise Rhetorical Structure Theory (RST) relations in RST Penn Treebank (Carlson, Marcu, and Okurowski 2003) for extracting the list of indicators of relations and co-occurring pairs of content words for each of the indicators of relations. In addition, the study tracks whether the first sentence of a post is a claim or argumentation. Starting a post with a claim or an argument indicates that it is stronger.

2.4.7. Phase-aware probabilistic model for ranking prominent microbloggers

Real-time information retrieval from microblogs plays a significant role during crisis events such as disasters. However, the volume and variety of shared information during such events over-complicates the situation. Unlike existing studies, Bizid et al. (2018) proposed a user-centric information retrieval (IR) algorithm for tracking and identifying prominent microblog users. These prominent users are susceptible to spread relevant information at the early stages of the crisis event. Consequently, emergency teams can have access to real-time valuable information.

Generally, the characteristics and importance of crisis events indicate a dynamic behaviour with respect to each event phase. The term 'phase' represents the evolving states of a crisis event over time. Each phase effects users' behaviour differently depending on their interests and the involvement in that phase. Similarly, the temporal distribution of users' activities over event phases reveals their real behaviour. The study predicts and ranks prominent users by learning from prior crisis events. In addition, the study considers the features which can be efficiently computed in real-time and learning a priori the identification algorithms which are adapted to each category of the crisis events.

The novelties of the study include a new user representation which considers both user and event-specific factors over time. The proposed user representation covers the following new aspects (1) modelling the behaviour dynamics of a microblog user, (2) depicting a microblog user's activity through a temporal sequence representation, (3) time-series-based selection of the highly discriminative features and (4) a priori learning, probabilistic, phase-aware models for predicting the prominent users over time. The results reveal that the proposed model significantly outperforms the existing phase-unaware models. In addition, the model

learns and predicts most of the prominent users at an early stage of each phase of the event being analysed.

2.4.8. Hub and authority topic (HAT)

On the other hand, discovering hidden topical hubs and authorities in OSNs is a novel generative technique. Using hubs and authorities, Lee, Hoang, and Lim (2018) proposed an effective model which directly and explicitly models the probabilistic relationships among hubs, authorities and topical interests. The study defines the proposed HAT model by first, developing a generative story for both the links and content in a social network. The generative process is based on Dirichlet-multinomial distribution which generates posts, topic-specific hubs and/or authorities and following links for each user. The topic distribution of each user is a k-dimensional multinomial distribution over k different topics, where k is a given parameter. A user''s posts are generated by sampling words of the posts from the topic distribution.

The study models the relationship between a user''s topical interests, hubs and authorities using exponential regression approach. Similarly, the link is generated from Bernoulli distribution and following link from a node u to a node v is determined by u''s hubs and v''s authorities. The post generation and topic distribution for each microblogger is performed independently. Therefore, multiple child processes can be used in parallel for simultaneously determining a user's posts and topic distribution. Finally, the model learning is based on Gibbs-Expectation-Maximisation (Bilmes 1998).

Next, the study utilises a data sub-sampling method which retains only the strong signals of a user''s hubs and authorities while eliminating the remaining noise. Resultantly, the complexity of the model is greatly reduced. HAT algorithm is evaluated in two sets of experiments: (i) topic modelling and (ii) link recommendation. In topic modelling, the topics learned by HAT are directly compared with that of baseline methods. On the other hand, in link recommendation, HAT predicts the missing links in Twitter and Instagram datasets.

One of the advantages of identifying hubs and authorities is enhanced user recommendation and effective marketing campaigns. In addition, HAT learns the hub, authority and topic-specific user interests simultaneously and outperforms other methods in recommending topical influential users by user link prediction.

2.4.9. Influence deep learning (IDL)

Similarly, predicting effective influential spreaders in OSNs plays an important role in a variety of applications such as viral marketing and online recommendation. However, conventional machine learning algorithms for influence prediction are mainly based on various hand-crafted characteristics and are difficult to generalise for different domains. To address these issues, Wang et al. (2019a) proposed a novel data-driven deep learning algorithm for learning the latent vector representation to predict influential spreaders. In addition, the authors designed a strategy for incorporating user features and network structure into graph convolutional neural network which overcomes the imbalance problem of the labelled training data. Finally, the evaluation process involves different datasets and the output of the proposed deep learning algorithm is compared with the ground truth for minimising the loss function.

The study includes novelties such as improved efficiency of the learning process through mini-batch learning model. During each iteration of the training process, a sub-network is randomly sampled to be a mini-batch. Next, the neural network learning model in sampled mini-batch is optimised for faster execution. In addition, IDL shows higher practical viability on large-scale datasets and it employs more information propagation data which improves the output of influence prediction.

2.4.10. WL (Weisfeiler-Lehman) + GAT (graph attention networks)

Engagement marketing is a relatively new phenomenon that is utilised by many businesses for maximising social actions to promote brand awareness. In OSNs, high sparsity in-feed ads pose a serious challenge which can be effectively addressed through the phenomenon of social influence prediction. However, current influence prediction approaches consider limited neighbour information due to sparsity (Wang et al. 2020). For instance, DeepInf (Qiu et al. 2018) is a state-of-the-art method that considers limited features and action status of a user's neighbours. To address this issue, Wang et al. (2020) proposed an effective approach which provides an end-to-end mechanism to leverage social influence for boosting social action prediction. The study focuses on in-feed ads and extracts a subgraph of each user with the near neighbour interactions. Later, a structure-aware graph encoding scheme is developed for learning the topological features of the subgraph. The underlying logic behind incorporating the topical features is that topological features reflect whether the user occupies an important place in the centre or remains in the peripheral of the neighbourhood which tends to influence that user.

Furthermore, the study analyses the process of how a user is influenced by its neighbours. In this context, the model utilises graph attention networks by combining features and action status of neighbours for learning the influence dynamics. Finally, the learned social influence is integrated with ad-exposure features and state-of-the-art classifiers are leveraged for predicting social actions.

The advantages of the study include extensive experiments on real-world datasets from commercial as well as public platforms. The experimental results prove the effectiveness of topological features for social action prediction even though the social network is highly sparse. In addition, the employed social influence learning significantly improves the social action prediction as compared to the baseline methods. The machine learning-based models for influential user analysis are summarised in Table 5.

2.5. Feature-based influential user analysis

Feature-based models utilise the feature of microbloggers and/or the features of tweet contents for identifying the influential users. The most relevant feature-based models for finding influential users are discussed in the following:

2.5.1. Ranking domain experts by combining text and non-text features

Existing feature-based approaches exploit authentication information (or profile descriptions) and post content analysis for ranking users in microblogs. However, users may publish post contents which is not related to their profile descriptions. Similarly, relying solely on post contents may cause biasness as spammers tend to relate posts and topics. Qi et al. (2015) proposed a novel method which combines text and non-text features for identifying experts from the domain of information technology in a microblog network. The study employs forward and backward greedy approaches for optimal feature combination and SVM algorithm for ranking domain experts. Results indicate that backward greedy approach outputs optimal feature combination. Subsequently, domain experts are ranked based on features in optimal feature combination which results in higher accuracy.

The advantages of the study include a comparison analysis between the three ranking algorithms for optimal feature selection and finding topical experts. The ranking algorithms are compared with and without optimal feature selection for a thorough analysis of the effects of features on the algorithm's performance. The results reveal that optimal feature combination improves the recognition rate of topical experts.

2.5.2. Information influence measurement

User quality is a relatively new aspect of influence analysis in microblogs. Research has been proposed which combines information attributes and user quality for effectively measuring message influence as compared to existing studies (Yu et al. 2016). In addition, the influence of spam users can be reduced using a punishment coefficient. Information attributes such as retweet and comment represent different influence effects. Retweet users play an active role in information diffusion whereas comment users mostly act as participators in promoting information diffusion. Therefore, the study considers different weights for retweets and comments and dynamically measures the message influence.

Subsequently, the study considers user quality of those users who participate in message diffusion process according to temporal order. In microblogs, a high influence message effects some high-quality users. A high-quality user has many followers and few followees. Next, users having user quality below a certain threshold are punished using a punishment function. Finally, message influence is measured using a number of retweets, comments and user quality. In addition, the model dynamically computes the influence decay in message diffusion process.

One of the advantages of the study includes a novel and dynamic information influence measurement process based on information attribute and user quality. In addition, a novel evaluation approach proves that the proposed study reduces spam user's influence for effective measurement of message influence.

2.5.3. Semantic quality mining

Similarly, semantic quality mining is an important quality-based approach (Mahalakshmi, Koquilamballe, and Sendhilkumar 2017). Unlike existing studies, semantic mining computes the influence score using a combination of network centrality and textual quality of the tweets. The process starts from validating and parsing tweets from the dataset and extracting relevant features for computing influence score of each user. The extracted features are later used for generating a connection graph which shows follower–followee relationships. Next, the connection graph is analysed for detecting communities using Louvain method. Louvain is a digraph library algorithm which exploits greedy approach for optimising the modularity of a network partition. The algorithm generates a higher-level network of nodes by identifying each small community in the connection graph. The algorithm continues to

Table 5. Machine learning-based models for influential user analysis.

Year	Model name	Model features	Baseline	Novelty	Dataset	Evaluation criteria	Results
2014	TIURank (Liu et al. 2014)	Follow, Retweet, List, Mention, Time, URL & Hashtag	TSPR (Haveliwala 2003) & RSTSPR (Xiang, Neville, and Rogati 2010)	Incorporating relationship strength, regression-based estimation of relationship strength & a novel TIURank framework	**Twitter** 5 topical lists[8] containing list members and followers with 200 recent tweets for each user	Precision, Relevance & User survey	TIURank shows higher precision and relevance on finding topical influentials as compared to baseline methods
2015	Personality + GA (Golkar Amnieh and Kaedi 2015)	Openness, Extroversion, Degree of a node, Two-steps, Average Path Length (APL) & Clustering Coefficient (CC)	Degree of nodes, two-steps of nodes & clustering coefficient of nodes (Even-Dar and Shapira 2007; Stonedahl, Rand, and Wilensky 2010; Stonedahl 2011)	Newly proposed features 'openness and extroversion' and using real coded genetic algorithm for optimal adjustment of weights	**Twitter** (Agent based implementation) Nodes (agents) = 1 K, Connections > 13 K	Net present value (NPV)	The proposed model shows 37% improvement in efficiency over baseline methods
2016	SSGR (Wei et al. 2016)	follower relation, user-list relation & list-list relation	TwitterRank (Weng et al. 2010) & Cognos (Ghosh et al. 2012)	Semi-Supervised Graph-based Ranking approach by jointly using different types of relations in Twitter	**Twitter** Users > 77 K, Lists – 5.5 M (From 4 April 2013 to 10 June 2013)	Precision at *top-k* & normalised discounted cumulative gain (Baeza-Yates and Ribeiro-Neto 1999)	Topic-specific users are ranked with higher accuracy as compared to baseline
2016	Predicting user influence in microblogs (Zhou et al. 2016)	Followers, friends, favourites, statuses, bi-followers, page-friends, URL, verification, length of name, length of description, days (date of joining), gender and follower to friend ratio	H-index (Mcdonald 2005)	Comparing three regression models	**Sina Weibo** Users > 244 K, Messages > 20 M From 01 July 2013 to 30 June 2015	*R*-squared	The study shows higher precision as compared to baseline methods
2017	Phrase Merging Algorithm (Zhou et al. 2017)	Statistical Features, 13 statistical features about the user & 06 statistical features about the message content Topic Feature Messages of each user	Random Forest, Decision Tree & Gradient Boosting	A novel phrase merging algorithm for topic extraction. In addition, a new method of finding topical influentials with higher precision	**Sina Weibo** No. of users – 1,262,518, No. of seed messages – 114,286,565 (between 1 April 2013 and 31 March 2015)	Root mean square error (RMSE) & Coefficient of determination (R^2)	The proposed algorithm identifies topical influential users with higher precision than baseline
2017	LLAC (Rosenthal and Mckeown 2017)	Opinion, Claims, Argumentation, Persuasion, Agreement, Author Traits, Dialogue Patterns and Credibility	Predicting everyone as the influential user, number of words a person writes & extension of similar work (Biran et al. 2012) of the same authors	Newly proposed features 'Author Traits' and 'Credibility' and development of rich dataset resource with detailed cross-genre analysis	**Twitter** Threads = 99, Authors >1 K, Posts >1 K, Influencers = 101 including datasets from Wikipedia, LiveJournal, Political Forum & Create Debate	F – score & cross-validation	The proposed supervised influencer detection system shows significant improvements than baseline
2018	Phase-aware probabilistic model	Machine learning-based feature extraction and selection	Pal (Pal and Counts 2011), Pal* (Pal and Counts 2011) & same phase-aware models without Boolean feature	New approach for temporal representation of user behaviour and machine learning-based features extraction and selection	**Twitter** Tweets >44 K, Shared tweets >3 K From 29 to 30 September 2014	Recall, Precision K & ROC curve	The proposed model outperforms the baseline methods in detecting prominent users at early stages of each phase
2018	HAT (Lee, Hoang, and Lim 2018)	Users, Posts, Links, Followers & Followees	LDA, Twitter-LDA, HITS & WTFW	Integrating topical interests with hub and authority users	**Twitter** Users > 9 K, Links > 316 K, Posts > 1130 K Instagram	Precision at *top-k* and Mean Reciprocal Rank (MRR) (Dang, Kelly, and Lin 2007)	HAT outperforms the state-of-the-art baseline models in recommending topical influential

(*Continued*)

Table 5. Continued.

Year	Model name	Model features	Baseline	Novelty	Dataset	Evaluation criteria	Results
					Users > 943, Links > 33 K, Posts > 38 K		
2019	Influence Deep Learning (IDL) (Wang et al. 2019a)	No. of users & Follower relationship	Naive Bayes, Support Vector Machine & Logistic Regression	IDL outperforms the baseline methods and shows higher practical viability on large-scale datasets and improves the output of influence prediction	**Sina Weibo** No. of Nodes – 63,641 No. of Edges – 1,391,718 Avg. Degree – 12.9 Max. Degree – 70 (From 03 January 2014 to 12 May 2014)	Precision, Recall, F1-measure, & Area Under the Curve	IDL outperforms the baseline methods and IDL has higher practical viability on large-scale social networks
2020	WL+GAT (Wang et al. 2020)	Intrinsic features, Structure features & Influence features	Logistic Regression (LR), DeepFM (Guo et al. 2017), ResFM (Guo et al. 2017), & DeepInf (Qiu et al. 2018)	Systematically learning the topological features of social influence structure and developing structure-aware graph encoding techniques	**Weibo** No. of users – 624,687 No. of edges – 4,681,881 Degree of SIN – 7.49 No. of ad exposure instances – 2,269,140	Area Under Curve (AUC) & Paired t-test	Social influence features can significantly improve the prediction performance. The proposed WL+GAT outperforms the baseline algorithms.

iterates until a hierarchy of the communities is formed and maximum possible level of modularity is achieved.

Subsequently, two different types of analyses are performed for identifying the influential users. Network analysis computes the influence score of network users using the extracted features. On the other hand, quality analysis is based on *Kincaidi* metric to compute average number of syllables per word as well as average sentence length. The *Flesch–Kincaid* metric, designed specifically for English, is in fact a readability test to judge the difficulty level of a passage. The quality analysis is further enhanced by combining *Kincaidi* score with word-count and *Dale-Chall* text analysis which uses word-length to determine the difficulty of a word for a reader. Finally, users are ranked based on the aggregated network and quality analysis.

The study includes several advantages such as a novel semantic quality analysis by extracting readability and polarity of a tweet as well as investigating a user's position in two different networks with adequate verifiability. In addition, accuracy of the work is computed using *Klout* score which measures the network reach of a user and presents a correlation analysis between the content created by a user and interaction patterns of other users with that content.

2.5.4. Multilevel models for influence type and number of retweets

Besides network and quality analysis for influential user mining, research has been proposed which examines the effects of various types of influential users on brand content diffusion (Araujo, Neijens, and Vliegenthart 2017). Based on retweet behaviour, users are categorised into highly influential users, information brokers and users who influence close friends. Highly influential users may include celebrities. Similarly, information brokers have the ability to connect groups and bring new details to users who are interested but do not follow brand content. In the first phase, the focus of the study is on brand content and drawing conclusions explicitly relevant to marketing communications. Secondly, the influence of an individual is tested to verify whether it is transferrable to brand content. Consequently, brands can focus their efforts on users who can actually help extend the reach of brand message. Finally, user influence is analysed for distinguishing different categories of influential users.

Subsequently, two different models are created for the analysis of data due to the difference in the number of retweets. The first model considers original brand tweets whereas the second model includes replies of the brand to other users. In addition, the study investigates the influence effects of influential users on

retweeting behaviour of the information brokers. When an influential user retweets a brand tweet, this content causes higher levels of overall retweeting. When a brand mentions an influential user in its tweet, this process causes a higher number of information brokers retweeting which ultimately results in higher levels of overall retweeting.

The evaluation was performed using covariance analysis (ANCOVA) with number of retweets by the information brokers as dependent variable and tested the differences between brand tweets mentioning influential users from the brand tweets without mentioning influential users. The results indicate that the influential users and information brokers contribute more in terms of number of retweets for diffusion of brand content. Although information brokers have greater influence in terms of retweeting, however, they are more likely to do so when brand tweets mention influential users in its tweets (Table 6).

One of the novelties of the work is an analysis of influence process using 30 top global brands with 10 market segments using real brand messages posted by actual consumers. The study employs multilevel regression model for brand tweet analysis. Multilevel regression models are most suitable to the scenario like this where the requirement is to differentiate between the influence of user-level attributes from the influence of group-level attributes (Rabe-Hesketh and Skrondal 2008).

2.5.5. Key user discovery model based on user importance calculation

In OSNs, the trend of publishing event-related views is rapidly increasing. Therefore, finding influential users in social networks can help in analysing the impact of hot events or real-world enterprise products. Most of the existing mainstream methods which either utilise relatively simple attributes, user behaviour relation or the content association relation for building a network. In addition, these methods are unable to consider user attributes or examine the event-targeted characteristics. The study of (Zhang et al. 2020) proposed a multi-aspect user importance computation approach with event-specificity. A user in the network is evaluated at four levels within user layer, fan layer, microblog layer and event layer, respectively. The authors compute the user importance of different layers by examining users' important attributes of each level as well as the influence of these attributes on the corresponding user importance. Finally, the study integrates the user importance at different levels for calculating overall importance within the event. The experimental results are in line with the 80–20 law i.e. 20% of the people hold 80% of the wealth. Therefore, the proposed method effectively computes the event-specific user importance.

One of the novelties of the study is to divide the analysis of user importance in social network events into four levels. In addition, the study analyses a user's own attributes and the role a user plays in the specific event from different layers for characterising user importance. The feature-based models for influential user analysis are summarised in Table 5.

3. Applications

Finding influential users in OSNs has several applications, from preventing virus and disease spread to maximising dissemination of targeted marketing and advertising. Following are some of the most relevant applications:

3.1. Education

Many of the educators are using microblogging platforms such as Twitter to keep their classes organised and well-structured e.g. upcoming, and due date of assignments, ask question, etc. Twitter use in education has been classified into six different categories (Tang and Hew 2017): (1) Capture and representation (2) communication (3) collaboration (4) class organisation and administration (5) reflection and (6) assessment. On one hand, Twitter mobile apps are being used by instructors and trainers to record media, showcase their activities and ideas (Charitonos et al. 2012), capture real-world designs and exhibit class appraisal. On the other hand, students are facilitated with course contents, complete course studies, assignments, communication and other learning opportunities (Blessing, Blessing, and Fleck 2012a; Evans 2014).

3.2. Health

Twitter is widely used in U.S.A., about 64% of hospitals use Twitter for communication and announcements. Twitter allows its users to share expertise about certain topics in healthcare. For instance, Twitter history is being used to recognised depression in the U.S.A. (Tsugawa et al. 2015). Similarly, Twitter content on healthcare topics such as dental pain (Heaivilin et al. 2011), physical activity (Zhang et al. 2013), vaccination (Love et al. 2013), breast cancer (Thackeray et al. 2013), H1N1 outbreak (Chew and Eysenbach 2010), drinking problem (West et al. 2012), Zika Outbreak (Khatua and Khatua 2016), etc. is being used by public health researchers and practitioners for valuable insight into public health problems.

Table 6. Feature-based models for influential user analysis.

Year	Model name	Model features	Baseline	Novelties	Dataset	Evaluation criteria	Results
2015	Ranking domain experts by combining text and non-text features (Qi et al. 2015)	Post number, forwarded post number, follower number, mutual follower number, attention number, average agreement number, average retweet number, average comment number & total microblog number	SVM & Cosine Similarity	Integrating textual and non-textual features including newly proposed 'user quality'. Computing dynamic influence decay in message diffusion	**Sina Weibo** Users > 1 K, Posts > 280 K (based on latest 20 post pages of each user)	Accuracy	Backward greedy approach shows higher precision in optimal feature combination and SVM ranking algorithm shows higher accuracy in identifying topical experts
2016	Information Influence Measurement (Yu et al. 2016)	Retweet, comment, user quality & time	Message Ranking Method	Proposed 'user quality', 'punishment coefficient' & novel evaluation approach	**Tencent Weibo** User > 187 K, Posts > 217 K, Retweets > 199 K, Comments > 16 K	Recall	The proposed model accurately measures the information influence and reduces the influence of spam messages than baseline
2017	Semantic Quality Mining (Mahalakshmi, Koquilamballe, and Sendhilkumar 2017)	Retweet-count, Mentions count, Twitter follower to followee (tff) ratio, Average retweet-count, Followers count, Average retweet frequency & Account age	Network & tweet content analysis	Semantic quality analysis & klout score for evaluation	**Twitter** Tweets >50 K	Accuracy	The proposed model shows higher accuracy as compared to baseline
2017	Multilevel models for influence type and number of retweets (Araujo, Neijens, and Vliegenthart 2017)	Tweets, Mentions, Retweets and Replies	motive-based segmentation (Hennig-Thurau et al. 2004) & brand content diffusion (Bronner and de Hoog 2011)	Examining diffusion of tweets of top global brands by actual consumers	**Twitter** Tweets >5 K, Users > 46 K, Retweets > 74 K (From 31 August 2011–19 February 2013)	Covariance Analysis (ANCOVA)	Highly influential users and information brokers contribute most in terms of number of retweets for diffusion of brand content.
2020	Key user discovery based on user importance calculation (Zhang et al. 2020)	User ID, Authentication status, No. of fans, No. of microblogs issued within a year, No. of times these microblogs were commented, forwarded, and linked by others, Posts, Comments, & Likes	Cosine Similarity	Dividing the user importance analysis in four levels. Analysis of a user''s own attributes and his/her role in social network events	**Sina Weibo** (Ma Hang MH370 missing Event) No. of users – 146,756 No. of posts – 62,119 (From 08 March 2014 to 06 April 2014)	User Importance Intensity	The proposed model is highly effective in computing event-specific importance of social networks user and obtained results are in line with the 80–20 world's wealth law

3.3. Politics

In U.S.A., politicians use Twitter to communicate with their parliament members (Parmelee and Bichard 2011). The Barack Obama (*@BarackObama*) was the first politician who used Twitter in his political campaign. He is the most followed politician with 97 Million followers. Hilary Clinton and David Cameron also used Twitter in their Political Campaigns (Parmelee and Bichard 2011; Knight and Pattison 2015). The Australian labor Party (Simms and Wanna 2013) used Twitter to attack their opponents in politics. In Pakistan, Imran Khan (*@ImranKhanPTI*) and Mariam Nawaz used Twitter to express their political views.

3.4. Marketing and corporate communication

Due to unprecedented marketing opportunities, some of the largest brands in the world are proactive on Twitter with millions of followers. These brands rigorously promote their brand content by frequently updating their followers. As influential users can publicise a business organisation, product or services positively or negatively, therefore, business organisations should seriously recognise the significance of influential users (Probst, Grosswiele, and Pfleger 2013). In this way, negative publicity about businesses or their business clients can be tackled efficiently (Williams and Buttle 2014). Moreover, identification of influential users provides various communication strategies which can effectively promote the products or services (Enginkaya and Yılmaz 2014).

3.5. Opinion dissemination

Influential users and the network of their influential friends play a significant role in expanding the follower-network for rapid information dissemination (Aral and Walker 2012). Influential users express their views gradually on social and political issues, social events and breaking news (Zhang et al. 2016). Resultantly, the users of OSNs are influenced by these individual or network of influential users who shape the public opinion (Watts and Dodds 2007).

3.6. Curbing negative behaviour

The spread of negative behaviour has surged significantly during last few years in OSNs (Kowalski et al. 2014). Particularly, rumours have become the cause of serious social harm (He et al. 2017). Most of the studies analyse the rumour spread using network topology (Nguyen et al. 2012; Zubiaga et al. 2018). The studies employ various approaches to tackle rumour spread, for instance, curbing rumour spread at maximum number of online users (Zubiaga et al. 2018), rumour clarification by disseminating the truth (Nguyen et al. 2012) and integrating the two approaches appropriately (He, Cai, and Wang 2015). However, the structure of OSNs assists in spreading negative behaviour such as suicidal tendencies (De Choudhury et al. 2016), online bullying (Nandhini and Sheeba 2015), etc. Therefore, these issues require serious attention and investigation for reducing the related problems in online social communities.

3.7. Recommender systems

As word-of-mouth marketing is rapidly developing via diverse media such as messages, emails, blogs and microblogs (Zhao et al., 2018). Recently, integration of social network services and the online social communities has gained a substantial success on a large number of web sites such as Yelp, Douban2, etc. where millions of users can rate books, music and movies as well as share their ratings with friends and followers (Zhao, Qian, and Xie 2016a). These online platforms play a significant role in shaping a user's behaviour through the phenomenon, called recommendations, where users can recommend their favourites to friends and followers. Therefore, a deep understanding of the word-of-mouth recommendations can help design successful marketing strategies (Zhang et al. 2019). In addition, recommendation process in online rating websites is enhanced through 'explainable recommendation' which improves the effectiveness of recommender systems and generates intuitive explanations to users or system designers (Zhang and Chen 2018). Identifying influential users in social networks can positively effect the recommendations process as the majority of the consumers trust the recommendations of other consumers, friends and acquaintances (Probst, Grosswiele, and Pfleger 2013).

3.8. Spreading awareness about natural disaster management

OSNs provide real-time information and strategies related to various events happening in the world, particularly, in emergency situations or mass crisis. Valuable information from social networks can provide useful insight and timely analysis of hazard at hand (Imran et al. 2015). People communicate beyond the borders during or after disastrous situations for help or emergency response. OSNs were actively utilised in natural calamities, such as earthquakes in Haiti (Gibson and Zapdramatic 2010) and Japan (Matsumura et al.

2016). Several scholars in the domain of crisis management recommend OSNs to establish flexible social community from natural catastrophes (Matsumura et al. 2016). In this context, influential users can play an important role in spreading awareness because of their strong connection strength.

3.9. Tourism

One of the prime role of social media platforms is to provide an interactive communication channel for both tourists and tourism developers. On one hand, social media presents a medium for tourists to express their requirements and demands. On the other hand, it equips tourism developers with specialised tools to get customer feedback (Öz 2015). Mostly, the tourism developers use social media for activities such as creating social communities of interest; collecting and maintaining user-generated content; displaying videos and high-quality photography, highlighting current events; promoting word-of-mouth recommendations and customer feedback (Morrison 2013). Table 7 presents a summary of application areas and relevant references given as

4. Current challenges and future research directions

Identifying influential users from OSNs presents various research challenges. Some of the most recent issues are discussed in the following:

4.1. Lack of ground truth

As influence analysis and measurement in microblogs is truly subjective, therefore, the domain of influential users lacks the ground truth (Yang et al. 2017b). Similarly, existing studies do not employ standard or benchmark datasets, therefore, difficult to verify and validate. This, in turn, makes the problem impossible to classify under supervised machine learning. Moreover, influence is a subjective measure which makes it very difficult to have a clear benchmark or ground truth. Therefore, instead of benchmark, some other measures are compared for results evaluation, for instance, in Twitter, number of followers is considered as a standard metric for measuring the influence in the social network.

4.2. Influence maximisation

The phenomenon of influence maximisation is concerned with analysing the spread of influence in a social network and identifying ways of improvement (Kuhnle et al. 2018). In this process, few of the preliminary nodes, called seed nodes, are identified to propagate information to adopt innovation. Influence maximisation has vast viral marketing applications. However, the researchers have classified influence maximisation (IM) as NP-hard problem (Alp and Öğüdücü 2018 Kuhnle et al. 2018;), therefore, IM poses a grave challenge and difficult to be resolved in polynomial time (Banerjee, Jenamani, and Pratihar 2020). However, recent development in graph theory and SNA can be applied on real-world datasets of social media such as Facebook, Twitter, etc. for finding models for influence maximisation. Some of relevant studies include knapsack seeding of network (Kuhnle et al. 2018), community-based seeds selection (Li et al. 2018), adaptive influence maximisation (Han et al. 2018), Technique for Order of Preference by Similarity to Ideal Solution (Zareie, Sheikhahmadi, and Khamforoosh 2018) and HEDVGreedy algorithm (Aghaee and Kianian 2020).

4.3. Partial network data

In OSNs, data acquisition is quite difficult, therefore, most of the previous studies have analysed social networks using partial network data (Ferrara et al. 2014). As partial network data are a complex dataset, therefore, the algorithms and techniques based on such partial data yields skewed results which are difficult to validate

Table 7. Summary of application areas with relevant references.

	Application scenario	References
1	Education	Charitonos et al. (2012), Blessing, Blessing, and Fleck (2012a), Evans (2014), Chew and Eysenbach (2010)
2	Health	Chew and Eysenbach (2010), Heaivilin et al. (2011), West et al. (2012), Zhang et al. (2013), Love et al. (2013), Thackeray et al. (2013), Tsugawa et al. (2015), Khatua and Khatua (2016)
3	Politics	Parmelee and Bichard (2011), Simms and Wanna (2013), Knight and Pattison (2015)
4	Marketing	Probst, Grosswiele, and Pfleger (2013), Williams and Buttle (2014), Enginkaya and Yılmaz (2014)
5	Information dissemination	Watts and Dodds (2007), Gibson and Zapdramatic (2010), Aral and Walker (2012), Morrison (2013), Imran et al. (2015), Öz (2015), Matsumura et al. (2016), Zhang et al. (2016)
6	Recommender systems	Probst, Grosswiele, and Pfleger (2013), Zhao, Qian, and Xie (2016a), Zhao et al. (2018), Zhang and Chen (2018), Zhang et al. (2019)
7	Curbing negative behaviour	Nguyen et al. (2012), Kowalski et al. (2014), He, Cai, and Wang (2015), Nandhini and Sheeba (2015), De Choudhury et al. (2016), He et al. (2017), Zubiaga et al. (2018)

as complete network data are unavailable. Consequently, extracting data efficiently from complex and dynamic social networks is difficult and challenging (Al-Garadi et al. 2018). However, studies such as vital nodes identification in complex networks (Lü et al. 2016a), HybridRank algorithm (Ahajjam and Badir 2018), community-based mediator (Tulu, Hou, and Younas 2018) and K-shell hybrid method (Namtirtha, Dutta, and Dutta 2018) can prove to be effective in identifying influential users from complex social networks.

4.4. Link/Relationship strength

In OSNs, relationship or connection strength is the most significant characteristic that effects the influence level in a social network (Guille et al. 2013). The connection strength diverges from strong to weak, from best friend to mere acquaintance (Bakshy et al. 2012). However, lack of knowledge about this particular domain results in having a network with diverse connection strength (Sherchan, Nepal, and Paris 2013). Therefore, in representing information, binary relationship (e.g. a relationship describing a relationship that exists even if connection strength is ignored) generates a confusing type of relationship that results in deceptive identification and measurement of influence. In network theory, a single static edge links the nodes and describes the relationship between them. Therefore, ignoring multiple relationships among the users alters topological as well as dynamic properties of the entire system (Boccaletti et al. 2014). Also, a node's importance varies with time and network changes that can result in incorrect identification of influencers (De Domenico et al. 2015). However, in recent studies such as link prediction method based on path length (Ayoub et al. 2020) and iterative degree penalty algorithm (Tang et al. 2020), the researchers have proposed effective strategies for link prediction in social networks.

4.5. Evolution in social networks

OSNs are dynamic and continuously evolving. Therefore, one of the challenging tasks is to understand the changing dynamics in a social network. The communication among users OSNs is based on characteristics such as proximity, common interest and mutual acquaintance (Hu et al. 2014). Analysis of network evolution helps in understanding information and influence spread in advance (Lü et al. 2016a). Therefore, factors effecting a node's influence as well as network evolution should be studied comprehensively. Some of the recent studies such as Analysis of the evolution of central nodes (Uehara and Tsugawa 2019), identification of community evolution by mapping (Mohammadmosaferi and Naderi 2020) and error accumulation sensitive incremental community detection (Xu et al. 2020) can be effective in detecting community evolution in social networks.

4.6. Validation issues

Most of the influence finding algorithms are validated through information spread based methods (Tan et al. 2016 Xia et al. 2016; Yang et al. 2017a;). However, the analysis of changing dynamics of real information spread as well as human behaviour in information diffusion is not adequately analysed. A recent research (Zhang et al. 2016) identifies three possible aspects that seriously affect the information contagion e.g. homophily (Aral, Muchnik, and Sundararajan 2013), human behaviour (Aral, Muchnik, and Sundararajan 2013) and social reinforcement (Min, Goh, and Vazquez 2011). Effectiveness is one of the key characteristics in effectively determining top influential in dynamic networks (Saito et al. 2012). Existing studies are either based on greedy (Mahalakshmi, Koquilamballe, and Sendhilkumar 2017; Song et al. 2017) or heuristic approaches (Wu and Liu 2008; Tan et al. 2016; Wu et al. 2017a, 2017b).

The selection of a measurement algorithm should be based on the application requirements. For instance, heuristic-based algorithms are more suited in rapid and real-time identification of influential users. On the other hand, the proposed techniques for finding influential users in various studies employ incomplete network data, therefore, many efficiency issues have been identified (Driscoll and Walker 2014). Similarly, limitations on data collection and a different source of skewness or bias exist such as selection bias as well as assortative bias in dynamic social networks (Muchnik, Aral, and Taylor 2013). Therefore, methods designed for traditional web pages or analysis of social networks should be carefully optimised to be applicable in the context of OSNs.

However, in recent years, researchers are opting to compare the ranked lists of proposed models with manually labelled ranked lists for higher levels of validity. The measures for comparison include accuracy, precision, recall and F1-score. Some of the relevant and recent studies which employ these measures include ontology-based domain discovery (Abu-Salih, Wongthongtham, and Chan 2018), TBMUR (Eliacik and Erdogan 2018), dynamic PageRank (Shi et al. 2020), influential marketer user detection (Zareie, Sheikhahmadi, and Jalili 2019a) and time-aware domain-based social influence prediction (Abu-Salih et al. 2020).

4.7. Understanding influential users and influence spread in OSNs

Most of the studies of OSNs consider the identification of influential users as a key factor in maximising information dissemination and minimising the spread of disinformation. Influential users are generally characterised by the presence of having strong connections. However, information brokers with low-degree connection strength can significantly disseminate information in a network (Lü et al. 2016b). Furthermore, influence dissemination is not derived by influential individuals but by many easily influenced users (Cha et al. 2010). Therefore, efficient, and accurate identification of influential users and influence spread poses a serious challenge. However, potential methods such as DERND D-hops & UERND D-hops (Sun et al. 2018 Alshahrani et al. 2018b;), Threshold estimation models based on influence-weight and degree distribution (Talukder et al. 2019), Interest Group Identification and Influence Propagation models (Abd Al-Azim et al. 2020) can be effective in the identification of influential users for spreading influence.

4.8. User privacy-related issues

In OSNs, privacy becomes an issue when user's account is set as private. Therefore, information extraction from user accounts is an impossible task unless permitted by account holders. Similarly, profile visibility in OSNs is critically important to expand friends' and followers' network. However, making profiles public can cause identity slander, spamming or other such attacks (Krasnova et al. 2010). Moreover, crawlers cannot extract and analyse data from private user accounts. This, in turn, leads to misleading results or extremely personal information (Ferrara et al. 2014). Therefore, influence analysis while preserving privacy is a real challenge. However, in recent years, several models have been proposed for preserving and/or minimising the leakage of user privacy. Some of well-known algorithms include LPT-DPk (Yin et al. 2018), CenLocShare (Xiao et al. 2018), Privacy Preserving System (Phan et al. 2018), a fair mechanism for private data publication (Zheng, Luo, and Cai 2018) and DP method (Wang et al. 2019b).

4.9. Capturing influential users on Instagram

Instagram is another popular OSN which has more than one billion active users.[1] Instagram provides a platform for giving an improved representation of post contents of user. However, as compared to Twitter that is widely used for sharing opinions, news or trends around the world, Instagram is mostly used for sharing photos and videos with other network users. In this survey, we have included only those models which are employed for mining influential users in microblog networks such as Twitter or SinaWeibo.

Nonetheless, finding influential users on Instagram is a trending research domain and various methods have been proposed for identifying influential users on Instagram. For instance, Segev, Avigdor, and Avigdor (2018) proposed a meta-algorithm expansion for well-known regression models using simple user statistics. The study employs recursive feature elimination process and generates a regression model for each subset of the Instagram data. On the other hand, MPHAT (Lee, Hoang, and Lim 2019) is a novel generative algorithm which jointly models topic-specific hubs, authorities, user interest and platform preferences simultaneously from a user's network links and textual content. Similarly, Alwan, Fazl-Ersi, and Vahedian (2020) posit that combining user interaction information with user-generated content (UGC) can result in higher accuracy in differentiating between images posted by influential users and ordinary users.

5. Conclusion

The popularity of OSNs presents unique opportunities to examine and understand various aspects of social interaction among its users. In this study, we investigated several state-of-the-art influential user finding algorithms in microblogs and presented a taxonomy of these algorithms based on their underlying techniques, baseline methodologies and research contributions. Subsequently, relevant studies have been classified into five main categories including PageRank-based, centrality-based, network diffusion-based, machine learning-based and semantic and quantitative feature-based models. A significant amount of existing literature is based on PageRank variants; however, fewer feature-based models exist for identifying influential users in microblog networks. This survey has reviewed various validation strategies employed for evaluating the performance of the algorithms and presented several real-world applications of identifying influential users. In addition, the survey highlighted various research challenges and potential research directions for addressing these challenges. Finally, the techniques reviewed in this study can further be explored for their applicability on other social networks such as Facebook, Instagram, etc.

Conflict of interest

The authors report no conflict of interest.

Notes

1. https://www.businessofapps.com/data/instagram-statistics/ date accessed on Feb 28, 2021.
2. http://weibo.com/ Accessed on 04-02-2019
3. http://www.stocktwits.com/ Accessed on 10-03-2019
4. http://socialcomputing.asu.edu/uploads/1251628491/TUAW-dataset.zip/ Accessed on 15-03-2019
5. https://snap.stanford.edu/data/higgstwitter.html/ Accessed on 21-05-2019
6. http://konect.uni-koblenz.de/networks/munmun _ twitter _ social
7. http://konec t.uni-koble nz.de/netwo rks/munmu n_twitt er_social.
8. http://listatlas.com/ Accessed on 03-02-2019

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abawajy, J. H., M. I. H. Ninggal, and T. Herawan. 2016. "Privacy Preserving Social Network Data Publication." *IEEE Communications Surveys & Tutorials* 18: 1974–1997.

Abd Al-Azim, N. A. R., T. F. Gharib, Y. Afify, and M. Hamdy. 2020. "Influence Propagation: Interest Groups and Node Ranking Models." *Physica A: Statistical Mechanics and its Applications* 124247.

Abu-Salih, B., K. Y. Chan, O. Al-Kadi, M. Al-Tawil, P. Wongthongtham, T. Issa, H. Saadeh, M. Al-Hassan, B. Bremie, and A. Albahlal. 2020. "Time-aware Domain-Based Social Influence Prediction." *Journal of Big Data* 7: 10.

Abu-Salih, B., P. Wongthongtham, and K. Y. Chan. 2018. "Twitter Mining for Ontology-Based Domain Discovery Incorporating Machine Learning." *Journal of Knowledge Management* 22: 949–981.

Abu-Salih, B., P. Wongthongtham, K. Y. Chan, and D. Zhu. 2019. "CredSaT: Credibility Ranking of Users in big Social Data Incorporating Semantic Analysis and Temporal Factor." *Journal of Information Science* 45: 259–280.

Aghaee, Z., and S. Kianian. 2020. "Efficient Influence Spread Estimation for Influence Maximization." *Social Network Analysis and Mining* 10: 1–21.

Ahajjam, S., and H. Badir. 2018. "Identification of Influential Spreaders in Complex Networks Using HybridRank Algorithm." *Scientific Reports* 8: 1–10.

Ahmed, N. K., N. Duffield, J. Neville, and R. Kompella. 2014a. "Graph Sample and Hold: A Framework for big-Graph Analytics." Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining: New York, USA. 1446–1455.

Ahmed, N. K., J. Neville, and R. Kompella. 2014b. "Network Sampling: From Static to Streaming Graphs." *ACM Transactions on Knowledge Discovery from Data* 8: 1–56.

Ai, J., H. Zhao, K. M. Carley, Z. Su, and H. Li. 2013. "Neighbor Vector Centrality of Complex Networks Based on Neighbors Degree Distribution." *The European Physical Journal B* 86: 163.

Akritidis, L., D. Katsaros, and P. Bozanis. 2009. "Identifying Influential Bloggers: Time Does Matter." Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society: 76–83.

Al-Garadi, M. A., G. Mujtaba, M. S. Khan, N. H. Friday, A. Waqas, and G. Murtaza. 2018. "Applications of big Social Media Data Analysis: An Overview. Computing, Mathematics and Engineering Technologies (iCoMET)." International Conference on, 2018, IEEE: 1–5.

Al-Garadi, M. A., K. D. Varathan, and S. D. Ravana. 2017. "Identification of Influential Spreaders in Online Social Networks Using Interaction Weighted K-Core Decomposition Method." *Physica A: Statistical Mechanics and its Applications* 468: 278–288.

Al-Garadi, M. A., K. D. Varathan, S. D. Ravana, E. Ahmed, and V. Chang. 2016. "Identifying the Influential Spreaders in Multilayer Interactions of Online Social Networks." *Journal of Intelligent & Fuzzy Systems* 31: 2721–2735.

Alp, Z. Z., and S. G. Öğüdücü. 2015. "Extracting Topical Information of Tweets Using Hashtags. Machine Learning and Applications (ICMLA)." IEEE 14th International Conference on, 2015, IEEE: 644–648.

Alp, Z. Z., and Ş. G. Öğüdücü. 2018. "Identifying Topical Influencers on Twitter Based on User Behavior and Network Topology." *Knowledge-Based Systems* 141: 211–221.

Alp, Z. Z., and Ş. G. Öğüdücü. 2019. "Influence Factorization for Identifying Authorities in Twitter." *Knowledge-Based Systems* 163: 944–954.

Alshahrani, M., Z. Fuxi, A. Sameh, S. Mekouar, and S. Huang. 2020. "Efficient Algorithms Based on Centrality Measures for Identification of Top-K Influential Users in Social Networks." *Information Sciences* 527: 88–107.

Alshahrani, M., F. Zhu, M. Bamiah, S. Mekouar, and S. Huang. 2018a. "Efficient Methods to Select top-k Propagators Based on Distance and Radius Neighbor." Proceedings of the 2018 International Conference on Big Data and Computing: 78–85.

Alshahrani, M., F. Zhu, L. Zheng, S. Mekouar, and S. Huang. 2018b. "Selection of top-k Influential Users Based on Radius-Neighborhood Degree, Multi-Hops Distance and Selection Threshold." *Journal of Big Data* 5: 1–20.

Alwan, W. H., E. Fazl-Ersi, and A. Vahedian. 2020. "Identifying Influential Users on Instagram Through Visual Content Analysis." *IEEE Access* 8: 169594–169603.

Aral, S., L. Muchnik, and A. Sundararajan. 2013. "Engineering Social Contagions: Optimal Network Seeding in the Presence of Homophily." *Network Science* 1: 125–153.

Aral, S., and D. Walker. 2012. "Identifying Influential and Susceptible Members of Social Networks." *Science* 337: 337–341.

Araujo, T., P. Neijens, and R. Vliegenthart. 2017. "Getting the Word Out on Twitter: The Role of Influentials, Information Brokers and Strong Ties in Building Word-of-Mouth for Brands." *International Journal of Advertising* 36: 496–513.

Ayoub, J., D. Lotfi, M. EL Marraki, and A. Hammouch. 2020. "Accurate Link Prediction Method Based on Path Length Between a Pair of Unlinked Nodes and Their Degree." *Social Network Analysis and Mining* 10: 9.

Backstrom, L., E. Bakshy, J. M. Kleinberg, T. M. Lento, and I. Rosenn. 2011. "Center of Attention: How Facebook Users Allocate Attention Across Friends." *ICWSM* 11: 23.

Backstrom, L., and J. Leskovec. 2011. "Supervised Random Walks: Predicting and Recommending Links in Social Networks." Proceedings of the fourth ACM International Conference on WEB SEARCH and Data Mining, ACM: 635–644.

Bae, J., and S. Kim. 2014. "Identifying and Ranking Influential Spreaders in Complex Networks by Neighborhood Coreness." *Physica A: Statistical Mechanics and its Applications* 395: 549–559

Baeza-Yates, R., and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. New York: ACM press.

Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts. 2011. "Everyone's an Influencer: Quantifying Influence on Twitter." Proceedings of the fourth ACM International Conference on Web Search and Data Mining: 65–74.

Bakshy, E., I. Rosenn, C. Marlow, and L. Adamic. 2012. "The Role of Social Networks in Information Diffusion." Proceedings of the 21st International Conference on World Wide Web, ACM: 519–528.

Banerjee, S., M. Jenamani, and D. K. Pratihar. 2020. "A Survey on Influence Maximization in a Social Network." *Knowledge and Information Systems* 62 (9): 3417–3455.

Bartoletti, M., S. Lande, and A. Massa. 2016. "Faderank: an Incremental Algorithm for Ranking Twitter Users." International Conference on Web Information Systems Engineering, Springer: 55–69.

Batagelj, V., and M. Zaversnik. 2003. "An O (m) Algorithm for Cores Decomposition of Networks." *arXiv preprint cs/0310049*.

Bi, B., Y. Tian, Y. Sismanis, A. Balmin, and J. Cho. 2014. "Scalable Topic-Specific Influence Analysis on Microblogs." Proceedings of the 7th ACM International Conference on Web Search and Data Mining, ACM: 513–522.

Bilmes, J. A. 1998. "A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models." *International Computer Science Institute* 4: 126.

Biran, O., S. Rosenthal, J. Andreas, K. Mckeown, and O. Rambow. 2012. "Detecting Influencers in Written Online Conversations." Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics: 37–45.

Bizid, I., N. Nayef, P. Boursier, and A. Doucet. 2018. "Detecting Prominent Microblog Users Over Crisis Events Phases." *Information Systems* 78: 173–188.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blessing, S. B., J. S. Blessing, and B. K. B. Fleck. 2012a. "Using Twitter to Reinforce Classroom Concepts." *Teaching of Psychology* 39: 268–271.

Bo, H., R. Mcconville, J. Hong, and W. Liu. 2020. "Social Network Influence Ranking via Embedding Network Interactions for User Recommendation." Companion Proceedings of the Web Conference 2020: 379–384.

Boccaletti, S., G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. 2014. "The Structure and Dynamics of Multilayer Networks." *Physics Reports* 544: 1–122.

Bronner, F., and R. de Hoog. 2011. "Vacationers and eWOM: Who Posts, and Why, Where, and What?" *Journal of Travel Research* 50: 15–26.

Cappelletti, R., and N. Sastry. 2012. "Iarank: Ranking Users on Twitter in Near Real-Time, Based on Their Information Amplification Potential." 2012 International Conference on Social Informatics, IEEE: 70–77.

Carlson, L., D. Marcu, and M. E. Okurowski. 2003. "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory." *Current and new Directions in Discourse and Dialogue*, 85–112. Kluwer, Dordrecht.

Cha, M., H. Haddadi, F. Benevenuto, and P. K. Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM* 10: 30.

Charitonos, K., C. Blake, E. Scanlon, and A. Jones. 2012. "Museum Learning via Social and Mobile Technologies: (How) Can Online Interactions Enhance the Visitor Experience?" *British Journal of Educational Technology* 43: 802–819.

Chen, W., A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, and Y. Yuan. 2011. "Influence Maximization in Social Networks When Negative Opinions may Emerge and Propagate." Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM: 379–390.

Chen, D.-B., H. Gao, L. Lü, and T. Zhou. 2013. "Identifying Influential Nodes in Large-Scale Directed Networks: The Role of Clustering." *PloS one* 8: e77455.

Chen, D., L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou. 2012. "Identifying Influential Nodes in Complex Networks." *Physica A: Statistical Mechanics and Its Applications* 391: 1777–1787.

Chen, L., and D. Wang. 2015. "An Improved Acquaintance Immunization Strategy for Complex Network." *Journal of Theoretical Biology* 385: 58–65.

Chen, W., Y. Wang, and S. Yang. 2009. "Efficient Influence Maximization in Social Networks." Proceedings of the 15th ACM SIGKDD International Conference on Knowledge discovery and data mining: 199–208.

Chen, W., Y. Yuan, and L. Zhang. 2010. "Scalable Influence Maximization in Social Networks Under the Linear Threshold Model." 2010 IEEE International Conference on Data Mining, IEEE: 88–97.

Chevalier, J. A., and D. Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews." *Journal of Marketing Research* 43: 345–354.

Chew, C., and G. Eysenbach. 2010. "Pandemics in the Age of Twitter: Content Analysis of Tweets During the 2009 H1N1 Outbreak." *PloS one* 5: e14118.

Cohen, E., D. Delling, T. Pajor, and R. F. Werneck. 2014. "Distance-based Influence in Networks: Computation and Maximization." *arXiv preprint arXiv:1410.6976*.

Cortez, P., N. Oliveira, and J. P. Ferreira. 2016. "Measuring User Influence in Financial Microblogs: Experiments Using Stocktwits Data." Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, ACM: 23.

Dang, H. T., D. Kelly, and J. J. Lin. 2007. "Overview of the TREC 2007 Question Answering Track." *Trec* 7: 63.

Daniel Tunkelang. 2015. *TunkRank* [Online]. Accessed May 9, 2018. http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G. & Kumar, M. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016. ACM: 2098–2110.

De Domenico, M., A. Lima, P. Mougel, and M. Musolesi. 2013. "The Anatomy of a Scientific Rumor." *Scientific Reports* 3: 2980.

De Domenico, M., A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas. 2015. "Ranking in Interconnected Multilayer Networks Reveals Versatile Nodes." *Nature Communications* 6: 6868.

Dijkstra, E. W. 1959. "A Note on Two Problems in Connexion with Graphs." *Numerische Mathematik* 1: 269–271.

Domingos, P., and M. Richardson. 2001. "Mining the Network Value of Customers." Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM: 57–66.

Dorogovtsev, S. N., A. V. Goltsev, and J. F. F. Mendes. 2006. "*k*-Core Organization of Complex Networks." *Physical Review Letters* 96: 040601.

Drakopoulos, G., A. Kanavos, P. Mylonas, and S. Sioutas. 2017. "Defining and Evaluating Twitter Influence Metrics: a Higher-Order Approach in Neo4j." *Social Network Analysis and Mining* 7: 52.

Driscoll, K., and S. Walker. 2014. "Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data." *International Journal of Communication* 8: 20.

Eliacik, A. B., and N. Erdogan. 2018. "Influential User Weighted Sentiment Analysis on Topic Based Microblogging Community." *Expert Systems with Applications* 92: 403–418.

Enginkaya, E., and H. Yılmaz. 2014. "What Drives Consumers to Interact with Brands Through Social Media? A Motivation Scale Development Study." *Procedia - Social and Behavioral Sciences* 148: 219–226.

Erosheva, E., S. Fienberg, and J. Lafferty. 2004. "Mixed-membership Models of Scientific Publications." *Proceedings of the National Academy of Sciences* 101: 5220–5227.

Evans, C. 2014. "Twitter for Teaching: Can Social Media be Used to Enhance the Process of Learning?" *British Journal of Educational Technology* 45: 902–915.

Even-Dar, E., and A. Shapira. 2007. "A Note on Maximizing the Spread of Influence in Social Networks." International Workshop on Web and Internet Economics, Springer: 281–286.

Ferrara, E., P. De Meo, G. Fiumara, and R. Baumgartner. 2014. "Web Data Extraction, Applications and Techniques: A Survey." *Knowledge-based Systems* 70: 301–323.

Freeman, L. C. 1977. "A Set of Measures of Centrality Based on Betweenness." *Sociometry*, 40: 35–41.

Fronczak, A., P. Fronczak, and J. A. Hołyst. 2004. "Average Path Length in Random Networks." *Physical Review E* 70: 056110.

Ghosh, S., N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. 2012. "Cognos: Crowdsourcing Search for Topic Experts in Microblogs." Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval: 575–590.

Gibson, M., and P. Zapdramatic. 2010. "The Haiti Earthquake Experience: A Case Study." Joint International Conference on Interactive Digital Storytelling, Springer: 236–239.

Girvan, M., and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99: 7821–7826.

Gjoka, M., M. Kurant, C. T. Butts, and A. Markopoulou. 2010. "Walking in Facebook: A Case Study of Unbiased Sampling of Osns." 2010 Proceedings IEEE Infocom, IEEE: 1–9.

Goldenberg, J., B. Libai, S. Moldovan, and E. Muller. 2007. "The NPV of Bad News." *International Journal of Research in Marketing* 24: 186–200.

Goldenberg, J., B. Libai, and E. Muller. 2001. "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth." *Marketing Letters* 12: 211–223.

Golkar Amnieh, I., and M. Kaedi. 2015. "Using Estimated Personality of Social Network Members for Finding Influential Nodes in Viral Marketing." *Cybernetics and Systems* 46: 355–378.

Goyal, A., W. Lu, and L. V. Lakshmanan. 2011. "Simpath: An Efficient Algorithm for Influence Maximization Under the Linear Threshold Model." 2011 IEEE 11th International Conference on Data Mining, IEEE: 211–220.

Granovetter, M. S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78 (6): 1360–1380.

Griffiths, T. L., and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101: 5228–5235.

Grover, A., and J. Leskovec. 2016. "node2vec: Scalable Feature Learning for Networks." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 855–864.

Guille, A., H. Hacid, C. Favre, and D. A. Zighed. 2013. "Information Diffusion in Online Social Networks." *ACM Sigmod Record* 42: 17–28.

Guo, L., J.-H. Lin, Q. Guo, and J.-G. Liu. 2016. "Identifying Multiple Influential Spreaders in Term of the Distance-Based Coloring." *Physics Letters A* 380: 837–842.

Guo, H., R. Tang, Y. Ye, Z. Li, and X. He. 2017. "DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction." *arXiv preprint arXiv:1703.04247*.

Hakimi, S. L. 1962. "On Realizability of a Set of Integers as Degrees of the Vertices of a Linear Graph. I." *Journal of the Society for Industrial and Applied Mathematics* 10: 496–506.

Hamzehei, A., S. Jiang, D. Koutra, R. Wong, and F. Chen. 2017. "Topic-based Social Influence Measurement for Social Networks." *Australasian Journal of Information Systems* 21.

Han, K., K. Huang, X. Xiao, J. Tang, A. Sun, and X. Tang. 2018. "Efficient Algorithms for Adaptive Influence Maximization." *Proceedings of the VLDB Endowment* 11: 1029–1040.

Hao, F., M. Chen, C. Zhu, and M. Guizani. 2012. "Discovering Influential Users in Micro-Blog Marketing with Influence Maximization Mechanism." 2012 IEEE Global Communications Conference (GLOBECOM), IEEE: 470–474.

Haveliwala, T. H. 2003. "Topic-sensitive Pagerank: A Context-Sensitive Ranking Algorithm for web Search." *IEEE Transactions on Knowledge and Data Engineering* 15: 784–796.

He, Z., Z. Cai, and X. Wang. 2015. "Modeling Propagation Dynamics and Developing Optimized Countermeasures

for Rumor Spreading in Online Social Networks. Distributed Computing Systems (ICDCS).” IEEE 35th International Conference on, 2015, IEEE: 205–214.

He, Z., Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li. 2017. “Cost-efficient Strategies for Restraining Rumor Spreading in Mobile Social Networks.” *IEEE Transactions on Vehicular Technology* 66: 2789–2800.

Heaivilin, N., B. Gerbert, J. Page, and J. L. Gibbs. 2011. “Public Health Surveillance of Dental Pain via Twitter.” *Journal of Dental Research* 90: 1047–1051.

Hennig-Thurau, T., K. P. Gwinner, G. Walsh, and D. D. Gremler. 2004. “Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?” *Journal of Interactive Marketing* 18: 38–52.

Hu, H., Y. Wen, T.-S. Chua, and X. Li. 2014. “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial.” *IEEE Access* 2: 652–687.

Huang, L., and Y. Xiong. 2013. “Evaluation of Microblog Users’ Influence Based on PageRank and Users Behavior Analysis.” *Advances in Internet of Things* 03: 34–40.

Imran, M., C. Castillo, F. Diaz, and S. Vieweg. 2015. “Processing Social Media Messages in Mass Emergency: A Survey.” *ACM Computing Surveys* 47: 1–38.

Inman, H. F., and E. L. Bradley Jr. 1989. “The Overlapping Coefficient as a Measure of Agreement Between Probability Distributions and Point Estimation of the Overlap of Two Normal Densities.” *Communications in Statistics – Theory and Methods* 18: 3851–3874.

Jianqiang, Z., G. Xiaolin, and T. Feng. 2017. “A new Method of Identifying Influential Users in the Micro-Blog Networks.” *IEEE Access* 5: 3008–3015.

Jin, L., Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. 2013. “Understanding User Behavior in Online Social Networks: A Survey.” *IEEE Communications Magazine* 51: 144–150.

Katsimpras, G., D. Vogiatzis, and G. Paliouras. 2015. “Determining Influential Users with Supervised Random Walks.” Proceedings of the 24th International Conference on World Wide Web, ACM: 787–792.

Katz, L. 1953. “A new Status Index Derived from Sociometric Analysis.” *Psychometrika* 18: 39–43.

Katz, E., P. F. Lazarsfeld, and E. Roper. 1955. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Glencoe, Illinois: The Free Press of Glencoe.

Kempe, D., J. Kleinberg, and É Tardos. 2003. “Maximizing the Spread of Influence Through a Social Network.” Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM: 137–146.

Khan, H. U., A. Daud, U. Ishfaq, T. Amjad, N. Aljohani, R. A. Abbasi, and J. S. Alowibdi. 2017. “Modelling to Identify Influential Bloggers in the Blogosphere: A Survey.” *Computers in Human Behavior* 68: 64–82.

Khan, H. U., A. Daud, and T. A. Malik. 2015. “MIIB: A Metric to Identify Top Influential Bloggers in a Community.” *PloS one* 10: e0138359.

Khatua, A., and A. Khatua. 2016. “Immediate and Long-Term Effects of 2016 Zika Outbreak: A Twitter-Based Study. e-Health Networking, Applications and Services (Healthcom).” IEEE 18th International Conference on, 2016, IEEE: 1–6.

Kim, H., K. Beznosov, and E. Yoneki. 2015a. “A Strategic Model for Network Formation.” *Computational Social Networks* 2: 1–15.

Kim, H., K. Beznosov, and E. Yoneki. 2015b. “A Study on the Influential Neighbors to Maximize Information Diffusion in Online Social Networks.” *Computational Social Networks* 2: 3.

Kim, P.-J., and H. Jeong. 2007. “Reliability of Rank Order in Sampled Networks.” *The European Physical Journal B* 55: 109–114.

Kim, H., and E. Yoneki. 2012. “Influential Neighbours Selection for Information Diffusion in Online Social Networks.” 21st International Conference on Computer Communications and Networks (ICCCN), 2012, IEEE: 1–7.

Knight, J., and M. Pattison. 2015. *British Politics for Dummies*. London: John Wiley.

Koren, Y., R. Bell, and C. Volinsky. 2009. “Matrix Factorization Techniques for Recommender Systems.” *Computer*, 42: 30–37.

Kowalski, R. M., G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner. 2014. “Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research among Youth.” *Psychological Bulletin* 140: 1073–1137.

Krasnova, H., S. Spiekermann, K. Koroleva, and T. Hildebrand. 2010. “Online Social Networks: Why We Disclose.” *Journal of Information Technology* 25: 109–125.

Kuhnle, A., M. A. Alim, X. Li, H. Zhang, and M. T. Thai. 2018. “Multiplex Influence Maximization in Online Social Networks With Heterogeneous Diffusion Models.” *IEEE Transactions on Computational Social Systems* 5: 418–429.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. “What is Twitter, a Social Network or a News Media?” Proceedings of the 19th International Conference on World Wide Web, ACM: 591–600.

Lee, R. K.-W., T.-A. Hoang, and E.-P. Lim. 2018. “Discovering Hidden Topical Hubs and Authorities in Online Social Networks.” Proceedings of the 2018 SIAM International Conference on Data Mining, SIAM: 378–386.

Lee, R. K.-W., T.-A. Hoang, and E.-P. Lim. 2019. “Discovering Hidden Topical Hubs and Authorities Across Multiple Online Social Networks.” *IEEE Transactions on Knowledge and Data Engineering* 33 (1): 70–84.

Lehmann, J., B. Gonçalves, J. J. Ramasco, and C. Cattuto. 2012. “Dynamical Classes of Collective Attention in Twitter.” Proceedings of the 21st International Conference on World Wide Web, ACM: 251–260.

Lei, S., S. Maniu, L. Mo, R. Cheng, and P. Senellart. 2015. “Online Influence Maximization.” Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 645–654.

Leskovec, J., L. A. Adamic, and B. A. Huberman. 2007. “The Dynamics of Viral Marketing.” *ACM Transactions on the Web* 1: 5.

Leskovec, J., and A. Krevl. 2011. “Stanford Large Network Dataset Collection.” http://snap.stanford.edu/data/index.html, 2011. Accessed 9 March 2020.

Leskovec, J., and J. Mcauley. 2012. “Learning to Discover Social Circles in Ego Networks.” *Advances in Neural Information Processing Systems* 25: 539–547.

Li, X., S. Cheng, W. Chen, and F. Jiang. 2013. “Novel User Influence Measurement Based on User Interaction in Microblog. Advances in Social Networks Analysis and

Mining (ASONAM).” IEEE/ACM International Conference on, 2013, IEEE: 615–619.
Li, X., X. Cheng, S. Su, and C. Sun. 2018. “Community-based Seeds Selection Algorithm for Location Aware Influence Maximization.” *Neurocomputing* 275: 1601–1613.
Li, Q., T. Zhou, L. Lü, and D. Chen. 2014. “Identifying Influential Spreaders by Weighted LeaderRank.” *Physica A: Statistical Mechanics and its Applications* 404: 47–55.
Liu, B. 2012. “Sentiment Analysis and Opinion Mining.” *Synthesis Lectures on Human Language Technologies* 5: 1–167.
Liu, Z., C. Jiang, J. Wang, and H. Yu. 2015. “The Node Importance in Actual Complex Networks Based on a Multi-Attribute Ranking Method.” *Knowledge-Based Systems* 84: 56–66.
Liu, X., H. Shen, F. Ma, and W. Liang. 2014. “Topical Influential User Analysis with Relationship Strength Estimation in Twitter.” 2014 IEEE International Conference on Data Mining Workshop, IEEE: 1012–1019.
Love, B., I. Himelboim, A. Holton, and K. Stewart. 2013. “Twitter as a Source of Vaccination Information: Content Drivers and What They Are Saying.” *American Journal of Infection Control* 41: 568–570.
Lü, L., D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou. 2016a. “Vital Nodes Identification in Complex Networks.” *Physics Reports* 650: 1–63.
Lü, L., L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley. 2015. “Toward Link Predictability of Complex Networks.” *Proceedings of the National Academy of Sciences* 112: 2325–2330.
Lü, L., T. Zhou, Q.-M. Zhang, and H. E. Stanley. 2016b. “The H-Index of a Network Node and Its Relation to Degree and Coreness.” *Nature Communications* 7: 10168.
Ma, J., W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. 2015. “Detect Rumors Using Time Series of Social Context Information on Microblogging Websites.” Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM: 1751–1754.
Mahalakshmi, G., K. Koquilamballe, and S. Sendhilkumar. 2017. “Influential Detection in Twitter Using Tweet Quality Analysis. Recent Trends and Challenges in Computational Models (ICRTCCM).” Second International Conference on, 2017, IEEE: 315–319.
Maiya, A. S., and T. Y. Berger-Wolf. 2011. “Benefits of Bias: Towards Better Characterization of Network Sampling.” Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining: 105–113.
Mao, D. 2012. “Improved Canopy-Ameans Algorithm Based on MapReduce.” *Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications)* 48 (27): 22–26.
Mao, G.-J., and J. Zhang. 2016. “A PageRank-Based Mining Algorithm for User Influences on Micro-Blogs.” Pacific Asia Conference on Information Systems (PACIS), Association For Information System.
Matsumura, N., A. Miura, M. Komori, and K. Hiraishi. 2016. “Media and Sentiments in the Great East Japan Earthquake Related Tweets – Social Media as ‘Meta Media’. Semantic Computing (ICSC).” IEEE tenth International Conference on, 2016, IEEE: 465–470.
Mccrae, R. R., and O. P. John. 1992. “An Introduction to the Five-Factor Model and Its Applications.” *Journal of Personality* 60: 175–215.
Mcdonald, K. 2005. “Physicist Proposes New Way to Rank Scientific Output.” *PhysOrg*. http://www.physorg.com/news7971.html. Accessed 8 March 2020.
Mcpherson, M., L. Smith-Lovin, and J. M. Cook. 2001. “Birds of a Feather: Homophily in Social Networks.” *Annual Review of Sociology* 27: 415–444.
Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv preprint arXiv:1301.3781*.
Min, B., K.-I. Goh, and A. Vazquez. 2011. “Spreading Dynamics Following Bursty Human Activity Patterns.” *Physical Review E* 83: 036102.
Mislove, A., M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. “Measurement and Analysis of Online Social Networks.” Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, ACM: 29–42.
Mohammadmosaferi, K. K., and H. Naderi. 2020. “Evolution of Communities in Dynamic Social Networks: An Efficient map-Based Approach.” *Expert Systems with Applications* 147: 113221.
Moro, S., P. Cortez, and P. Rita. 2014. “A Data-Driven Approach to Predict the Success of Bank Telemarketing.” *Decision Support Systems* 62: 22–31.
Morrison, A. M. 2013. *Marketing and Managing Tourism Destinations*. New York: Taylor & Francis.
Muchnik, L., S. Aral, and S. J. Taylor. 2013. “Social Influence Bias: A Randomized Experiment.” *Science* 341: 647–651.
Namtirtha, A., A. Dutta, and B. Dutta. 2018. “Identifying Influential Spreaders in Complex Networks Based on Kshell Hybrid Method.” *Physica A: Statistical Mechanics and its Applications* 499: 310–324.
Nandhini, B. S., and J. Sheeba. 2015. “Online Social Network Bullying Detection Using Intelligence Techniques.” *Procedia Computer Science* 45: 485–492.
Nguyen, H. T., T. N. Dinh, and M. T. Thai. 2016a. “Cost-aware Targeted Viral Marketing in Billion-Scale Networks.” IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE: 1–9.
Nguyen, H. T., M. T. Thai, and T. N. Dinh. 2016b. “Stop-and-stare: Optimal Sampling Algorithms for Viral Marketing in Billion-Scale Networks.” Proceedings of the 2016 International Conference on Management of Data: 695–710.
Nguyen, N. P., G. Yan, M. T. Thai, and S. Eidenbenz. 2012. “Containment of Misinformation Spread in Online Social Networks.” Proceedings of the 4th Annual ACM Web Science Conference, ACM: 213–222.
Öz, M. 2015. “Social Media Utilization of Tourists for Travel-Related Purposes.” *International Journal of Contemporary Hospitality Management* 27: 1003–1023.
Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. “The PageRank Citation Ranking: Bringing Order to the Web.” Technical report, Stanford University, Stanford, CA.
Pal, A., and S. Counts. 2011. “Identifying Topical Authorities in Microblogs.” Proceedings of the fourth ACM International Conference on Web Search and Data Mining, ACM: 45–54.
Parmelee, J. H., and S. L. Bichard. 2011. *Politics and the Twitter Revolution: How Tweets Influence the Relationship*

Between Political Leaders and the Public. Lanham, MD: Lexington Books.

Pei, S., L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse. 2015. "Searching for Superspreaders of Information in Real-World Social Media." *Scientific Reports* 4: 5547.

Phan, T. N., T. K. Dang, T. A. Truong, and T. H. Lam. 2018. "A Context-Aware Privacy-Preserving Solution for Location-Based Services." 2018 International Conference on Advanced Computing and Applications (ACOMP), IEEE: 132–139.

Prabhakaran, V., O. Rambow, and M. Diab. 2010. "Automatic Committed Belief Tagging." Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics: 1014–1022.

Probst, F., L. Grosswiele, and R. Pfleger. 2013. "Who Will Lead and Who Will Follow: Identifying Influential Users in Online Social Networks." *Business & Information Systems Engineering* 5: 179–193.

Qi, L., Y. Huang, L. Li, and G. Xu. 2015. "Learning to Rank Domain Experts in Microblogging by Combining Text and non-Text Features." 2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC), IEEE: 28–31.

Qiu, J., J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang. 2018. "Deepinf: Social Influence Prediction with Deep Learning." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining: 2110–2119.

Rabe-Hesketh, S., and A. Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. Berkeley: STATA press.

Rand, W., and R. T. Rust. 2011. "Agent-based Modeling in Marketing: Guidelines for Rigor." *International Journal of Research in Marketing* 28: 181–193.

Rehman, A. U., A. Jiang, A. Rehman, A. Paul, and M. T. Sadiq. 2020. "Identification and Role of Opinion Leaders in Information Diffusion for Online Discussion Network." *Journal of Ambient Intelligence and Humanized Computing*, 1–13.

Ren, X., and L. Linyuan. 2014. "Review of Ranking Nodes in Complex Networks." *Chinese Science Bulletin* 59: 1175–1197.

Riquelme, F., and P. González-Cantergiani. 2016. "Measuring User Influence on Twitter: A Survey." *Information Processing & Management* 52: 949–975.

Rosenthal, S., and K. Mckeown. 2017. "Detecting Influencers in Multiple Online Genres." *ACM Transactions on Internet Technology* 17: 1–22.

Rosenthal, S., K. Mckeown, and A. Agarwal. 2014. "Columbia nlp: Sentiment Detection of Sentences and Subjective Phrases in Social Media." Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014): 198–202.

Saito, K., M. Kimura, K. Ohara, and H. Motoda. 2012. "Efficient Discovery of Influential Nodes for SIS Models in Social Networks." *Knowledge and Information Systems* 30: 613–635.

Salamanos, N., E. Voudigari, and E. J. Yannakoudakis. 2017. "Deterministic Graph Exploration for Efficient Graph Sampling." *Social Network Analysis and Mining* 7: 1–14.

Sarma, A. D., A. R. Molla, G. Pandurangan, and E. Upfal. 2013. "Fast Distributed Pagerank Computation." International Conference on Distributed Computing and Networking, Springer: 11–26.

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18: 429–434.

Segev, N., N. Avigdor, and E. Avigdor. 2018. "Measuring Influence on Instagram: A Network-Oblivious Approach." The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval: 1009–1012.

Shakarian, P., A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo. 2015. *Diffusion in Social Networks*, 47–58. Cham, Switzerland: Springer International Publishing.

Sheikhahmadi, A., and M. A. Nematbakhsh. 2017. "Identification of Multi-Spreader Users in Social Networks for Viral Marketing." *Journal of Information Science* 43: 412–423.

Sheikhahmadi, A., M. A. Nematbakhsh, and A. Zareie. 2017. "Identification of Influential Users by Neighbors in Online Social Networks." *Physica A: Statistical Mechanics and its Applications* 486: 517–534.

Sherchan, W., S. Nepal, and C. Paris. 2013. "A Survey of Trust in Social Networks." *ACM Computing Surveys* 45: 1–33.

Shi, L.-L., L. Liu, Y. Wu, L. Jiang, and J. Hardy. 2017. "Event Detection and User Interest Discovering in Social Media Data Streams." *IEEE Access* 5: 20953–20964.

Shi, L.-L., L. Liu, Y. Wu, L. Jiang, J. Panneerselvam, and R. Crole. 2020. "A Social Sensing Model for Event Detection and User Influence Discovering in Social Media Data Streams." *IEEE Transactions on Computational Social Systems* 7: 141–150.

Shi, L., Y. Wu, L. Liu, X. Sun, and L. Jiang. 2018. "Event Detection and Identification of Influential Spreaders in Social Media Data Streams." *Big Data Mining and Analytics* 1: 34–46.

Silva, A., S. Guimarães, W. Meira Jr, and M. Zaki. 2013. "ProfileRank: Finding Relevant Content and Influential Users Based on Information Diffusion." Proceedings of the 7th Workshop on Social Network Mining and Analysis, ACM: 2.

Simms, M., and J. Wanna. 2013. *Julia 2010: The Caretaker Election*. Canberra: ANU Press.

Song, G., Y. Li, X. Chen, X. He, and J. Tang. 2017. "Influential Node Tracking on Dynamic Social Network: an Interchange Greedy Approach." *IEEE Transactions on Knowledge and Data Engineering* 29: 359–372.

Srivatsa, M., L. Xiong, and L. Liu. 2005. "TrustGuard: Countering Vulnerabilities in Reputation Management for Decentralized Overlay Networks." Proceedings of the 14th International Conference on World Wide Web: 422–431.

Stonedahl, F. J. 2011. *Genetic Algorithms for the Exploration of Parameter Spaces in Agent-Based Models* (Unpublished doctoral dissertation). Northwestern University, Evanston, IL.

Stonedahl, F., W. Rand, and U. Wilensky. 2010. "Evolving Viral Marketing Strategies." Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, ACM: 1195–1202.

Su, S., Y. Wang, Z. Zhang, C. Chang, and M. A. Zia. 2018. "Identifying and Tracking Topic-Level Influencers in the Microblog Streams." *Machine Learning* 107: 551–578.

Sun, H., R. Cheng, X. Xiao, J. Yan, Y. Zheng, and Y. Qian. 2018. "Maximizing Social Influence for the Awareness Threshold Model." International Conference on Database Systems for Advanced Applications, Springer: 491–510.

Sung, J., S. Moon, and J.-G. Lee. 2013. "The Influence in Twitter: Are They Really Influenced?" *Behavior and Social Computing,* 95–105. Springer International Publishing.

Szell, M., R. Lambiotte, and S. Thurner. 2010. "Multirelational Organization of Large-Scale Social Networks in an Online World." *Proceedings of the National Academy of Sciences* 107: 13636–13641.

Talukder, A., M. G. R. Alam, N. H. Tran, D. Niyato, G. H. Park, and C. S. Hong. 2019. "Threshold Estimation Models for Linear Threshold-Based Influential User Mining in Social Networks." *IEEE Access* 7: 105.

Tan, C. W., P.-D. Yu, C.-K. Lai, W. Zhang, and H.-L. Fu. 2016. "Optimal Detection of Influential Spreaders in Online Social Networks. Information Science and Systems (CISS)." Annual Conference on, 2016, IEEE: 145–150.

Tang, Y., and K. F. Hew. 2017. "Using Twitter for Education: Beneficial or Simply a Waste of Time?" *Computers & Education* 106: 97–118.

Tang, R., S. Jiang, X. Chen, H. Wang, W. Wang, and W. Wang. 2020. "Interlayer Link Prediction in Multiplex Social Networks: an Iterative Degree Penalty Algorithm." *Knowledge-Based Systems* 194: 105598.

Tang, Y., Y. Shi, and X. Xiao. 2015. "Influence Maximization in Near-Linear Time: A Martingale Approach." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data: 1539–1554.

Tang, Y., X. Xiao, and Y. Shi. 2014. "Influence Maximization: Near-Optimal Time Complexity Meets Practical Efficiency." Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data: 75–86.

Thackeray, R., S. H. Burton, C. Giraud-Carrier, S. Rollins, and C. R. Draper. 2013. "Using Twitter for Breast Cancer Prevention: An Analysis of Breast Cancer Awareness Month." *BMC Cancer* 13: 508.

Tsugawa, S., Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. 2015. "Recognizing Depression from Twitter Activity." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM: 3187–3196.

Tsugawa, S., and K. Kimura. 2018. "Identifying Influencers from Sampled Social Networks." *Physica A: Statistical Mechanics and Its Applications* 507: 294–303.

Tulu, M. M., R. Hou, and T. Younas. 2018. "Identifying Influential Nodes Based on Community Structure to Speed Up the Dissemination of Information in Complex Network." *IEEE Access* 6: 7390–7401.

Tunkelang, D. 2009a. "TunkRank: A Twitter Analog to PageRank." http://thenoisychannel.com/2009/01/13/atwitter-analog-to-pagerank.

Tunkelang, D. 2009b. A Twitter Analog to Pagerank. The Noisy Channel, 44.

Twitter. 2019. *REST API Resources* [Online]. Accessed November 11, 2020. https://dev.twitter.com.

Uehara, M., and S. Tsugawa. 2019. "Analysis of the Evolution of the Influence of Central Nodes in a Twitter Social Network." 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), IEEE: 892–895.

Wang, Q., Y. Jin, S. Cheng, and T. Yang. 2017. "ConformRank: A Conformity-Based Rank for Finding Top-k Influential Users." *Physica A: Statistical Mechanics and its Applications* 474: 39–48.

Wang, H., Q. Meng, J. Fan, Y. Li, L. Cui, X. Zhao, C. Peng, G. Chen, and X. Du. 2020. "Social Influence Does Matter: User Action Prediction for In-Feed Advertising." Proceedings of the AAAI Conference on artificial intelligence: 246–253.

Wang, W., M. Min, L. Xiao, Y. Chen, and H. Dai. 2019b. "Protecting Semantic Trajectory Privacy for Vanet with Reinforcement Learning." ICC 2019–2019 IEEE International Conference on Communications (ICC), IEEE: 1–5.

Wang, F., J. She, Y. Ohyama, and M. Wu. 2019a. "Deep-learning-based Identification of Influential Spreaders in Online Social Networks." IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, IEEE: 6854–6858.

Wang, X., Y. Su, C. Zhao, and D. Yi. 2016. "Effective Identification of Multiple Influential Spreaders by DegreePunishment." *Physica A: Statistical Mechanics and its Applications* 461: 238–247.

Watts, D. J., and P. S. Dodds. 2007. "Influentials, Networks, and Public Opinion Formation." *Journal of Consumer Research* 34: 441–458.

Watts, D. J., and S. H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393: 440.

Wei, W., G. Cong, C. Miao, F. Zhu, and G. Li. 2016. "Learning to Find Topic Experts in Twitter via Different Relations." *IEEE Transactions on Knowledge and Data Engineering* 28: 1764–1778.

Weng, J., E.-P. Lim, J. Jiang, and Q. He. 2010. "Twitterrank: Finding Topic-Sensitive Influential Twitterers." Proceedings of the third ACM International Conference on Web Search and Data Mining: 261–270.

Weng, L., F. Menczer, and Y.-Y. Ahn. 2013. "Virality Prediction and Community Structure in Social Networks." *Scientific Reports* 3: 2522.

West, J. H., P. C. Hall, C. L. Hanson, K. Prier, C. Giraud-Carrier, E. S. Neeley, and M. D. Barnes. 2012. "Temporal Variability of Problem Drinking on Twitter." *Open Journal of Preventive Medicine* 02: 43–48.

White, D. R., and S. P. Borgatti. 1994. "Betweenness Centrality Measures for Directed Graphs." *Social Networks* 16: 335–346.

Williams, M., and F. Buttle. 2014. "Managing Negative Word-of-Mouth: An Exploratory Study." *Journal of Marketing Management* 30: 1423–1447.

Wu, Y., N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui. 2016. "Evaluation of Graph Sampling: A Visualization Perspective." *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 401–410.

Wu, Y., N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui. 2017a. "Evaluation of Graph Sampling: A Visualization Perspective." *IEEE Transactions on Visualization and Computer Graphics* 23: 401–410.

Wu, X., and Z. Liu. 2008. "How Community Structure Influences Epidemic Spread in Social Networks." *Physica A: Statistical Mechanics and Its Applications* 387: 623–630.

Wu, J., Y. Sha, R. Li, Q. Liang, B. Jiang, J. Tan, and B. Wang. 2017b. "Identification of Influential Users Based on Topic-Behavior Influence Tree in Social Networks." National CCF Conference on Natural Language Processing and Chinese Computing, Springer: 477–489.

Xia, Y., X. Ren, Z. Peng, J. Zhang, and L. She. 2016. "Effectively Identifying the Influential Spreaders in Large-Scale Social Networks." *Multimedia Tools and Applications* 75: 8829–8841.

Xiang, R., J. Neville, and M. Rogati. 2010. "Modeling Relationship Strength in Online Social Networks." Proceedings of the 19th International Conference on World Wide Web: 981–990.

Xiao, X., C. Chen, A. K. Sangaiah, G. Hu, R. Ye, and Y. Jiang. 2018. "CenLocShare: A Centralized Privacy-Preserving Location-Sharing System for Mobile Online Social Networks." *Future Generation Computer Systems* 86: 863–872.

Xu, Z., X. Rui, J. He, Z. Wang, and T. Hadzibeganovic. 2020. "Superspreaders and Superblockers Based Community Evolution Tracking in Dynamic Social Networks." *Knowledge-Based Systems* 192: 105377.

Yamaguchi, Y., T. Takahashi, T. Amagasa, and H. Kitagawa. 2010. "Turank: Twitter User Ranking Based on User-Tweet Graph Analysis." International Conference on Web Information Systems Engineering, Springer: 240–253.

Yang, L., Y. Tian, J. Li, J. Ma, and J. Zhang. 2017a. "Identifying Opinion Leaders in Social Networks with Topic Limitation." *Cluster Computing* 20: 2403–2413.

Yang, Y., Z. Wang, J. Pei, and E. Chen. 2017b. "Tracking Influential Individuals in Dynamic Networks." *IEEE Transactions on Knowledge and Data Engineering* 29: 2615–2628.

Yildiz, E., C. Tirkaz, H. B. Sahin, M. T. Eren, and O. Sonmez. 2017. "A Morphology-Aware Network for Morphological Disambiguation." *arXiv preprint arXiv:1702.03654*.

Yin, C., J. Xi, R. Sun, and J. Wang. 2018. "Location Privacy Protection Based on Differential Privacy Strategy for Big Data in Industrial Internet of Things." *IEEE Transactions on Industrial Informatics* 14: 3628–3636.

Yu, M., W. Yang, W. Wang, and G. W. Shen. 2016. "Information Influence Measurement Based on User Quality and Information Attribute in Microblogging." 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), IEEE: 603–608.

Zareie, A., A. Sheikhahmadi, and M. Jalili. 2019a. "Identification of Influential Users in Social Networks Based on Users' Interest." *Information Sciences* 493: 217–231.

Zareie, A., A. Sheikhahmadi, and M. Jalili. 2019b. "Influential Node Ranking in Social Networks Based on Neighborhood Diversity." *Future Generation Computer Systems* 94: 120–129.

Zareie, A., A. Sheikhahmadi, and K. Khamforoosh. 2018. "Influence Maximization in Social Networks Based on TOPSIS." *Expert Systems with Applications* 108: 96–107.

Zhang, N., S. Campo, K. F. Janz, P. Eckler, J. Yang, L. G. Snetselaar, and A. Signorini. 2013. "Electronic Word of Mouth on Twitter About Physical Activity in the United States: Exploratory Infodemiology Study." *Journal of Medical Internet Research* 15 (11): e261.

Zhang, Y., and X. Chen. 2018. "Explainable Recommendation: A Survey and New Perspectives." *arXiv preprint arXiv:1804.11192*.

Zhang, L., D. Jiang, R. Xue, Y. Yi, and X. Luo. 2020. "The key User Discovery Model Based on User Importance Calculation." *International Journal of Computational Science and Engineering* 21: 315–323.

Zhang, Z.-K., C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang. 2016. "Dynamics of Information Diffusion and Its Applications on Complex Networks." *Physics Reports* 651: 1–34.

Zhang, S., L. Yao, A. Sun, and Y. Tay. 2019. "Deep Learning Based Recommender System: A Survey and New Perspectives." *ACM Computing Surveys (CSUR)* 52: 1–38.

Zhao, G., X. Lei, X. Qian, and T. Mei. 2019a. "Exploring Users' Internal Influence from Reviews for Social Recommendation." *IEEE Transactions on Multimedia* 21: 771–781.

Zhao, Y., S. Li, and F. Jin. 2016b. "Identification of Influential Nodes in Social Networks with Community Structure Based on Label Propagation." *Neurocomputing* 210: 34–44.

Zhao, G., X. Qian, and X. Xie. 2016a. "User-service Rating Prediction by Exploring Social Users' Rating Behaviors." *IEEE Transactions on Multimedia* 18: 496–506.

Zhao, L., J. Wang, Y. Chen, Q. Wang, J. Cheng, and H. Cui. 2012. "SIHR Rumor Spreading Model in Social Networks." *Physica A: Statistical Mechanics and its Applications* 391: 2444–2453.

Zhao, Z., H. Zhou, B. Zhang, F. Ji, and C. Li. 2019b. "Identifying High Influential Users in Social Media by Analyzing Users' Behaviors." *Journal of Intelligent & Fuzzy Systems* 36: 6207–6218.

Zhaoyun, D., J. Yan, Z. Bin, and H. Yi. 2013. "Mining Topical Influencers Based on the Multi-Relational Network in Micro-Blogging Sites." *China Communications* 10: 93–104.

Zheng, X., G. Luo, and Z. Cai. 2018. "A Fair Mechanism for Private Data Publication in Online Social Networks." *IEEE Transactions on Network Science and Engineering* 7 (2): 880–891.

Zhou, J., G. Wu, M. Tu, B. Wang, Y. Zhang, and Y. Yan. 2017. "Predicting User Influence Under the Environment of Big Data. Cloud Computing and Big Data Analysis (ICCCBDA)." IEEE 2nd International Conference on, 2017, IEEE: 133–138.

Zhao, G., X. Lei, X. Qian, and T. Mei. 2018. "Exploring Users' Internal Influence from Reviews for Social Recommendation." *IEEE Transactions on Multimedia* 21 (3): 771–781.

Zhou, J., Y. Zhang, B. Wang, and Y. Yan. 2016. "Predicting User Influence in Microblogs." 2016 first IEEE International Conference on Computer Communication and the Internet (ICCCI), IEEE: 292–295.

Zhu, H., X. Yin, J. Ma, and W. Hu. 2016. "Identifying the Main Paths of Information Diffusion in Online Social Networks." *Physica A: Statistical Mechanics and Its Applications* 452: 320–328.

Zubiaga, A., A. Aker, K. Bontcheva, M. Liakata, and R. Procter. 2018. "Detection and Resolution of Rumours in Social Media." *ACM Computing Surveys* 51: 1–36.