# Models of Social Influence: A Survey

David Oriedi, Zahia Guessoum, Cyril de Runz, Amine Ait Younes

▶ **To cite this version:**

David Oriedi, Zahia Guessoum, Cyril de Runz, Amine Ait Younes. Models of Social Influence: A Survey. [Research Report] CReSTIC, Université de Reims. 2020. hal-03128157

**HAL Id: hal-03128157**
**https://hal.science/hal-03128157**

Submitted on 30 Dec 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Models of Social Influence: A Survey

David Oriedi, Zahia Guessoum, Cyril de Runz, Amine Aït Younes
University of Reims Champagne-Ardenne, CReSTIC Reims, France

**Abstract**

Online Social Networks (OSN) have become an integral part of human life in the world today. Social media applications have permeated almost every aspect of our lives. The concept of social influence has emerged from social networks and has brought with it stimulating research interest areas like influence maximization, viral marketing and sentiment analysis. Most researchers provide specific contexts for analyzing influence such as topics, opinions or intensity of interactions. As a result many different views exist of what constitutes influence on social network. In general, empirical social influence research has three phases: collection of social network data, formation of a social graph out of the collected data and use of the social graph to compute influence scores through specific metrics. While many surveys on Social Influence exist, there is hardly a survey that has taken an approach that explores various techniques and metrics that have been used particularly to abstract the social graph and to calculate influence out of the social graph. In this survey, we review past and current approaches that have been applied in each of these phases of social influence study over time. We also review some attributes of social network data and preprocessing activities undertaken to make such data suitable for social influence analysis. The review concludes with a summary of current trends in social influence research, existing open issues for further investigation and future opportunities for exploitation.

*Keywords*
**Social Influence, Online Social Network, Influence Metrics, Viral Marketing**

## 1 Introduction

A Social Network is a group of people connected through a defined relationship. Online Social Networks are social networks through which people interact over a communication network. In the recent past, millions of people have been connected through online social networks, for example friendship in facebook, following in Twitter or photo liking in Instagram. As a result, researchers have shown a lot of interest in social network analysis. With a huge volume of data generated by online social networks such as Twitter, Facebook and Google, a number of research areas have come up such as influence maximization, sentiment analysis and viral marketing. There is rich literature on scientific identification of influential users in an online social network. This is due to the vast amount of data generated from social networks.

Research in Social Influence has various classic applications in everyday life including protection against malware propagation[1], finding of Opinion Leaders[2], Sentiment Analysis[3], finding of Expert Persons[4] and Influence Maximization[5, 6]. Apart from these application advantages, social influence analysis on social network data reveals patterns to do with the behavior of people directs public decision making and promotes national security and economic control [7].

According to Almgren and Lee[8], research on influence measurement has been largely done on two fronts, namely *prediction based measures* and *observation based measures*. Prediction based

measures rely on network structural measures such as centrality measures to *predict* influential users in a social network. On the other hand, observation based measures seek to *quantify* the amount of influence attributed to users in the social network. However, social influence is usually viewed as an end product of an information diffusion process. Banerjee *et al.*[9] have identified two major diffussion models namely the *Independent Cascade Model* and the *Linear Threshold Model*. The *Independent Cascade Model* traces individual propagation behaviour while the *Linear Threshold Model* exhibits a collective diffusion behaviour. It is therefore important to view social influence as a twin phenomena of information diffussion and aggregation of the effect of that diffussion as it relates to particular network members of interest. There exists in literature several approaches that have been developed by researchers in order to approximate node influence on social networks. Since Social Influence analysis contexts vary from one author to another, many models have been proposed to provide definitions and metrics for computation [10]. The common motivation in each of these works is the need for models that are accurate, optimal, scalable and computationally efficient. Nevertheless, as far as we know, there are still no surveys that have been done from the perspective of the three major phases of social influence study i.e social network data collection, abstraction of the social graph and quantification of influence scores from the social graph. Figure 1 summarizes this framework of social influence research.

This survey reviews past and current approaches adopted by researchers in collecting social network data, building the social graph through abstraction and determining influence scores for members of the social network. The survey is divided into three parts:

We begin by exploring existing approaches through which influence analysis is done on the social graph. Specifically, we identify metrics that have been used through literature to carry out quantitative definitions and experiments on the social graph. The value of influence score for a particular user on the network is determined by the kind of metrics used in measuring that influence score. For example, an influence score based on the number of links to a node may not be the same as an influence score derived from the number of retweets received by a user upon retweeting.

Secondly, we review several ways through which entities and relationships on the social network can be abstracted into a social graph. The topology of a social graph is determined by the interacting entities and relationship types extracted from the social network data. Most works in literature abstract social network members as nodes on the graph and relationships as links between the edges. The survey will review frequency of user interactions, similarity of user activity times, similarity of user topics,user opinions and other common criteria for abstracting node relationships.

Thirdly, this survey explores a variety of ways through which social network data can be collected and preprocessed for social network analysis. The bulk of social network data comes from offline records of online interactions among social network users, online shopping, advertising, instant messaging as well as mobile communications[10]. Another source of this data is data that comes through live streaming from different social media applications such as Facebook or YouTube.

Finally, a summary of the survey will be provided including the current research trends in influence analysis, identified open issues and opportunities for further research in influence analysis. Figure 2 shows a summarized view of this survey.

The major contributions of this paper are summarized as follows:

- We take a *Data-Social Graph-Influence Analysis* approach of the review of the state of the art. As far as we know, we are the first to take this explicit approach to reviewing the state of the art literature. We therefore offer a broader and a clearer perspective on the phases of social influence research.

- We approach the review from a perspective that helps a reader to identify the various approaches available for adoption at different phases of social influence research. By doing this, we offer

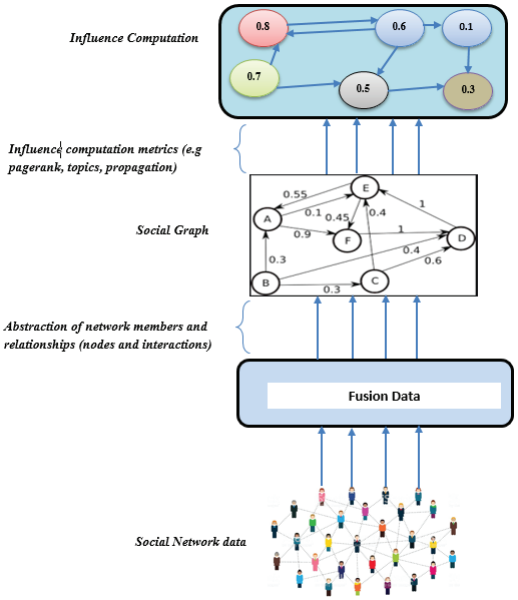


Figure 1: Framework for Social Influence Study

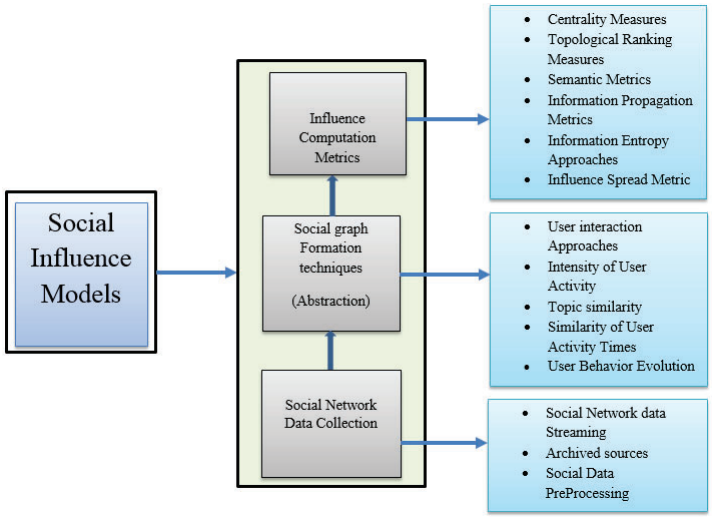


Figure 2: Survey Overview

a clear view to the reader to be able to select appropriate methods for addressing the open issues associated with each phase of social influence research.

- We identify specific strengths and weaknesses associated with the most common social influence metrics of computation. We believe this forms a good basis for identifying unresolved open issues and scaling up influence research further.

The rest of this paper is organized as follows. In Section 2, we give a detailed introduction of Social Influence. In Section 3, we provide techniques available for quantitative analysis of influence on the social graph and explore the applications associated with each. In Section 4, we explore the specific metrics used in the abstraction of social network data into a social graph while discussing the types of social graphs that are realizable for each approach. In Section 5, we review the options available for obtaining and cleaning social network data for research. Finally in Section 6, we conclude the paper and highlight some current open issues.

## 2 Definition of Social Influence

According to Merriam-Webster dictionary, Influence is the power or capacity of causing an effect in indirect or intangible way. Social Influence refers to the way in which individuals change their behavior in order to meet the demands of a social environment. Social Influence takes many forms including conformity, peer pressure, obedience, leadership or persuasion. This shows that influence takes many forms and therefore, the understanding of its analysis can vary from one context to the other.

Influencers in a social network play important roles in everyday life these days. They have special characteristics related to their individual activity, social background and connection with other members of the network that give them crucial roles in scientific and business domains[11]. Kardara *et al.* [12] define *influencers* as prominent individuals that have special traits that enable them to affect proportionately a large number of their peers with their actions. These traits are related to their activity in the network, social background and their position in the network.

Peng *et al.* [7] define Social Influence as a relationship established between two entities for a specific action. In such a relationship, one entity is said to influence another entity if the former is able to alter the opinion or behavior of the latter through their own actions. On Social Network platforms, this behavior change comes through interactive actions such as posts, shares or replies of content. Sun *et al.* [13] describe Social Influence as a behavioral change of individuals affected by others in a network. They observe that the strength of Social Influence depends on many factors like relationship strengths, temporal effects, network characteristics and individuals in the network.

From these definitions, it is evident that influencers are always active on the network. Infact in [13], it is suggested that most influencers have strong ties in their neighborhoods. According to [14], the tie strength between two nodes depends on the overlap of their neighborhoods meaning that if the overlap between nodes A and B is large, then the two nodes are considered to have a strong tie.

Some attributes of Social Influence as outlined by Peng *et al.* [7] include being dynamic, transitive, measurable, subjective, asymetric and event sensitive. Of these properties, subjectivity is probably more interesting because it shows why there has been a lot of effort from research to come up with many influence measurement metrics and models. In fact, according to Riquelme and Gonzalez [10], there is no agreement on what is meant by an influential user and hence new influence measures are constantly emerging, each with different measurement criteria.

# 3 Influence Analysis Metrics

As has been mentioned, there are several metrics that are currently being used to carry out quantitative analysis of influence on social networks. These metrics differ based on the definition of influence that a researcher chooses to go by. According to [12], a metric of an influence model is the kind of information that such a model takes into consideration when analyzing influence. The metric determines the kind of network events or objects that are used in approximating Social Influence and the relationships thereof.
In this section, we review the most common metrics that have been used in computing influence on social networks. Table 1 shows a summary of these metrics and the datasets that they have been used by various authors.

## 3.1 Centrality Based Metrics

Centrality Based models of influence mostly rely on topological metrics like the position of the node in the network and and the network community structure [15]. The general assumption being that if a node is located in an appropriate location in the graph then it could be an influential node. They are mostly used for node-level ranking through centrality metrics [16]. Centrality metrics give weight to the number of links that a node is incident to and the position of such a node in relation to the immediate neighborhood or the global network.
For the analysis of influence, centrality measures reward nodes with higher number of links and better positioning within the graph. However, it is important to note that the edge weight formation here is not determined by the frequency of interaction events but rather just the existence of interactions. In other words, the number of events between the nodes is not a major factor in forming node relationships. Even then, they have been used in recent studies for baseline comparisons, like in [17] and [18]. A major motivation for using local centrality metrics as opposed to the global ones is that local metrics are less computationally expensive and therefore can be evaluated in a relatively shorter time especially in large graphs [16]. We now review the major centrality measures that rely on metrics from this category.

### 3.1.1 Degree Centrality

The *Degree Centrality (DC)* measures influence as calculated from the number of links incoming to and outgoing from a node. If the graph is oriented, then separate values for degree centrality can be computed based on the number of inbound and outbound links. The Degree Centrality of a node $v$, is calculated as:

$$DC(v_i) = \frac{1}{N-1} \sum_{j=1}^{N} \alpha_{i,j} \tag{1}$$

where $N$ is the number of vertices and $\alpha_{i,j} = 1$, if there exists a direct link between $v_i$ and $v_j$ s.t $i \neq j$. According to this metric therefore, the more the number of links incident to a node, the more influential a node is. Several works have distinguished between indegree and outdegree to differentiate between measures based on incoming links and outgoing links respectively. The degree centrality has the limitation of being able to consider only a node's incoming edges irrespective of the centrality of its neighbors [16]. This limitation makes it unsuitable for use in cases where a node's influence is evaluated against the influence performance of its neighbors.

Table 1: **Summary of Social Influence Metrics**

| Model | Metrics Used | Metric Category | Dataset |
|---|---|---|---|
| SoRec (Li et al, 2019) | Interaction frequency | semantic | MIT reality Mining data |
| SAIM (Azzouzi and Romdhane, 2018) | retweets,replies, likes | semantic | twitter |
| CIM (Almgren and Lee, 2015) | degree, interaction frequency | centrality, semantic | Flickr Group |
| CentralityIBM (Arazkhani et al, 2019) | centrality measures | centrality | Facebook |
| GT (Li et al., 2017) | retweets | semantic | Sina Weibo, Flickr |
| OpinionRank (Zhou et al., 2009) | user opinion | opinion | Epinions dataset |
| Huang's model (Huang et al, 2014) | user emotions, topic similarity | semantic | Weibo News |
| TSR (Wang et al., 2016) | interaction frequency, topic similarity | topic | Sina Weibo |
| PHYSENSE (Sathanur et al., 2013) | user opinions | opinion | Synthesised Graphs |
| ClusterRank (Chen et al., 2013) | followers, followees, interaction frequency | topological | Delicious, SIM |
| Sheikhhahmadi's model (Sheikhhahmadi et al., 2017) | reply, retweet, mention | semantic | twitter |
| TDN (Zhao et al.,2019) | retweets, other relationships | propagation | Twitter-Higgs, Brightkite |
| Al-garadi's model (Al-garadi et al., 2016) | interaction frequency, propagation strength | propagation | twitter |
| LAIM (Wu et al., 2018) | node-node hop | propagation | NetHEPT, NetPHY, Epinions, Amazon, DBLP |
| Peng's model (Peng et al., 2017) | node-node entropy | information entropy | Cellular Network Data |
| Lu's model (Lu et al., 2019) | topic similarity, activity similarity | topic, activity similarity | Sina weibo |
| ProfileRank (Silva et al., 2013) | content, interaction frequency | semantic | twitter |
| Socialrank (Yang et al., 2012) | forwarded contents, created contents | semantic | Sina weibo |
| Yulan's model (Yulan and Ling, 2013) | replies, retweets | semantic | Sina Weibo |
| AWI(Yin and Zhang, 2012) | retweets, retweet willingness | semantic | Sina Weibo |

### 3.1.2 Closeness Centrality

*Closeness Centrality(CC)* is a global metric that is able to show how fast a node is able to communicate with others in the network [15]. A node with high Closeness Centrality value is an effective spreader of information in the network [19]. The value of closeness centrality ($CC$) for a node $v_i$ is calculated using the formular:

$$CC(v_i) = \frac{N-1}{\sum_{j=1}^{N} c(v_i, v_j)} \tag{2}$$

where $N$ is the global number of vertices and $c(v_i, v_j)$ is a function that defining the distance between nodes $v_i$ and $v_j$ s.t $i \neq j$ (i.e *min, max, mean, or median*). Essentially closeness centrality of a node measures the distance between that node and all other nodes in the network. Yang *et al.* [19] however point out that Closeness Centrality is not suitable for common influence measurement problems because it assumes that any flow on the network happens only through the shortest paths, a situation that is only possible if there is prior knowledge of the network. This is rarely the case for real-world spreading processes. Empirical experiments have shown that closeness centrality underperforms with respect to local metrics in identifying influential spreaders [20].

### 3.1.3 Betweenness Centrality

Arazkhani *et al.* [18] describe Betweenness Centrality as the number of shortest paths that pass through a node. Nodes with high betweenness centrality has capacity to facilitate interaction among nodes in different partitions of the network. Betweenness Centrality ($BC$) for a node $v_i$ is given as:

$$BC(v_i) = \frac{\sum_{i \neq k \neq j} \sigma_{i,j}(V_k)}{\sum_{j=1}^{N} \sigma_{i,j}} \tag{3}$$

where $N$ is the number of vertices, $\sigma_{i,j}$ is the geodesic paths connecting $v_i$ and $v_j$ while $\sigma_{i,j}(V_k)$ represents the number of geodesic paths including $V_k$. Although betweenness centrality is key in enhancing interactions among different segments of a network, it suffers from high computational complexity since it calculates shortest paths between all pairs of nodes in the network [15].

### 3.1.4 Eigenvector Centrality

Eigenvector Centrality ($EC$) of a node is a global measure of the extent to which a node is connected to important nodes. Li *et al.* [21] define global influence as the influence strength of a node $v$ over the whole network. This means that the $EC$ of a node is proportional to a location near the most significant nodes or communities in a graph. Given an adjacency matrix $A = (a_{i,j})$, the eigenvector centrality $x_i$ of a node $i$ is calculated as:

$$x_i = \frac{1}{\lambda} \sum_{k}^{N} a_{k,i} x_k \tag{4}$$

where $\lambda \neq 0$. A major shortcoming of Eigenvector centrality is that it is possible to have zero Eigenvector values for nodes that have no incoming links [15].

## 3.2 Topological Ranking Measures

Most Centrality Measures, except Eigenvector Centrality, assume that all nodes and links are the same in terms of their importance. However, this is not usually the case. Topological Ranking

measures provide an opportunity to reward nodes that are more connected than others and also improve the ranking of nodes that are linked to important nodes [7]. The metrics in this topological ranking category are meant to address this major weakness of centrality measures. Specifically, the metrics used here are the number of links to a node and the ranking score of a node at a particular time. Additionally, nodes that have been in the network longer are rewarded compared to more recent nodes.

### 3.2.1 PageRank

PageRank algorithm is a variant of Eigenvector centrality and is measured through PageRank algorithm [22], which is used to calculate the importance of web pages according to the number of links received by each. The importance assigned to a web page (or a node) is based on the count and quality of links to a node. In PageRank algorithm, the links that are coming from nodes that have high number of outgoing links are less important compared to those coming from nodes with less outgoing links. The PageRank Centrality $PR(v_i)$ of a node $v_i$ is:

$$PR(v_i) = \beta + \alpha \sum_{j \in W(i)} \frac{a_{k,i}}{k_j^{out}} PR(v_j) \tag{5}$$

where $W(i)$ neighbors with ingoing links to node $i$, $k_j^{out}$ is the outdegree of node $j$ if such degree is positive. Otherwise $k_j^{out} = 1$ in case $j$ is null.

The total influence is controlled through a dumping factor $\beta$. The value $\beta = (1 - \alpha)$ indicates that even if a page has no ingoing links, it will still get a small $PR(v)$. There is no universal criterion for choosing the value of the dumping factor [16]. However, Avrachenkov *et al.* [23] recommended values around 0.5 instead of the original 0.85 [22] on the basis that higher values of $\beta$ might lead to a ranking that is highly sensible to small perturbations on the structure of the network.

There are several variants of PageRank Centrality in literature. For example, it has been used to formulate topical models of influence such as InfluenceRank [2], OpinionRank [24], dynamic OpinionRank [25] as well as TopicSimilarRank [26]. It has also been popular for use in user interaction models such as in [27]. However, PageRank algorithm is not able to determine the number of nodes that can be influenced by a vertex i.e the influence spread of a node $v$, neither can it compute joint influence by a group of nodes[28]. Moreover, it is static in nature and as a result exhibits shortcomings when applied to real networks that rapidly evolve in time [16]. Finally, PageRank algorithm is based purely on the number of incoming connections and is therefore inadequate for predicting node influence based on social activity potential [29].

Some weaknesses of PageRank algorithm have been addressed through some of the variants. For example, Li *et al.* [21], developed a model in which influence of one user over the other is associated with a payoff value. The payoff is assigned a time delay that represents the time difference between the choices made by users $u$ and $v$ respectively. For example, if node $v$ adopts choice $A$ at time $t_v$, then node $v$ will have the social payoff defined as follows:

$$P_A^{soc}(v, t_v) = \sum_{u \in N_{A(v)}} a_{uv}^{soc}(t_v - t_u) + \sum_{u \in N_{B(v)}} c_{uv}^{soc}(t_v - t_u) \tag{6}$$

where $N_{A(v)}$ and $N_{B(v)}$ are the sets of neighbors of $v$ who adopt choices $A$ and $B$ respectively and $a_{uv}^{soc}$ represents the case that user $v$ made choice $A$ while user $u$ made choice $B$. After the computation of node payoffs, they use PageRank algorithm to calculate global influence for each of the nodes.

Zhang *et al.* [30] propose the use of various interactive social actions like retweets, comments and followers to model influence. According to this work, each social action is associated with a weight that reflects how important it is in the definition of an influence value. In this case, the social action consideration provides the content based evaluation for influence. A summation of the weighted social actions therefore provides content based notation for node influence. However, even though the definition of influence by Zhang *et al.* is global, important metrics like the engagement and propagation capacities have not been explicitly incorporated in this model. This makes the model limited from a user interaction perspective.

#### 3.2.2 LeaderRank

LeaderRank [31] is similar to PageRank with the exception of a *ground node* which plays a role similar to that of the dumping factor in PageRank algorithm thereby making LeaderRank parameter free. The ground node is connected to every node in the network through *bidirectional links*. A *random walk* is performed on the resulting network. The LeaderRank score is given by the fraction of time the random walker spends on a given node. The ranking process is implemented in the way of random walk described by a stochastic matrix $p$ with elements $p_{i,j} = \alpha_{i,j}/k_i^{out}$ which represents the probability of the next step of a random walker moving from node $i$ to node $j$, with $k_i^{out}$ being the out-degree. An initial score of 1 is assigned to every node except the ground node. Thereafter, the random walk step of LR is defined as the score of the node $i$ at time $t$ as follows:

$$S_i(t+1) = \sum_{j=1}^{N+1} \frac{\alpha_{i,j}}{k_i^{out}} S_i(t+1) \tag{7}$$

Lu *et al.* [31] have shown that LeaderRank is less sensitive to network perturbations and malicious manipulations of the system. A major weakness of LeaderRank is that it does not take into account the intensity of interactions that occur between users in a network.

#### 3.2.3 ClusterRank

Chen *et al.* [32] argue that although PageRank and LeaderRank take into account the influence of a node's neighbors, they do not directly make use of the social interactions among the neighbors. In other words, uncommon neighbors are likely to diffuse information further than common neighbors. For such purposes ClusterRank becomes a better ranking measure. The ClusterRank of a node $i$ is given as follows:

$$S_i = f(c_i) \sum_{j \in \Gamma_i} (k_j^{out} + 1) \tag{8}$$

where $s_i$ is the ClusterRank, $f(c_i)$ accounts for the effect of $i's$ local clustering, the term $+1$ results from the contribution of $j$ itself, $k_j^{out}$ is the outdegree of $j$ and $\Gamma_i$ is the set of followers of $i$. Although ClusterRank takes advantage of local clustering to diffuse information further into the network, it assumes same levels of influence among the nodes which is not always the case in real networks.

#### 3.2.4 FollowerRank

One of the simplest popularity measures, known as *FollowerRank*, was proposed by Nagmoti *et al.*[33]. This measure combines the number of followees and the number of followers as follows:

$$FollowerRank(i) = \frac{F1}{F1 + F3}; \tag{9}$$

in which $F1$ and $F2$ are number of followers and the number of followees respectively. This metric may not be accurate in cases where the number of followers is too high due to spammers. To mitigate this, some authors like [34], use many followees and few followers. This is meant to punish spammers.

### 3.3 Semantic Metrics of Influence

The nature of social engagement on the social network platform is such that users exchange information about various aspects of life that are of interest to them. The medium of this exchange is a set of social actions such as posts, comments, likes or shares. These social actions play an important role in enabling social network users to express themselves. Infact, Yang and Pei [28] suggest that social actions are a good way to build the edge weight between two nodes and as a result express the strength of the relationship between them. Given the crucial role played by social actions, there has been a lot of research work dedicated to defining influence from the perspective of user interaction semantics. In other words categorizing interactions such as likes, tweets, retweets or mentions according to the influence weight they have. Most authors argue that the more frequent these interactions are between nodes, the stronger the tie between them. A node receiving lots of these interactions on their posts is also perceived to be more influential.

Semantics modeling is an abstraction of social actions that users in a network use to exchange information (and the content thereof) with one another. Such social actions are ordinary posts, shares, comments and likes.

Li *et al.* [35] use a combination of retweets, comments, mentions and keyword similarity to define influence. An average of these metrics is used to represent influence. For the global approximation of node influence, they use PageRank algorithm. In [8], a hybrid of both the context and the content aspect of the network is used to define influence.

The content property of a network deals with the type and the semantics of interactions that take place among the users in the network while the context aspect brings the topological attributes of the network. For the interactive social actions both the comments and the likes have been considered. Sheikhahmadi *et al.* [36] propose NeighborRank model that calculates influence based on the number and the type of the social action in building the edge weight. To build an edge weight, they used all three interactive metrics, i.e, following, retweets and comments. A weighted summation is used to define the edge weight $w(i, j)$.

In [10], *active users* are defined as those users that are able to maintain their participation in the network in a manner that is constant and frequent for a period of time regardless of whether they receive attention for their participation. Participation in this case means carrying out actions that can be measured such as tweets, retweets, mentions and replies.

Although some authors, like Yin and Zhang [37] define activity in terms of users' ability to see tweets, it is difficult to tell if a user has seen a tweet unless they respond to through a reply, retweet, like or mention. The main strength of metrics that rely on these responses is in the assurance that these actions are a true reflection of what actually happens in the network in terms of user interactions. The metrics reviewed in this section are derived from actual participation of network members.

Azzouzi and Romdhane [27] propose a model that considers social actions such as retweets, replies and favorites in twitter. They associate each social action with a weight in the range [0,1] to differentiate the influence effect of each. According to this metric, the Influence $W(u_x, u_y)$ of a node $u_y$ on another node $u_x$ is calculated based on the weighted summation of the social actions that

node $u_x$ has generated in response to $u_y$. The influence value is calculated as shown in equation 10. In the equation, $N_{py}$ is the number of published contents by $u_y$, $N_{ai}(u_x, u_y)$ is the number of social actions $a_i$ performed by $u_x$ on the published contents of $u_y$ and $\alpha_i$ is the the weight associated a type of social action.

$$W(u_x, u_y) = \frac{\sum_{i=1}^{n} \alpha_i \times N_{ai}(u_x, u_y)}{N_{py}} \tag{10}$$

To calculate the influence of a node, this value is then incorporated into a personalized PageRank algorithm as shown in equation 10.

$$IP(u_x) = d \times \left( \sum_{u_y \in Followers(u_x)} \frac{W(u_y, u_x) \times IP(u_y)}{Followees(u_y)} \right) + (1-d)\frac{|Followers(u_x)|}{N} \tag{11}$$

where $d$ is the dumping factor, $N$ is the number of nodes and IP is the Influence Power.

A new metric called *Aquintance-Affinity Score* (AA) is introduced in [10] which measures how important a user is by gauging how well known are the users that are interested in him/her. The metric is calculated for a user $j$ as:

$$AA(j) = \sum_{i \in E_{RP}} A(i) \cdot \frac{\#replies\ of\ i\ to\ j}{\#replies\ of\ i} + \sum_{i \in E_M} A(i) \cdot \frac{\#mentions\ of\ i\ to\ j}{\#mentions\ of\ i} + \sum_{i \in E_{RT}} A(i) \cdot \frac{\#retweets\ of\ i\ to\ j}{\#retweets\ of\ i} \tag{12}$$

where $E_{RP}$, $E_M$ and $E_{RT}$ are the set of users who reply, mention and retweet the tweets of $j$ respectively. $A(i)$ is called the *Acquintance-Score* of node $i$ and is computed as follows:

$$A(i) = \frac{F1 + M4 + RP3 + RT3}{n} \tag{13}$$

with $F1$ being the number of followers, $M4$ the number of users mentioning the author and $RP3$ the number of users who have replied the authors tweets. $RT3$ and $n$ represent number of users who have retweeted the author's tweets and the number of considered user accounts respectively.

In [38], Yulan and Ling introduce user activity in PageRank algorithm by arguing that the original PageRank algorithm does not take into account user activity and the duration of such activities. Their proposal was based on the premise that since user activity levels are usually not the same on a social network, relationship strengths cannot be the same. They introduced the parameter $A = \frac{a}{T}$ into the PageRank formular, with $a$ representing the total number of posts by the user over a time period $T$. Consequently, the modified PageRank formular, as presented in equation 10 is not only ranking users but most importantly doing so based on their activity.

$$PageRank(p_i) = d + (1-d)A \sum_{p_j \subseteq M(p_i)} \frac{PageRank(p_j)}{L(p_j)} \tag{14}$$

The parameter $d$ is the dumping factor, $L(p_j)$ are the followees of $p_j$, and $M(p_i)$ are the followers of $p_i$.

Yang *et al.* define a model called *SocialRank* in [39] that uses content forwarded($C_f$) and content created($C_c$). They argue that influence score of users are dependent on the amount of content that

they either create or forward. According to this model, the influence of the created content depends on the number of reviews and forwarding that it attracts, and they created a relationship between the content created and the content shared.. The factor $\beta$ is added to make created contents more important than forwarded contents.

$$\beta_k = \begin{cases} 1, k \in C_c \\ \beta_t, k \in C_f \end{cases}$$

In this relation, $k$ represents content, $\beta_t \leq 1$ is the ratio of the score of the forwarded content and created content.The value of $\beta_k$ is then used as a major factor in the computation of node influence.

A model known as *PHYSENSE*, developed in [29], is anchored on the idea of *activity potential*. Activity potential of a user $i$ refers to the total probability of the user engaging in activity on a given topic at a given time. According to this study, user activity is divided into intrinsic and influenced activities. An intrinsic activity being an activity that shows that the user is not susceptible to interpersonal influence and influenced activity is the state of choosing to be influenced by each of the connections according to certain conditional probabilities. Intrinsic and influenced activities are related as shown in equation (15):

$$p_A^T(t,\omega)_i = (1-\alpha_{ii})(p_A^I(t,\omega)_i) + \alpha_{ii}(p_A^S(t,\omega)_i) \tag{15}$$

in which $p_A^T(t,\omega)$ is the probability of user $i$ engaging in an activity on topic $\omega$ at time $t$ over the online social network. $p_A^I$ and $p_A^S$ are the influenced activity potential and the intrinsic activity potential respectively. $\alpha_{ii}$ is the probability of user $i$ choosing to post intrinsically (equivalent to the lack of susceptibility to interpersonal influence) and is given as:

$$\alpha_{ii} = \frac{SA_i}{TA_i}; TA_i = SA_i + IA_i \tag{16}$$

where $TA_i$, $SA_i$ and $IA_i$ refer to the total, intrinsic (self) and influenced activity on the part of user $i$ in the time interval of interest. User activity is measured as number of own *tweets, replies, mentions, retweets, shares, likes or comments*. Edge weights are built based on high conversion rate to the influenced activity rather than just the amount of activity i.e:

$$w_{ij} = \frac{CF_{ij}}{\sum_{j \to i} CF_i j} \tag{17}$$

where $CF_{ij}$ is the fraction of tweets from user $j$ retweeted or shared or liked or commented by user $i$.

A common concern for researchers investigating user interaction modeling is the high computational overhead that comes with the re-computation of cumulative interactive social action effects especially in dynamic networks [40]. It is for this reason that Zhao *et al.* [40], propose a streaming algorithm that processes an interaction stream directly in a streaming fashion.

Secondly, as observed by Zhiyuli *et al.* [41], social networks are inherently hierarchical. This means that the strength of influence along a path that connects any pair of nodes fades with additional hops. This seems to partly challenge the transitive property of influence upon which propagation of influence through user interactions depends.

## 3.4 Information Propagation Measures of Influence

These metrics are based on the propagative nature of information flow on social networks. They are used to measure influence of nodes based on how well a node is able to activate other nodes through information propagation process. As observed by Silva *et al.* [42], influence can also be viewed as a measure of the ability to popularize information through diffusion. This is the context in which the models in this section have been discussed. They have been extensively applied in the study of marketing research [43], epidemiology [44] and behavioral research [45].

Al-garadi *et al.* [46] use the concept of propagation and engagement to quantify the activity of a node within its neighborhood. According to Al-garadi *et al*, propagation is the ability of a node to consistently post contents that compel its neighbors to share further down in the network. Engagement power describes the ability of a user to share contents that evoke reactive tendencies among the neighbors. To get the global influence value for each node, this model uses the k-shell algorithm together with measures of content engagement and propagation ability. A node's ability to spread influence is determined using a combination of imprecision function and a recognition rate.

### 3.4.1 Independent Cascade Model

In the *Independent Cascade* model, each edge is associated with a probability of infection which can be assigned based on frequency of infections, geographic proximity or historical infection traces[47]. An activated node infects its neighbor based on its infection probability assigned on the edge connecting with the neighbor. In each step $i \geq 1$, each node activated in step $i-1$ has a single chance to influence its inactive out-neighbor $v$ with an independent probability $p_{uv}$. According to the general cascade model, when a node $u$ attempts to activate another node $v$, it succeeds with probability $p_v(u, S)$, where $S$ is the set of neighbors that have already tried to activate $v$ and failed. The *independent Cascade Model* is the special case where $p_v(u, S)$ is a constant $p_{u,v}$, independent of $S$[6]. The propagation process usually terminates when there are no more new nodes to activate. However, there are concerns in literature that the Independent Cascade Model does not reflect the real life propagation of information due to its reliance on randomly assigned values to represent node to node influence.

### 3.4.2 Linear Threshold Model

Under the *Linear Threshold* model, a node $v$ gets to select a uniformly random threshold influence value $\theta_v$ which is in the interval [0,1]. At each time step $t$, where $H_{t-1}$ represents the set of nodes that have been activated at time $t-1$ or earlier, each inactive node becomes active on condition that:

$$\sum_{u \epsilon \eta^{in}(v) \cap H_{t-1}} b(u, v) \geq \theta_v \tag{18}$$

where $b(u, v)$ is the edge weight of the edge (u,v) and $\eta^{in}$ is the set of incoming edges. Unlike in the $IC$ model, this model is able to incorporate negative influence effects in the propagation process [9]. Although in most cases the diffusion probability is assumed, there are studies that have proposed the computation of this probability [48, 49, 50]. Similarly, this model relies on probabilistic appproximations of node to node influence and so does not accurately reflect real life formation of influence which is based on actual user interactions.

### 3.4.3 Continuous Time Independent Cascade Model

The *Continuous Time Independent Cascade* model has a length distribution associated with each edge [51]. During influence diffussion process, this model first samples a length $\tau_{uv}$ for each edge $(u, v)$ and then activates any vertex that can be reached from the seed set through a path whose total length is no more than $t$. $\tau_{uv}$ represents how long $u$ takes to influence $v$. Accordingly, given a seed set $S$ and a time threshold $t$, the influence of $S$ is the expected number of vertices activated within time $t$ during the influence propagation started from $S$. Lin *et al.* [52] propose a hybrid model that considers cumulative influence of all generated seed nodes in each iteration as opposed to what Hill Climbing and Greedy algorithms do. Specifically, this work uses a variant of the $LT$ model to reduce the performance weaknesses of both the Hill Climbing and Greedy algorithm. According to this model, to calculate the influence of a node $u$ at the $i^{th}$ iteration after successfully activating its neighbor $v$, the following relation is used:

$$val(u) = val(u) + \frac{active(v) - inf(u)}{active(v)} \tag{19}$$

where $val(u)$ is the present influence value of node $u$, $active(v)$ is the threshold value of node $v$ assigned according to the $LT$ model and $inf(v)$ is the total influence attributed to node $v$. Wu *et al.* [53] propose a model called *LAIM* that uses a local computation of influence to approximate a global influence value for a node. The local computation of influence involves an iterative process in which node influence values are calculated for each node at different neighbor levels denoted as $\lambda$. In other words, to determine a local influence value for a node $u$, a scope $i \leq \lambda$ is defined within which the cumulative influence of the node $u$ will be determined. This local influence value is then used to approximate the global influence value. This model of propagation is based on time although it does not address the issues associated with inactivity over a period of time or pheromone.

### 3.4.4 Majority Threshold Model

In an attempt to address the randomness of the values used to represent node to node influence in both the *Independent Cascade Model* and *Linear Threshold Model*, this model is introduced to provide a heuristic way of defining such a value. This model is proposed by Valente [54] and requires that the threshold node influence be defined in terms of the number of neighbors that are already influenced. In this case, that threshold is that half the neighbors must be influenced for a node to be influenced.

### 3.4.5 Shortest Path Model

Kimaru *et al.* [55] introduce this model as a variant of the Linear Threshold model in which an active node gets a chance to be activated only through shortest paths from the initially active nodes at a given time, $t$. That is:

$$t = \left( u \in \mathcal{A}_0, v \overset{min}{\in} V(G) \setminus \mathcal{A}_0 \right) dist(u, v) \tag{20}$$

in which $\mathcal{A}_0$ denotes the set of active nodes at time $t = 0$. In this way, there is some gain on the amount of time taken to process the whole graph.

While diffusion models have been extensively adopted in influence analysis and maximization research, using them in their original form to compute influence is NP-hard [6]. Research has therefore concentrated on how these models can be made to be more optimally efficient [56, 57] and

scalable in approximating influence in large networks [58, 59]. The other shortcoming associated with this category of models is that they use assumed uniform probabilistic values to represent influence values along vertex edges. This situation can easily lead to over estimation of influence[60]. Furthermore, while classifying them as *theory-centric models*, Li *et al.* [21] observe that this category of models usually use randomly distributed parameters that are not learned from actual diffusion data and are not therefore representative of real life situations.

To address the challenges that come with influence maximization in dynamic networks, Liqing *et al.* [61] propose an approach in which they greedily select the most influential node based on a *Power Law* delay distribution. This algorithm incorporates the temporal factor by considering the delayed influence propagation. Each node is associated with a distribution for the influence delays which utilizes the Power Law delay.

## 3.5 Influence Spread Measures

Influence Spread is an influence measurement metric that returns the number of nodes that have been activated or influenced by a node directly or indirectly. The activation process is based on a process of propagation usually implemented through a diffusion model. In the context of social influence research, influence spread is one of the outputs of information propagation process in the social network [62, 63, 64].

It is important to clarify that a majority of empirical works on the determination of influence spread (Influence Maximization) heavily rely on diffusion models some of which have been described in Section 3.4. Additionally, there are diffusion models that do not necessarily approach information diffusion from a probabilistic perspective. A case in point is [27], in which, instead of randomly assigning edge or node influence thresholds, the authors use actual social actions between pairs of users to generate such values.

In [65], node influence spread is calculated using *Influence Spread Paths(ISP)*. An ISP is a path that links nodes from the seed set $S$ to other nodes in the graph under study. With a given seed set and a social graph $G = (V, E)$, a simple path $p = (u_1, u_2, u_3, ...u_k)$ in graph $G$ is an $ISP\ iff\ u_1 \in S$ and $u_i \notin S$ for $i \neq 1$ and $k > 1$.

For an $ISP\ p$, the length of $p$ is $\sum_{i=1}^{k-1} length(e_i)$ and the probability of $p$ is $\prod_{i=1}^{k-1} prob(e_i)$. Further, the authors introduced a time constraint that requires the activation to occur within a time limit, beyond which the activated node will not be considered as part of those contributing to the influence spread of a node $u \in S$. Thereafter, Depth First Search algorithm is used to get all *ISPs* starting from $S$. The paths are then divided into disjoint sets based on their ending nodes. Finally, activation probabilities for all nodes are summed together and returned as the expected influence spread.

Azzouzi and Romdhane [27], while computing influence spread values for influential nodes in a network, implement an *Influence Breadth First Search* tree. In their work, the authors generate the seed set as opposed to having a seed set as the input.
To do this, a set of seed set candidates $B$, is generated from the graph based on the condition that the influence power of such nodes must be greater than the average of the influence power of its zone of influence, $I_{L_o}(u)$.

$$B = \{u : u \in V \wedge IP(u) > I_{L_o}(u)\} \tag{21}$$

where $IP(u)$ is the Influence Power of node $u$. For each node $u \in B$, an Influence BFS tree is formed in which the root is node $u$. Eventually, redundant the number of trees generated is equal to the size of set $B$. A ranking is then defined for each of the Influence BFS trees. The ranking is based

on the individual ranking of the vertices in such a tree. The reader is welcomed to refer to [27] for a better understanding. Equation 22 summarizes the ranking,

$$Rank(T^v) = \frac{1}{\mid A_v \mid} \sum_{u_i \in T^v} rank(u_i, T^v) \tag{22}$$

in which $A_v$ is the set of vertices in the BFS tree. The tree with the least ranking is selected (and by extension the seed node) since it has the quickest broadcasting of information from its root to the rest of the nodes hence a better spreader of information.

In [66], a *Maximum Gap Selecting*(MGS) algorithm is proposed to deal with the problem of approximating influence maximization in dynamic networks. Given a network graph transition from $G^0$ to $G^t$, the $MGS$ algorithm applies the *Degree Discount Algorithm* algorithm [67] on $G^0$ and $G^t$ to determine influence spread performance gap for each of the nodes $v \epsilon G^0$ as shown in the equation below:

$$PG_v = \sigma_v(S) - \sigma_v(S_0) \tag{23}$$

in which $\sigma_v(S)$ and $\sigma_v(S_0)$ represent influence spread values at time $t$ and time $t-1$ respectively. The influence spread is computed by taking the sum of the indegree in the seed set $S$. Emerging edges are built based on the time weight importance of network structures at previous time stamps using the relation $Tw(t) = e^{-\psi(T-t)}$. Influence spread measures have been very popularly used in determining node influences. But they ignore a key element of influence namely the level of node activity. This is what information entropy based metrics address.

### 3.6 Information Entropy Measures of Influence

Information Entropy provides a way to gauge the level of activity of a node based on how unpredictable the information coming from the node is. It measures how unpredictable and disorderly sources of information can be [68]. In other words, there is more information coming from an unpredictable event or source than there would be from an event or source whose outcome is known. For a random variable $X$ made up of $n$ symbols, each symbol $x_i$ with a probability $P_i$ of appearance, the *Entropy* $H$ of the source $X$ is defined as:

$$H_b(X) = -\mathbb{E}[log_b P(X)] = \sum_{i=1}^{n} P_i log_b \left(\frac{1}{P_i}\right) = -\sum_{i=1}^{n} P_i log_b P_i \tag{24}$$

where $\mathbb{E}$ denotes mathematical expectation and $log_b$ the logarithm to base $b$ and the value of $b$ is 2. Shannon [68] argue that an information transmission system has five parts namesly the information source, a transmitter, a channel, a receiver and an intended destination. Wang *et al.* [69] has argued that effective nodes receive, comprehend and transfer information through a diffussion process along social network structures. It can therefore be seen that there is a link between user activity and entropy. Influence computation through information entropy is therefore based on how predictable the events in the social network are and how this affects the stability of such a network over a period of time. Ma *et al.* [70] formulate a measure of network stability using user behavioral stability measure. They defined the entropy of a behavior of a network in terms of the ability of such a network to remain stable or otherwise from the perspective of user behavior. The network stability is expressed as a measure of the entropy of its user behavior as follows:

$$\mathbb{E} = -\sum_{i=1}^{n} P_i log_b P_i, \quad P_i = d_i / \sum_{j=1}^{n} d_j \tag{25}$$

where $\mathbb{E}$ is the entropy of behavior network and $p_i$ is the proportion of the degree $d_i$ of node $i$ to the sum of the degree of all nodes. For popular posts, there is lots of discussion by a large number of users leading to several behavior types and therefore high entropy. A higher entropy value associated with a node can therefore be interpreted as a high influence score on the node since the entropy reflects the unpredictable intensity of interactions around the node.

In one of their recent works, Yang *et al.* [71] propose a heterogeneous definition of both direct and indirect influence from a node. To quantify influence they used link entropy and interaction entropy to represent the degree for a node $i$ and node interaction with its neighborhood respectively. Equations (27) represents the link entropy, with $N_i^{(l)}$ being the degree of node $i$, while equation (28) shows interaction entropy with the variable $M_{p_i}$ representing an interaction matrix.

$$I_i^{f^{(l)}} = -\sum_{j=1}^{N_i^{(l)}} \frac{1}{N_i^{(l)}} log_{10} \frac{1}{N_i^{(l)}} \tag{26}$$

$$I_i^{c^{(l)}} = -\sum_{i=1}^{N_i^{(l)}} \frac{M_{p_i}(i,j)}{\sum_{k=1}^{N_i^{(l)}} M_{p_i}(i,k)} log_{10} \frac{M_{p_i}(i,j)}{M_{p_i}(i,k)} \tag{27}$$

As a result of these separate metrics, the total direct influence that node $i$ has on its one-hop friend nodes, $DI_i^{(l)}$ is then represented as:

$$DI_i^{(l)} = \alpha I_i^{f^{(l)}} + \beta I_i^{c^{(l)}} \tag{28}$$

where $\alpha$ and $\beta$ denote the weight assigned to $I_i^{f^{(l)}}$ and $I_i^{c^{(l)}}$ respectively and $\alpha$+$\beta$=1

Similarly, Peng *et al.* [72] applied information entropy in determining influential users in a smartphone network. They defined the edge weight as:

$$W_{ij}(t) = min\left\{C_{ij}(t), C_{ji}(t)\right\}; \tag{29}$$

where $C_{ij}(t)$ is the number of SMS/MMS sent from node $i$ to node $j$ after time duration $t$. The entropy of friend nodes $I_i^f(t)$ for a node $i$ is then given as:

$$I_i^f(t) = -\sum_{i=1}^{N_i(t)} \frac{1}{N_i(t)} log_{10} \frac{1}{N_i(t)}; \tag{30}$$

in which $N_i(t) = \sum_{j=1}^{N} f_{ij}(t)$ and $f_{ij}(t)$ represents friendship between $i$ and $j$ at time $t$.

For interaction entropy the interaction frequency among friend nodes is taken into consideration as follows:

$$I_i^c(t) = -\sum_{j=1}^{N_i(t)} \frac{C_{ij}(t)}{\sum_{k=1}^{N_i(t)} C_{ik}(t)} log_{10} \frac{C_{ij}(t)}{\sum_{k=1}^{N_i(t)} C_{ik}(t)} \tag{31}$$

The two entropy types are combined to give the total direct influence of $i$ on its direct friends is then given as $DI_i(t) = \alpha I_i^f(t) + \beta I_i^c(t)$.

In summary, this section has explored several metrics that have been adopted by various authors in quantifying influence at the social graph level. It is imperative to understand that the metrics used in computing influence on the social graph have a direct connection with the definition that has

been adopted for influence. For example if influence has been interpreted to mean the followers of a user, then its computation will be limited to the calculation of node degree. This cannot be the same for a case where influence has been defined in terms of the frequency of interactions. The theoretical definition of influence therefore determines the kind of metrics that are used in formalizing the model for calculating influence.

# 4 Social Graph Formation Approaches

There are several techniques that have been used in literature in defining an abstraction methodology for the building of a social graph from the social network data. The fundamental role of the techniques used in this case is to provide a means of building the node to node edge weight and representing the nodes themselves. Since the node attributes do not change much, a majority of these metrics are mostly dedicated to the realization and maintenance of an edge $w(i, j)$ between nodes. Generally, the value of the edge weight reflects the strength of the relationship between the nodes.

As expected, the existence of these different types of metrics translates into different interpretations of node to node edge weights and different authors give definitions suitable for their research needs and the format of data that they are using in their experiments.

## 4.1 Homophily

Homophily has been formally defined as the tendency of users in a social graph to associate with others who are similar to them along certain attribute lines such as gender, race, occupation or political views [73]. Usually, a pair of users can be described as homophilous if one or more of their attributes match in a proportion greater than other relationships within that network.

Zardi *et al.* use *static* homophily to build edge weights between nodes based on the similarity of node attributes such as age, gender education, occupation and families. For every similar pair of attributes, the edge weight is increased by a factor $\alpha$ that represents the importance of that attribute i.e

$$w(i, j) = w(i, j) + \alpha_x \tag{32}$$

where $x$ represents a node attribute. Homophily provides an attribute based relationship building among nodes in the network although this leaves out intensity of interactions among users.

## 4.2 Intensity of User Activity

In this case, intensity and frequency of interactions among network members is given prominence in forming node relationships on the graph. This means that the strength of node relationships (edge weights on the graphs) may increase or decrease depending on the frequency and intensity of their interactions. Therefore network members that do not engage one another would have relatively weaker relationship strengths compared to those that actively engage one another.

In [72], a network of smartphone communication is created out of Short Message Service (SMS) exchanges. This work abstracts the graph based on the presence and intensity of the messages exchanged among smartphone users to create both the nodes and the edges between them. The relationship weights are denoted as $W_{ij}(t)$ indicating that at a given time $t$ node $i$ sent a message to node $j$. Therefore, the more messages are exchanged between nodes $i$ and $j$, the more the edge weight between them. In order to prevent possible spamming effects in this kind of weighting, a

minimum value of the bidirected edge weights is considered, that is:

$$W_{ij}(t) = min\{C_{ij}(t), Cji(t)\} \tag{33}$$

where $C_{ij}(t)$ denotes the number of messages sent from node $i$ to node $j$.

In [27] the number of tweet replies, retweets and favorites are used to define the strength of user relationships forming the edge weights. Furthermore, the authors associated each of these actions with a weight of importance. For example, a retweet carries more weight than a favorite. In this way user whose tweets attract a lot of retweets scores higher influence than the one whose tweets only attract favorites. This proposal is shown in equation 34.

$$W(u_x, u_y) = \frac{\sum_{i=1}^{n} \alpha_i \times N_{ai}(u_x, u_y)}{N_{py}} \tag{34}$$

In order to model model node relationships, Chen *et al.* [74] use both reply relationships and the time at which users are making posts on the social network. A reply relationship is established if two users reply to the same post. They were able to check the neighborhood similarity for both nodes through Jaccard Similarity index. With the similarity index, the posting time and reply relationship, the edge weight is built as follows:

$$w(u,v) = \begin{cases} \frac{sim(u,v)}{|T_u - T_v|}, if T_u \neq T_v \\ sim(u,v), if T_u = T_v \end{cases}$$

where $sim(u, v)$ is the Jaccard Similarity Index between the adjacent node set of nodes $u$ and $v$.

## 4.3 Topic and Opinion Based Techniques

Online discussions are always on various topics that trend from time to time. A trending topic in turn attracts lots of opinion expression from users. Since some users post and others air opinions, relationships naturally develop. The abstraction of relationships in a network is based on the content of topics shared by users. This abstraction relies on topic contents and the kind of interest that such topics generate from the users. According to this approach, user relationships are tracked on the basis of topical interest or the similarity of topic interests. According to [10], unlike most of the other models that use only relational interactions among network members, this abstraction approach models relationships by analyzing the content and similarity of the information shared among the members of the network. The abstraction uses subject topics as their main approach to graph abstraction.

Bogdanov *et al.* [75], proposed a model called *genotype* through which they were able to summarize a user's topic-specific footprint in the information dissemination process. In this model, a user's topic distribution is monitored based on their interest in Twitter's topical hashtags. The *genotype* provides a multi-dimensional feature space that summarizes the observable behavior of user $u$ with respect to different hashtag topics on a 1:1 basis. To determine an influential topic, an *influence edge* $e_i(u, v)$ is defined between a *followee* $u$, who has adopted at least one hashtag $h$ within a topic $T_i$ before the corresponding follower $v$. A subnetwork $N_i(U, E_i)$ for topic $T_i$ is extracted in which the weight of the edges is determined by the number of hashtags adopted by the followee after the corresponding follower within the same topic.

In [76], a model based on a variant of the $IC$ model is used to determine the activation probability of nodes based on a user's topic popularity ranking [77]. Topic popularity $TP^t_{u,v}$ between two users

can be calculated using the following equation:

$$TP^{t}_{u,v} = \frac{Hub^{t}_{u,v}}{Hub^{t_i}_{max} + Hub^{t_i}_{min}}, (u, v \in V, ti \in t) \tag{35}$$

where $Hub^{t}_{u,v}$ denotes the hub value of a user in topic $t$, $Hub^{t_i}_{max}$ is the maximum hub value of a user's topics and $Hub^{t_i}_{min}$ is the minimum hub value of a given user's topics.
In addition, the authors define *user intimacy* which is the frequency of connection between two users $u$ and $v$. This relationship is expressed as follows:

$$C_{u,v} = \frac{R_{u,v}}{\sum_{i=1}^{n} R_{u,V_i} + \sum_{i=1}^{n} R_v, V_i}, (u, v, V_i \in V) \tag{36}$$

where $R_{u,V_i}$ represents the connection time between the users $u$ and $V_i$, $R_{u,v}$ denotes the connection time between the users $u$ and $v$, $V$ denotes all the users and $n$ denotes the size of $V$. The topic activation probability $P^{t}_{ij}$ is calculated by combining the user intimacy $C_{u,v}$ and the topic popularity $TP^{t}_{u,v}$, as shown below:

$$P^{t}_{ij} = C_{u,v} \times TP^{t}_{u,v}, (P^{t}_{ij} \in [0, 1]). \tag{37}$$

Lu *et al.* [78] believe that a user's ability to spread information on the social network and therefore become influential is dependent on that information being shared or forwarded by other users. However, they also believe that influence is tied to topics and therefore varies depending on the topics of interest. To this end, they propose two influence measurement metrics namely *Topical User Intimacy* which measures the possibility that a user will forward some information on a given topic and *Social Circle Difference* which is a measure of the scope to which information posted by one user may be subsequently spread. These two metrics are put together in order to measure the global user influence thus:

$$w_{uv,k} = [\lambda \cdot w^{e}_{uv,k} + (1 - \lambda) \cdot w^{i}_{uv,k}] \cdot SocialDiv_{uv} \tag{38}$$

where $w_{uv,k}$ is the topical influence of user $v$ over user $u$ on topic $k$. $w^{e}_{uv,k}$ and $w^{i}_{uv,k}$ denote the explicit influence and implicit influence of user $v$ to user $u$ on topic $k$ respectively while $SocialDiv_{uv}$ is the social circle difference. According to this work, the explicit influence between two users informed by the topic distribution of the forwarded content, the forwarding scope and forwarding frequency. On the other hand, implicit influence occurs when there is similarity in their topic interests and activity times.

While formalizing a relationship between content producers and content readers, Herzig *et al.* recognize that every user can play the dual role of an author and a reader or both. In this case, relationships are created through citation, i.e during interactions, readers make reference to contents generated by some other readers and vice versa. And so the nodes in the graph represent users, who can be authors, readers or both. An edge $e_{uv}$ shows that user $u$ has cited content posted by user $v$. According to this model, the influence of a user is measured by the ability to generate relevant and unique content that can be exposed to as many readers as possible both directly and indirectly.

Modeling of node to node relationships based on opinions has also been popular. In doing this, most authors recognize the evolving nature of opinions over a period of time. Xu *et al.* [79] model user influence based on the evolving nature of opinion dynamics. To model this influence relationship, they define an *opinion distance*, a measure that estimates how easily a node is influenced by its neighbor. The smaller the opinion distance between node $i$ and node $j$, the higher the influence that node $i$ has over node $j$. The opinion distance at a time $t$ is expressed as $d^{t}_{ij} = |x_i(t) - x_j(t)|$ where

$x_i(t)$ and $x_j(t)$ are the opinions of node $i$ and node $j$ at time $t$ respectively. The opinion influence of node $i$ over node $j$ forms the edge weight between the two nodes at time $t$ and is therefore expressed as:

$$w_{ij}^{t+1} = \frac{w_{ij}^t}{w_{ii}^t + \sum_{j=1}^{n} w_{ij}^t} \times 1 \tag{39}$$

where $w_{ij}^t = \frac{1}{d_{ij}^t}$ and $w_{ii}^t$ represents node $i's$ own opinion.

Liang *et al.* [80] however introduce a new dimension to opinion based influenced. They argued that there are two opinion types associated with each node on the network namely expressed opinion and innate opinion. An expressed opinion is an opinion that a node expresses to its neighbors in an attempt to influence them while an innate opinion is an opinion that a node holds within itself. The expressed opinion of a user is $u$ can be represented as follows:

$$y_u = (1 - \alpha_u) \cdot z_u + \alpha_u \cdot \sum_{v \in \cup 0 \leq i \leq (t-1)^{A_i}} y_v \cdot w_{v,u} \cdot r_{v,u} \tag{40}$$

where $w_{v,u}$ is the influence edge weight, $r_{v,u}$ is the expressed opinion from neighbor $v$ to $u$ and $z_u$ is the innate opinion. The factor $\alpha_u \in [0, 1]$ indicates how much $u$ is influenced by its neighbors compared to how it is influenced by its own opinion. A large value of $\alpha_u$ shows that user $u$ is influenced more by opinions from its neighbors rather than by its own opinions and vice versa.

There are other works that have considered different opinion aspects including opinion uncertainty [81], selective exposure to information when presented with several opinion sources [82] and PageRank algorithm with opinion component [83].

In general, a recurring concern for topic based influence computation is the fact that most of the models developed are text based. There are not many works that have investigated pictures and videos that otherwise contain richer topical content [84].

## 4.4 Similarity of User Activity Times

Similarity of user activity is a graph abstraction idea in which relationship ties are formed based on whether users engage on social interaction on the network at the same time or otherwise. This similarity points to interest in the same activities at similar times and possible homophylic tendencies.

In [78], the authors believe that if two users $u$ and $v$ are active at the same time, they are more likely to read each other's post and therefore influence each other. Activity time similarity therefore is a major metric that points at the existence of a relationship between two users in the network. To measure this similarity, they use the cosine similarity for the comparison of the activity times with $AT_u$ and $AT_v$ denoting the activity time distribution of users $u$ and $v$ respectively.

$$ActivitySim_{uv} = sim(AT_u, AT_v) = \frac{AT_u . AT_v}{\parallel AT_u \parallel . \parallel AT_v \parallel} \tag{41}$$

The key thing to address in this metric is the possibility of coordinated spamming actions that may appear as users engaging at the same time.

## 4.5 User Interaction Approaches

The cases reviewed under this section are those that are used to create the social graph based on user interactions irrespective of the intensity or the frequency of such interactions among users. For example, if a user $u$ replies to a tweet by another user $v$ then a link relationship is created on the graph between the two users. However, the strength of that relationship does not increase nor

decrease irrespective of how many more replies or less user $u$ gives to user $v$. In the same way, the representation of each user as a node on the graph is independent of any individual attributes that may be associated with such a user. It therefore means that the edge weights throughout the graph are assumed to be the same. This representation approach is what has been adopted by most centrality based metrics of social influence.
In [85] node indegree and outdegree values are used to represent starters and followers. Starters being bloggers who generate content and followers are the users that comment on and link to posts generated by other users. The authors define the *degree* $deg_G(n_p)$ of a node $n_p$ in graph $G$ as the difference between its indegree and outdegree, that is:

$$deg_G(n_p) = inDeg_G(n_p) - outDeg_G(n_p) \tag{42}$$

Agarwal *et al.* [86] propose a visualization of an influence graph (also called *i-graph*) based on the idea of influence flow of a blog post among the nodes in a graph. To do this this, they model each node as a single blog post characterized by four properties namely incoming influence $\iota$, outgoing influence $\theta$, number of comments $\gamma$ and length of the post $\lambda$. The influence flow is then calculated as:

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n) \tag{43}$$

where $w_{in}$ and $w_{out}$ are the weights that can be used to adjust the contribution of incoming and outgoing influence respectively. $p_m$ denotes the blog posts that link to the blog post while $p_n$ refers other blog posts referred to by this blog post. The *influenceFlow* is a metric that measures the difference between the total incoming influence of all inlinks and the total outgoing influence of all outlinks of the blog post $p$. Based on this abstraction, the Influence attributed to a blog post is then computed as:

$$I(p) = w(\lambda) \times (w_{com}\gamma_p + InfluenceFlow(p)) \tag{44}$$

where $W_{com}\gamma_p$ is a weight to regulate the number of comments $\gamma_p$.

In [87], node link relationships are used to define a dominating set. A dominating set is a collection of nodes that has a large positive influence over the network. Given a graph $G = (V, E)$, a node $v$ is influenced by a set $D$ if at least $\left\lceil \frac{deg(v)}{2} \right\rceil$ of its neighbors is in the influence dominating set D. A set $D$ is called a dominating set of $G$ if each node not in $D$ has at least one neighbor in $D$. Therefore, a set $P$ is called an influenced set of $D$ if each $v \in P$ is influenced by $D$. These relationships are used to define Time bounded Positive Influence Dominating sets which have major applications in controlling the spread of negative publicity in social networks [88]. Although this metric is good for representing positive influence on the network, it would be interesting to have a way of abstracting both the positive and the negative relations at the same time.

## 4.6 User Behaviour Evolution

A lot of research on influence has been dedicated to static networks [28]. However there are not as many works dedicated to the analysis of influence on dynamic networks. The very nature of real social networks is such that the network evolves thereby affecting its structure and content over time [89, 90]. Dynamic analysis of influence seeks to track such changes as the network continues to evolve [28]. Infact, [21] argues that since the influence strength between two nodes $u$ and $v$ varies over time, introducing time variable $t$ may lead to accurate descriptions of the influence strength between the

two nodes. An evolving directed social network at time $t$ is defined as a graph $G^t = (V^t, E^t)$, where $V^t$ is a node set and $E^t$ is a set of edges consisting of every pair of node in $G^t$ at time $t$. Yang *et al.* [28] observes that incrementally tracking a set of influential vertices in a dynamic network is a key problem.

Graph formation approaches provide ways of building the social graph either from simulations of social network data or from real social network data. Basically, the building of the social graph has two parts: The representation of the interacting objects within the social network - usually as nodes - and the abstraction of the relationships between them including the temporal evolution associated with that evolution. These two tasks are what put together a social graph.

# 5 Social Network Data

## 5.1 Social Network Data Collection

While a lot of research work on social influence is relying on social network data, little work in the context of social network has been dedicated to addressing issues that surround social network data acquisition and usage. When the data is intended for purposes involving semantic examination like topic, opinion, or semantic analysis then care should be taken to clean the data before being used for more reliable results. Social Network Data collection generally refers to the process of collecting hyperlinks from web pages associated with seed URLs from various servers[91]. Collection of online social data is not exactly easy since the collection process tends to face some problems such as the data being unstructured, heterogeneous, dynamic and bulky [92]. Additional challenges are the fragmented nature of the data (dispersed), frequent changes in the data format, absence of universal software interfaces for crawling and emerging legal hurdles regarding social network data usage [93]. Since social network users communicate freely by expressing their opinions, the data collected is not always without some unwanted part. Jiang *et al.* [94] propose a framework for using social network data. The steps include representation of social users, discovery of frequently connected friends and tracking of friendships or friend recommendations.

## 5.2 Attributes of Social Network Data

Depending on the purpose for which the data has been collected, the target data has four attributes which are *content, structure, usage and user profile*. Content describes the form in which the data is available such as text, images, or video. The structure of the data is the technology of representation like HTML or XML, usage is the purpose for which the data is being collected and user profile represents the demographic information about the users of the web services. However, Vutrapu *et al.* [95] suggest that any purpose for which social data is being collected must support both the conceptual and mathematical modeling of such data through software. They provide two perspectives to the use of social data namely notation - which is largely graph abstraction and operational semantics, which is about modeling of social interactions.

## 5.3 Social Network Data PreProcessing

Given the challenges associated with social network data, it is always necessary that the collected data undergoes a cleaning process to make it suitable for the role for which it has been collected. This is why the data needs to undergo preprocessing. Sharma *et al.* [91] define Data preprocessing

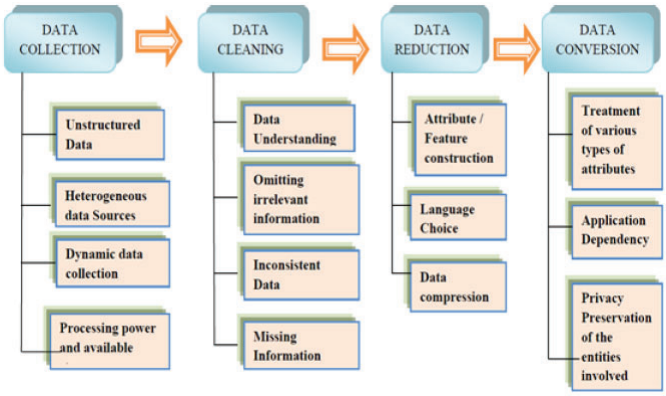


Figure 3: Social Network Data PreProcessing Phases [92]

as a process that represents data in a format that is best suitable for mining purposes. In particular, when the data in question is intended for use in text related analysis such as influence analysis through through topic, opinion or sentiment examination, it becomes crucial to clean such text to ensure its suitability for use. One of the ways through which the suitability of data is ensured is data preprocessing. As outlined by Gupta *et al* [92], the data preprocessing task has four phases namely *data collection, data cleaning, data reduction* and *data conversion*. For a detailed explanation on each of these phases, the reader can refer to [92]. Preprocessing of social network data is an important stage in experiments that rely on social network data. This step is meant to clean the data and make it suitable for a specific experimental activity. Fig 3 shows an overview of the phases for the preparation of social network data for use.

Although the generic framework for preprocessing social network data suffices for most cases, the heterogeneous nature of social network data still makes it difficult to achieve adequate cleaning for all various sources of social network data. this is because social media applications serve different needs and as a result attract different forms of data.

# 6 Conclusion and Recommendations

In this survey, we have provided a summarized categorization of various approaches that have been adopted by researchers for building the social graph from social network data and analyzing the social graph to determine the most influential members of a social network. We have reviewed specific metrics and techniques in each of the phases of influence analysis namely social graph abstraction and social graph analysis.

The review reveals that a majority of variables used to define user relationships in the social network come from their mode of association such as social actions, similar topics, opinions, topological links and frequency of interactions. These, and others form a basis for formalizing relationships among network members and hence are used to provide frameworks for influence computation. Furthermore, it is clear that the metrics that are used to compute influence on the social graph must have their basis on the theoretical definition of influence for the identified context. In this way, it is easier to show the reliability of the used metrics.

Social influence analysis cannot be performed without social network data. for this we have provided a brief on the sources of such data. A majority of the reviewed literature reveal that

social network data suitability may vary from one work to the other. However, such data undergoes preprocessing to fit the role for which it has been collected.

Finally, in each of the identified phases of social influence study, we have identified challenges associated with the techniques adopted and we have suggested, through appropriate citations, how some of the challenges have been addressed in literature. However, one of the major challenges that still stand out is the absence of a distributed approach to influence analysis. Since social networks are large in size, this is a major area that needs to be addressed in future studies since it will reduce challenges that come with insufficient computing power when analyzing influence on networks with enormous sizes. Other challenges include the absence of ground truth data and metrics for evaluation, differences in social data formats and varying scientific definitions for influence from author to author.

We hope that these challenges form a good basis for the formulation of new research ideas for future researchers in the field of Social Influence.

# References


[1] W. Yang, H. Wang, and Y. Yao, "An immunization strategy for social network worms based on network vertex influence," *China Communications*, vol. 12, no. 7, pp. 154–166, 2015.

[2] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 971–974, ACM, 2007.

[3] D. Li, X. Shuai, G. Sun, J. Tang, Y. Ding, and Z. Luo, "Mining topic-level opinion influence in microblog," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1562–1566, 2012.

[4] X. Tang and C. C. Yang, "Ranking user influence in healthcare social media," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, pp. 1–21, 2012.

[5] C.-T. Lu, H.-H. Shuai, and P. S. Yu, "Identifying your customers in social networks," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 391–400, 2014.

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

[7] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: A survey," *Journal of Network and Computer Applications*, vol. 106, pp. 17–32, 2018.

[8] K. Almgren and J. Lee, "Who influences whom: Content-based approach for predicting influential users in social networks," in *International conference on advances in big data analytics*, pp. 89–99, 2015.

[9] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *arXiv preprint arXiv:1808.05502*, 2018.

[10] F. Riquelme and P. González-Cantergiani, "Measuring user influence on twitter: A survey," *Information Processing and Management*, vol. 52, no. 5, pp. 949–975, 2016.

[11] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65–74, 2011.

[12] M. Kardara, G. Papadakis, A. Papaoikonomou, K. Tserpes, and T. Varvarigou, "Large-scale evaluation framework for local influence theories in twitter," *Information processing and management*, vol. 51, no. 1, pp. 226–252, 2015.

[13] J. Sun and J. Tang, "A survey of models and algorithms for social influence analysis," in *Social network data analytics*, pp. 177–214, Springer, 2011.

[14] M. S. Granovetter, "The strength of weak ties," in *Social networks*, pp. 347–367, Elsevier, 1977.

[15] S. M. H. Bamakan, I. Nurgaliev, and Q. Qu, "Opinion leader detection: A methodological review," *Expert Systems with Applications*, vol. 115, pp. 200–222, 2019.

[16] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, "Ranking in evolving complex networks," *Physics Reports*, vol. 689, pp. 1–54, 2017.

[17] B. Li, Z. Gao, X. Shan, W. Zhou, and E. Ferrara, "Sorec: A social-relation based centrality measure in mobile social networks," *arXiv preprint arXiv:1902.09489*, 2019.

[18] N. Arazkhani, M. R. Meybodi, and A. Rezvanian, "Influence blocking maximization in social network using centrality measures," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pp. 492–497, IEEE, 2019.

[19] L. Yang, Y. Qiao, Z. Liu, J. Ma, and X. Li, "Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm," *Soft Computing*, vol. 22, no. 2, pp. 453–464, 2018.

[20] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Physics Reports*, vol. 650, pp. 1–63, 2016.

[21] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, "Modeling information diffusion over social networks for temporal dynamic prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1985–1997, 2017.

[22] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[23] K. Avrachenkov, N. Litvak, and K. S. Pham, "A singular perturbation approach for choosing the pagerank damping factor," *Internet Mathematics*, vol. 5, no. 1-2, pp. 47–69, 2008.

[24] H. Zhou, D. Zeng, and C. Zhang, "Finding leaders from opinion networks," in *2009 IEEE International Conference on Intelligence and Security Informatics*, pp. 266–268, IEEE, 2009.

[25] B. Huang, G. Yu, and H. R. Karimi, "The finding and dynamic detection of opinion leaders in social network," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[26] C. Wang, Y. J. Du, and M. W. Tang, "Opinion leader mining algorithm in microblog platform based on topic similarity," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pp. 160–165, IEEE, 2016.

[27] M. Azaouzi and L. B. Romdhane, "An efficient two-phase model for computing influential nodes in social networks using social actions," *Journal of Computer Science and Technology*, vol. 33, no. 2, pp. 286–304, 2018.

[28] Y. Yang and J. Pei, "Influence analysis in evolving networks: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[29] A. V. Sathanur, V. Jandhyala, and C. Xing, "Physense: Scalable sociological interaction models for influence estimation on online social networks," in *2013 IEEE International Conference on Intelligence and Security Informatics*, pp. 358–363, IEEE, 2013.

[30] Y. Zhang, J. Mo, and T. He, "User influence analysis on micro blog," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 3, pp. 1474–1478, IEEE, 2012.

[31] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PloS one*, vol. 6, no. 6, p. e21202, 2011.

[32] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: the role of clustering," *PloS one*, vol. 8, no. 10, 2013.

[33] R. Nagmoti, A. Teredesai, and M. De Cock, "Ranking approaches for microblog search," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 153–157, IEEE, 2010.

[34] D. Gayo-Avello, "Nepotistic relationships in twitter and their impact on rank prestige algorithms," *Information Processing and Management*, vol. 49, no. 6, pp. 1250–1280, 2013.

[35] X. Li, S. Cheng, W. Chen, and F. Jiang, "Novel user influence measurement based on user interaction in microblog," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 615–619, ACM, 2013.

[36] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Zareie, "Identification of influential users by neighbors in online social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 486, pp. 517–534, 2017.

[37] Z. Yin and Y. Zhang, "Measuring pair-wise social influence in microblog," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 502–507, IEEE, 2012.

[38] H. Yulan and L. Ling, "Analysis of user influence in social network based on behavior and relationship," in *Proceedings of 2013 2nd International Conference on Measurement, Information and Control*, vol. 1, pp. 682–686, IEEE, 2013.

[39] Z. Yang, X. Huang, J. Xiu, and C. Liu, "Socialrank: Social network influence ranking method," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 2, pp. 591–595, IEEE, 2012.

[40] J. Zhao, S. Shang, P. Wang, J. C. Lui, and X. Zhang, "Tracking influential nodes in time-decaying dynamic interaction networks," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1106–1117, IEEE, 2019.

[41] A. Zhiyuli, X. Liang, Y. Chen, and X. Du, "Modeling large-scale dynamic social networks via node embeddings," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[42] A. Silva, S. Guimarães, W. Meira Jr, and M. Zaki, "Profilerank: finding relevant content and influential users based on information diffusion," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, pp. 1–9, 2013.

[43] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[44] R. M. Anderson, B. Anderson, and R. M. May, *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[45] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[46] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method," *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 278–288, 2017.

[47] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo, *Diffusion in social networks*. Springer, 2015.

[48] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda, "Learning diffusion probability based on node attributes in social networks," in *International Symposium on Methodologies for Intelligent Systems*, pp. 153–162, Springer, 2011.

[49] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *International conference on knowledge-based and intelligent information and engineering systems*, pp. 67–75, Springer, 2008.

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

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

[52] Y. Lin, X. Zhang, L. Xia, Y. Ren, and W. Li, "A hybrid algorithm for influence maximization of social networks," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 427–431, IEEE, 2019.

[53] H. Wu, J. Shang, S. Zhou, Y. Feng, B. Qiang, and W. Xie, "Laim: A linear time iterative approach for efficient influence maximization in large-scale networks," *IEEE Access*, vol. 6, pp. 44221–44234, 2018.

[54] T. W. Valente, "Social network thresholds in the diffusion of innovations," *Social networks*, vol. 18, no. 1, pp. 69–89, 1996.

[55] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *European conference on principles of data mining and knowledge discovery*, pp. 259–271, Springer, 2006.

[56] A. Yadav, H. Chan, A. Xin Jiang, H. Xu, E. Rice, and M. Tambe, "Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty," in *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pp. 740–748, International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[57] A. Yadav, B. Wilder, E. Rice, R. Petering, J. Craddock, A. Yoshioka-Maxwell, M. Hemler, L. Onasch-Vera, M. Tambe, and D. Woo, "Influence maximization in the field: The arduous journey from emerging to deployed application," in *Proceedings of the 16th conference on autonomous agents and multiagent systems*, pp. 150–158, International Foundation for Autonomous Agents and Multiagent Systems, 2017.

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

[59] Y. Zhao, S. Li, and F. Jin, "Identification of influential nodes in social networks with community structure based on label propagation," *Neurocomputing*, vol. 210, pp. 34–44, 2016.

[60] N. Sumith, B. Annappa, and S. Bhattacharya, "A holistic approach to influence maximization in social networks: Storie," *Applied Soft Computing*, vol. 66, pp. 533–547, 2018.

[61] Q. Liqing, Y. Jinfeng, F. Xin, J. Wei, and G. Wenwen, "Analysis of influence maximization in temporal social networks," *IEEE Access*, vol. 7, pp. 42052–42062, 2019.

[62] J. Li, T. Cai, A. Mian, R.-H. Li, T. Sellis, and J. X. Yu, "Holistic influence maximization for targeted advertisements in spatial social networks," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 1340–1343, IEEE, 2018.

[63] W. Yang, L. Brenner, and A. Giua, "Influence maximization by link activation in social networks," in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1, pp. 1248–1251, IEEE, 2018.

[64] F.-H. Li, C.-T. Li, and M.-K. Shan, "Labeled influence maximization in social networks for target marketing," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 560–563, IEEE, 2011.

[65] B. Liu, G. Cong, Y. Zeng, D. Xu, and Y. M. Chee, "Influence spreading path and its application to the time constrained social influence maximization problem and beyond," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1904–1917, 2013.

[66] S. Xu, N. Xu, J. Zhang, F. Li, and S. Li, "Seed set selection in evolving social networks," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 2323–2328, IEEE, 2017.

[67] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 199–208, ACM, 2009.

[68] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[69] X. Wang, O. Liu, *et al.*, "A new model to measure the knowledge diffusion via information entropy in virtual communities," in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–3, IEEE, 2015.

[70] Q. Ma, X. Luo, and Y. Luo, "Information entropy based the stability measure of user behaviour network in microblog," in *2014 10th International Conference on Semantics, Knowledge and Grids*, pp. 67–74, IEEE, 2014.

[71] Y. Yang, L. Zhou, Z. Jin, and J. Yang, "Meta path-based information entropy for modeling social influence in heterogeneous information networks," in *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pp. 557–562, IEEE, 2019.

[72] S. Peng, A. Yang, L. Cao, S. Yu, and D. Xie, "Social influence modeling using information theory in mobile social networks," *Information Sciences*, vol. 379, pp. 146–159, 2017.

[73] M. De Choudhury, "Tie formation on twitter: Homophily and structure of egocentric networks," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 465–470, IEEE, 2011.

[74] Y.-C. Chen, Y.-H. Chen, C.-H. Hsu, H.-J. You, J. Liu, and X. Huang, "Mining opinion leaders in big social network," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pp. 1012–1018, IEEE, 2017.

[75] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 236–242, ACM, 2013.

[76] L.-L. Shi, L. Liu, Y. Wu, L. Jiang, J. Panneerselvam, and R. Crole, "A social sensing model for event detection and user influence discovering in social media data streams," *IEEE Transactions on Computational Social Systems*, 2019.

[77] Y. Li, C. Jia, and J. Yu, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 321–334, 2015.

[78] M. Lu, Z. Wang, and D. Ye, "Topic influence analysis based on user intimacy and social circle difference," *IEEE Access*, vol. 7, pp. 101665–101680, 2019.

[79] H. Xu, W. Cai, and G. Chen, "Opinion evolution model based on the node influence on the internet," in *2015 2nd International Conference on Information Science and Control Engineering*, pp. 359–362, IEEE, 2015.

[80] W. Liang, C. Shen, X. Li, R. Nishide, I. Piumarta, and H. Takada, "Influence maximization in signed social networks with opinion formation," *IEEE Access*, vol. 7, pp. 68837–68852, 2019.

[81] S. Malinchik, "Framework for modeling opinion dynamics influenced by targeted messages," in *2010 IEEE Second International Conference on Social Computing*, pp. 697–700, IEEE, 2010.

[82] R. Das, J. Kamruzzaman, and G. Karmakar, "Opinion formation in online social networks: Exploiting predisposition, interaction, and credibility," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 554–566, 2019.

[83] A. N. Ngaffo, W. El Ayeb, and Z. Choukair, "Mining user opinion influences on twitter social network: Find that friend who leads your opinion using bayesian method and a new emotional pagerank algorithm," in *2019 15th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 680–685, IEEE, 2019.

[84] M. Kretschmer, B. Göschlberger, and R. Klamma, "Using topical networks to detect editor communities in wikipedias," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 102–109, IEEE, 2019.

[85] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th international conference on World Wide Web*, pp. 491–501, 2004.

[86] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 207–218, 2008.

[87] T. Shi, J. Wan, S. Cheng, Z. Cai, Y. Li, and J. Li, "Time-bounded positive influence in social networks," in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, pp. 134–139, IEEE, 2015.

[88] A. Dhawan and M. Rink, "Positive influence dominating set generation in social networks," in *2015 International Conference on Computing and Network Communications (CoCoNet)*, pp. 112–117, IEEE, 2015.

[89] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[90] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguná, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, p. 537, 2012.

[91] S. Sharma and A. Bhagat, "Data preprocessing algorithm for web structure mining," in *2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS)*, pp. 94–98, IEEE, 2016.

[92] P. Gupta and V. Bhatnagar, "Data preprocessing for dynamic social network analysis," in *Data Mining in Dynamic Social Networks and Fuzzy Systems*, pp. 25–39, IGI Global, 2013.

[93] L. Rudikowa, O. Myslivec, S. Sobolevsky, A. Nenko, and I. Savenkov, "The development of a data collection and analysis system based on social network users' data," *Procedia Computer Science*, vol. 156, pp. 194–203, 2019.

[94] F. Jiang, C. K. Leung, and A. G. Pazdor, "Big data mining of social networks for friend recommendation," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 921–922, IEEE, 2016.

[95] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics," *Ieee Access*, vol. 4, pp. 2542–2571, 2016.