Expert Systems With Applications 115 (2019) 200–222

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Opinion leader detection: A methodological review

Seyed Mojtaba Hosseini Bamakan[a], Ildar Nurgaliev[a,b], Qiang Qu[a,*]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[b] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 16 April 2018
Revised 8 June 2018
Accepted 30 July 2018
Available online 31 July 2018

Keywords:
Opinion leader
Social network analysis
Flow of influence
Influential users/nodes
Graph mining

ABSTRACT

A social network as an essential communication platform facilitates the interactions of online users. Based on the interactions, users can influence or be affected by the opinions of others. The users being able to influence and shape the opinions of others are considered as opinion leaders. The problem of identifying opinion leaders is an important task due to its wide applications in reality, including product adoption for marketing and societal analytics. The problem has been attracting proliferating studies over the recent years. To overview and provide insights of the methodologies and enlighten the future study, we review the well-known techniques for opinion leader detection problems. These techniques are classified into descriptive approaches, statistical and stochastic methods, diffusion process based approaches, topological based methods, data mining and learning methods, and approaches based on hybrid content mining. The advantages and drawbacks of each method are systematically analyzed and compared, to provide deep understanding into the existing research challenges and the direction of future trends. The findings of this review would be useful for those researchers are interested in identifying opinion leaders and influencers in social networks and related fields.

1. Introduction

We are living in an era that the ways of people communicate with each other have been changed dramatically because of the advent and expansion of social media sites. By propagation of social networking sites, opinion-sharing websites, blogs, and microblogs people can easily and freely interact and express their personal experiences, opinions, emotions, and feelings regard to a specific product, service or even in a political sphere and economic issues (Arrami, Oueslati, & Akaichi, 2017; Kostkova, Mano, Larson, & Schulz, 2017; Song, Cho, & Kim, 2017). In such an environment that information flows smoothly, some users/persons have a high capacity to influence the others opinion or lead them toward a particular topic because of their experience, mentality, pursue goals or probably because of their charismatic personality to provoke the emotions of their followers.

Nowadays, social networks provide an essential communication platform that facilitates the interactions of people in the society. Based on these interactions, individuals will be affected by others opinions and also can influence on them. Among them, who have the ability to influence and shape the opinions of other people are considered as opinion leaders or referred as thought leaders. The task of identifying these non-ordinary and influential individuals is defined as opinion leaders detection (OLD) (Potolea, 2016).

The significance of identifying opinion leaders may become clearer by revealing their powerful influence on changing and shaping a trend in business and marketing, to promote and demote products, to lead a politic stream and to raise the awareness of society on a public health issue or environmental problem. Here, the question is that who these opinion leaders are and how we can detect them in a social network. In the last few years, there has been a growing effort to propose some persuasive and precise solutions to this problem. Hence, there is a need for a better understanding of constraints in existing approaches for identifying and modeling opinion leaders to ensure a possible future informations spreading pattern. In the following, the main contributions of this paper are mentioned.

1. A comprehensive and systematic survey with a critical perspective on the state-of-the-art approaches for opinion leader detection is presented.
2. A deep insight into the existing research challenges is presented and future directions in identifying opinion leaders are discussed.

The main objective of this methodological review is to provide a deep insight into the methods of analysis and techniques applied to identify opinion leaders with the focus on the social network.

* Corresponding author.
E-mail addresses: hosseini@siat.ac.cn (S.M.H. Bamakan), ildar@siat.ac.cn (I. Nurgaliev), qiang@siat.ac.cn (Q. Qu).

https://doi.org/10.1016/j.eswa.2018.07.069
0957-4174/

Meanwhile, by comprehensive considering the existing literature, we provide a framework of understanding at different levels such as theory and concept of opinion leaders, the domain of applicability, research approaches, data collection and analysis methods and techniques.

The research methodology of this survey is to identify literature related to "opinion leader" consists of a systematic search in the journal and conference papers. For this purpose, we first identify a wide range of search terms including "opinion leader, influential leader/users, influencer, influential node, dominant node, influence power". In addition, we apply synonyms and alternative keywords with adding AND/OR Boolean expressions to find more related papers in various academic databases and search engines such as Science Direct, SpringerLink, IEEE Xplore, arXiv, ACM digital library, and Google scholar. The next step is to assess the quality of the collected papers and decide whether they are within the scope of our survey or not. Finally, based on critical evaluation of collected literature and the properties of their methods, we categorize the opinion leader detection techniques into six groups.

The remainder of the paper is organized as follows. Section 2 defines the opinion leader and the possible application domains. In Section 3, we categorize and compare the approaches are proposed to deal with the opinion leader detection problem. The public datasets for opinion leader detection are presented in Section 4. The strengths and limitations of these methods and the possible challenges are discussed in Section 5. Finally, we conclude this review and provide some future research directions in Section 6.

2. Background

2.1. Opinion leader

The role of leaders in constructing and understanding the group members opinions can be traced back to some works in the 1950s (Chowdhry & Newcomb, 1952; Katz, 1957). Chowdhry and Newcomb (1952) believed that individuals are not only selected for the position of leadership based on their personal qualities but also they should have some characteristics which match with a particular situation of the group and the set of interests shared by its members. Based on their hypothesis, a leader of a group has substantial ability to affect the group opinions on some familiar and relevant issues compare to the other group members, but there is no such a result for relatively non-familiar and non-relevant issues (Chowdhry & Newcomb, 1952). Based on the hypothesis of "the two-step flow of communication" (Katz, 1957; Lazarsfeld, Berelson, & Gaudet, 1944), the flow of opinion and information is transferred from mass media to the public audience is transferred through the mediating role of opinion leaders. Katz (1957) supposed the influence as a function of three elements, including personification of certain values, competence, strategic social location. The findings of his research also support the results of Chowdhry and Newcomb (1952) that the role of influentials and influencees is possible to change during the time and in different domains (Potolea, 2016).

In different situations and various contexts of research domains, a number of definitions have been proposed to describe the opinion leaders. Lazarsfeld et al. (1944) defined the opinion leader as an attractive person with outstanding features in his/her psychological, physical and social aspects which has a credible knowledge in a given domain. By Rogers (2010) an opinion leader is a participant with higher socioeconomic status compared to his/her followers because of great exposure to mass media and maintains a strong contact with change agents, which makes him/her an influential social participant. The common characteristics which are mentioned for an opinion are usually specified as, likable, trustworthy, educationally influential, self-confidence (Chan & Misra, 1990; Flodgren et al., 2011).

Some authors discriminate between influencer and opinion leader, in such a way that influencer is considered as a broader concept includes who has a potential ability to influence others such as celebrities, a family member or a friend in a friendship group (Arrami et al., 2017), but from our point of view, opinion leader is narrow and deep term which frequently is considered by the context of the research areas. By the way, by considering influencer in a certain domain, both concepts have the same meaning. As an example, influencer marketing is defined as a kind of marketing in which the marketing activities are arranged around the influential people instead of targeting a market as a whole. In summary, an influencer can derive his value from some sources include social reach, original content and consumer trust. In marketing studies the value of influencer is measured in multiple ways, such as Earned Media Value (EMV), Cost Per Action (CPA) and tracking the obtained impressions.

In general, opinion leaders based on their characteristics and the structure of social network are categorized as shown in Fig. 1.

1. *Local vs. global opinion leader*; based on the influence scope of the opinion leaders in their communities, they can be grouped as local and global leaders. From the marketing perspective, the global opinion leaders have a greater ability to exert the global influence on the multiple international markets, unlike the local market influencers. According to the network structures, a global opinion leader has an access to a mature and dense online community, comparable to a local opinion leader, which mostly engage with some disparate or fragmented communities. Furthermore, because of the critical role of global opinion leaders in big networks, they have the highest chance of controlling the information flow between the large numbers of followers.
2. *Monomorphic vs. polymorphic opinion leader*; this kind of classification is mostly considered in the context of marketing and advertising. Here, monomorphic opinion leaders are those with specialty and considerable knowledge within single-topic areas, whereas the polymorphic opinion leaders are considered across multiple-topic areas, in fact, based on their access to the mass media tend to disseminate information in a broad range of domains (Richmond, 1980; Rogers & Shoemaker, 1971). Rogers and Shoemaker (1971) believe that in a traditional society, opinion leaders tend to be more polymorphic, controversy these individuals are considered as the monomorphic opinion leaders in a modem society.
3. *Positive vs. destructive opinion leader*; by reviewing the literature of opinion leader, it is clearly obvious that focuses mostly were on the positive side of leadership. However, opinion leaders can be destructive as much as they are positive (Einarsen, Aasland, & Skogstad, 2007; Krasikova, Green, & LeBreton, 2013; Padilla, Hogan, & Kaiser, 2007) and can behave in manipulation, persuasion, destructive manner. Padilla et al. (2007) did a comprehensive research on defining the destructive leaders. From their point of view, a destructive leader has a selfish personality, tends to use control and coercion instead of persuasion and commitment, and deflects the follower from their main goals and directions. Totally, Padilla et al. (2007) discussed three main elements of destructive leadership, including the destructive leaders, susceptible followers, and conducive environments.
4. *Long-term vs. short-term opinion leader*; long-term or short-term influence of opinion leaders on their community can be viewed from the perspective of opinion dynamics theory (Zhao, Kou, Peng, & Chen, 2018). Based on this theory, the opinions, beliefs, and judgments of individuals are shaped based on their

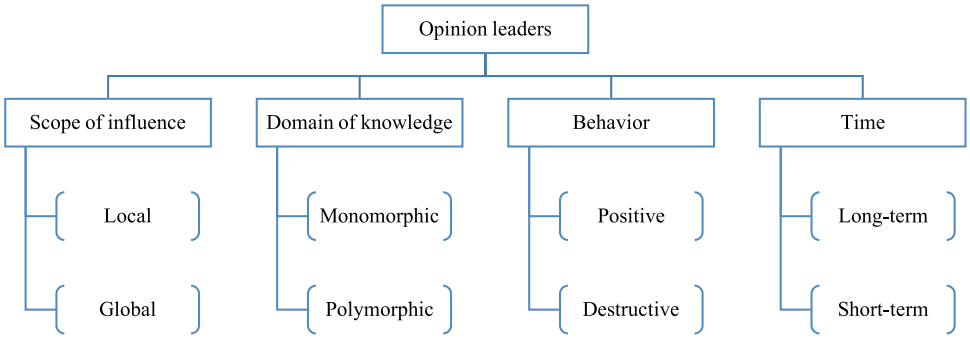


Fig. 1. Opinion leaders categorization.

accessibility to the information and the opinions of their influential neighbors. However, the influence power of opinion leaders during the formation of followers opinions is varying because of dependency of this process to the level of users trust and confidence toward their opinion leaders (Das, Gollapudi, & Munagala, 2014).

2.2. Domain of applicability

Since the opinion leaders can facilitate the dissemination of information more effective throughout their networks, opinion leader detection has many applications in different domains. In the following, we mentioned some of them briefly.

Business and marketing; by observing these influential individuals from the marketing perspective, there are increasing interests from companies to find such users who can increase their profits from advertising, marketing, and new product adoption by attracting potential customers. Thus, opinion leaders are considered as an absorbent goal for marketing, advertising and brand evaluation (Grissa, 2016; Iyengar, Van den Bulte, Eichert, & West, 2011; Mohammadi & Andalib, 2017; Tejavibulya & Eiamkanchanalai, 2011).

Political science; in the context of politic, opinion leaders are the key individuals or organizations with a prominent status in their communities which consolidate, change and propagate a specific political trend to the target population in society in order to promote social harmony and stability. There are some studies that explore the role of opinion leader in political science such as Rocha et al. (2016), Horio and Shedd (2016), Kingdon (1970) and Lees-Marshment (2012). For example, Rocha et al. (2016) is the study of understanding and modeling the opinion leaders regard to Brazilian political protests in 2015. Here, the authors assume that opinions and opinion leaders are two completely dependent concepts, so they proposed an opinion leader detection methodology based on a combination of sentiment analysis (SA) and influential users detection (IUD).

Public health; in this domain, opinion leader can help to promote evidence-based practice, speed up the diffusion of health promoting and disease preventing innovations (Flodgren et al., 2011; Gentina, Kilic, & Dancoine, 2017; Gotecha & Patwardhan, 2016). As an example, Guldbrandsson, Nordvik, and Bremberg (2012) is the case of identifying opinion leaders to promote the child health in Sweden. Researchers in this domain believe in the external influences of opinion leaders on the learning process of society and the ability of them to modify others behavior (Carpenter & Sherbino, 2010). Furthermore, opinion leaders play a vital role to influence on physicians and consequently on prescribing a specific behavior and treatment guidelines, also to identify the new healthcare research areas and to give notice about the benefits and side effects of some drugs (Kumar, 2015).

Sociology and psychology; the work of Cialdini (2001) is referenced as a good reference of social influence with the focus on compliance and persuasion in a social environment. He believed that in the modern society, people are overwhelmed with an intense flow of information which they cannot take it into account in their decision making process. Therefore, people mostly act with the limited amount of thought and time, which it provided a situation to be manipulated by who can influence on their opinions. Moreover, the work of Tucci, González-Avella, and Cosenza (2016), is an example of dynamic modeling for cultural interaction among social agents, which tried to find the influence of opinion leaders in the collective behavior of a social system.

Education; the impact of opinion leaders on the educational system is an important meanwhile interesting domain (Koeslag-Kreunen, Van der Klink, Van den Bossche, & Gijselaers, 2017). Here, opinion leaders mostly called "Teacher Leaders" that play a wide range of roles that may assign formally or shared informally to improve the quality of teaching, shape the thought of students, provide extra instructional resources, serve as a model for both students and new teachers and so on (Harrison & Killion, 2007). For example, the role of school leader to shape the procedure of reculturing is examined by Geijsel, Meijers, and Wardekker (2007) and opinion leaders network-level characteristics based on users behavior and topic contents in online learning communities are investigated by Li, Ma, Zhang, Huang et al. (2013).

3. Algorithms

This section presents the proposed approaches to deal with the opinion leader detection problem. Based on the characteristics of algorithms, we group them into six categories as shown in Fig. 2 including (I) descriptive approaches; (II) statistical and stochastic approaches; (III) diffusion process based approaches; (IV) topological based methods; (V) data mining and machine learning approaches, and finally (VI) hybrid content mining approaches. In each group, we present relevant methods and highlight the advantages of models and metrics employed. However, since opinion leader detection is considered as an interdisciplinary topic and it is an interesting subject for a wide range of audiences, we provide a complete yet simple overview of existing approaches. In Table 1, we present the mathematical preliminaries and notations are used throughout this review paper.

For the whole section, we suppose G as a graph which is denoted as a pair $G(V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$, is a set of nodes and $E = \{e_1, e_2, \ldots, e_M\}$ is a set of edges. In each graph, the entities are called nodes, vertices or actors, and the connections among the nodes are called edges or links.

3.1. Descriptive approaches

As mentioned in Section 2, an opinion leader plays significant roles such as a pioneer of innovations, an accelerator of behavior changes, an assistant for adoption of new social norms, a promoter

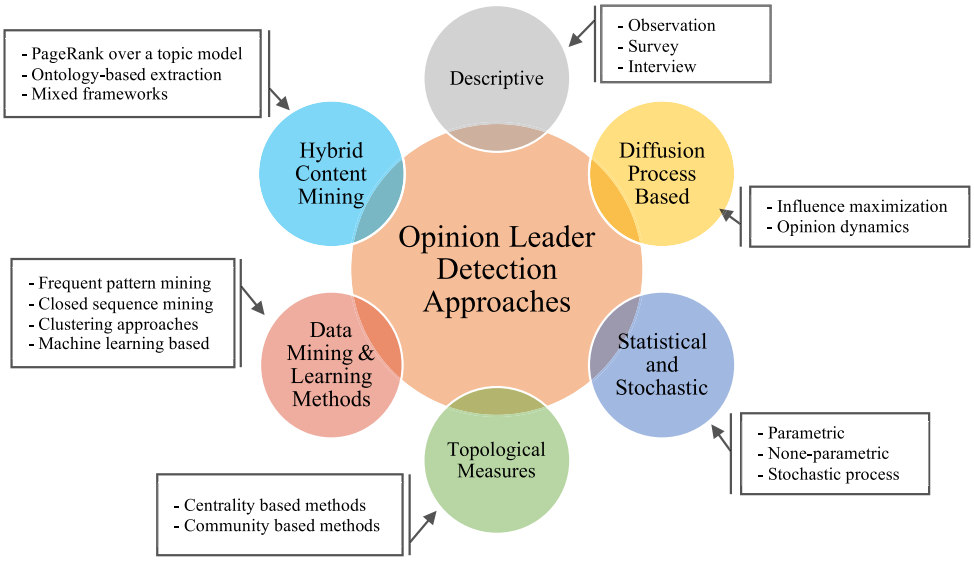


Fig. 2. Opinion leader detection approaches structure.

Table 1
Mathematical preliminaries and notations.

Symbol	Description	Symbol	Description
G	A graph of a social network	$c(v_i, v_j)$	Distance between nodes
V	Set of users in one graph	$\sigma_{i,j}$	Number of geodesic paths connecting v_i and v_j
E	Set of relationship between users	$\sigma_{j,j}(v_i)$	Number of geodesic paths including v_k
N	Number of vertices in G	k_j^{out}	The out-degree of node j
v_i	A node $v_i \in V$	k_{int}	Number of connections to other vertices
α	Tunable parameter [0,1]		Within the community
A	The adjacency matrix of a graph	x	Left-hand eigenvector of A
$a_{i,j}$	$a_{i,j} \in A$	λ	Eigenvalue
D	Diagonal matrix	k_{in_i}	Number of edges inside the community of v_i
$\mathcal{W}_k(v)$	Nearest and next $k-1$ neighbors of v	$N_k(v)$	$\lvert\mathcal{W}_k(v)\rvert$
$\overline{k_s}$	Average vertex degree in community s	c_v	Cluster coefficient

for the success of community-based efforts and so on (Valente & Pumpuang, 2007). Hence, this topic attracts the attention of many researchers to identify the influential individuals and takes the advantages of them. One of the primary approaches to identify the opinion leaders is the descriptive methods that frequently are utilized in the social and life sciences. The authors in Valente and Pumpuang (2007) conducted a complete survey by covering close to 200 articles in the field of sociology and distinguished ten descriptive techniques for opinion leader detection.

3.1.1. Observation, survey, interview

For simplicity purpose, we categorized the descriptive methods into three groups and the advantages and disadvantages of each group are discussed as follows.

1. *Observation methods*; this group of methods is based on the observation of an expert to the subject field which including selecting celebrities (Erdogan, Baker, & Tagg, 2001; Jin, 2017), self-selection (Aghdam & Navimipour, 2016; Gates & Kennedy, 1989), staff selected (Earp et al., 2002; Weenen, Pronker, Commandeur, & Claassen, 2013) and positional approach (Valente & Pumpuang, 2007; Weenen et al., 2013). Although these methods are easy to implement and have a high visibility and low cost, they suffer from some problems such as observation bias, tendency to select well-known members, uncertain abilities of the observer, and selecting an important but not relevant and motivated person as an opinion leader. Therefore, this group of methods is often applicable in small or local communities (Kumar, 2015).
2. *Survey methods*; in this group of methods, participants are asked to select the opinion leaders through a short questionnaire or to rate themselves and their peers based on some predefined measures. Methods such as self-identification (Childers, 1986; Katz, 2015), judges ratings, and expert identification can be fell in this group (Valente & Pumpuang, 2007). It should be noted that the cost and time of survey grow up by increasing the sample size. Other common problems are sampling bias and volunteer response bias, which avoiding these drawbacks depends on the proficiency of surveyors (Kumar, 2015).
3. *Interview methods*; which here means a person-to-person, either face-to-face interaction between two parties to collect information and many ways exist depending on defined aims. The methods of this group include snowball method (Van den Nieuwboer, Van De Burgwal, & Claassen, 2016; Pronker, Weenen, Commandeur, Claassen, & Osterhaus, 2015), sample sociometric (Costenbader & Valente, 2003), and sociometric which are based on asking participants about their opinions (Buchler, Rajivan, Marusich, Lightner, & Gonzalez, 2017; Schneider, Zhou, & Laumann, 2015). There are not only the same limitations of survey methods in this group, but also they are more expensive based on the time and cost. In fact, it would be extremely expensive to interview with all or most of community members in a big network of people.

3.1.2. Summary

In summary, although the descriptive methods summarized in (Table 2) are favorable to carry out in the social sciences because of their simplicity. These methods are increasingly ineffective in

Table 2
Descriptive methods.

Sub category	Key references
Observation methods	(Erdogan et al., 2001), (Jin, 2017), (Aghdam & Navimipour, 2016), (Earp et al., 2002), (Gates & Kennedy, 1989), (Weenen et al., 2013), (Valente & Pumpuang, 2007)
Survey methods	(Katz, 2015), (Childers, 1986)
Interview methods	(Van den Nieuwboer et al., 2016), (Pronker et al., 2015), (Costenbader & Valente, 2003), (Buchler et al., 2017), (Schneider et al., 2015)

the big communities and are favorably impressed with the risk of experience bias. Thereby, social network analysis as an alternative approach can provide rich contextual information such as the position of individuals within a network and the degree of importance of each leader in the community. In total, scalability is the most important property that provides the ability of considering the entire community rather than just a sample size (Kumar, 2015) to analyze the social networks.

3.2. Statistical and stochastic approaches

Based on the definition of opinion leaders, these people have different characteristics, which make them distinctive among the rest of the community. From this perspective, thought, opinion, and behavior of these non-ordinary individuals do not follow the patterns of the majority of society. Hence, the opinion leader detection can be formulated as an anomaly or exception detection problem. Probabilistic and statistical models are considered as the earliest methods for solving anomaly detection problem, which date back to the 19th century. According to Chandola, Banerjee, and Kumar (2009), this batch of methods is based on a main assumption that *"Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model"*. In order to examine whether a data example is an exception or not by using statistical techniques, a statistical model is fitted to the given data, then the decision will be made by computing the probability of generating this sample by the stochastic model assumed.

3.2.1. Statistical models

While searching the literature, we found a few research directly related to the application of statistical models to detect the opinion leaders. For example, Wang and Zhai (2017) recently published a survey on the application of statistical models for sentiment analysis and opinion mining. To be more specific, the authors in Guimera and Amaral (2005) utilize a stochastic model to detect the "Modules" in complex networks. Here modules referred to those communities, which have highly interconnected nodes, and less connected to other communities. In order to define the roles of each node in its own community and with respect to other communities, they proposed the within-module degree and participation coefficients, respectively.

1. *Within-module degree*: this measure that indeed is a parametric statistical model, compute the strangeness of connectivity of a node to the other nodes within its own module or community. This measure is also named as the maximum normed residual test or z-score which is formulated as follows:

$$z_i = \frac{k_{in_i} - \overline{k_{s_i}}}{\sigma_{k_{s_i}}} \tag{1}$$

where k_{in_i} is the internal degree of vertex i within its module, $\overline{k_{s_i}}$ is the average internal degree of all the vertices in community s of the node i and $\sigma_{k_{s_i}}$ its standard deviation. The problem of this measure is that it does not consider the connections to other communities, so two vertices within the same community, with the equal internal degree but the different external degree, will get the same z-score. Hence, participation coefficient is defined as follows:

1. *Participation coefficient*: this is also a statistical model which measure the connectivity of a node to other communities that is presented in Eq. (2).

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{s_i}}{k_{in_i}}\right)^2 \tag{2}$$

where N_M is the number of modules and $0 \le P_i \le 1$ defines if the connection of a node is uniformly distributed among all the communities, this value would be close to one, otherwise, it would be zero, in a case of just fully inter-community connections (Guimera & Amaral, 2005). The authors came up with seven roles for vertices w.r.t. joint evaluation of these measures. The z-score of vertices classify them as hubs or non-hubs, which are additionally classified by the participation coefficient value. In increasing order of participation coefficient, non-hubs were classified as ultra-peripheral (I), peripheral (II), non-hub connectors (III) and non-hub kinless (IV) and hubs were classified as provincial (V), a connector (VI) or kinless hubs (VII).

A particular network is actually a *single observation* that consist of numerous components such as nodes and links. Accordingly, the case of *parametric bootstrap* is used to fit the data by resampling from the components. *Parametric bootstrap resampling* is utilized in Rosvall and Bergstrom (2010) to quantify changes in large networks. For this purpose, the authors used a statistical significance metric to identify the prominent communities and to detect *significant vertices* that are the most adherent to a cluster. To get the statistical significance, they gather a collection of *bootstrap networks* by resampling from the parametrized link weights, that consequently does not impair the individual characteristics of the nodes. Finally, the final distinct cluster is that one which does not have 95% significant similarity with other cluster's subset among all bootstrap networks. As a result, for opinion leader detection such reduced distinct clusters could be considered as a set of the most influential nodes. Actually, in practice, it was not used for opinion leader detection, the proposed method is applied to discover the significant structural changes in the scientific research field by doing experiments on the papers citation network collected from about 7000 scientific journals (Rosvall & Bergstrom, 2010).

3.2.2. Stochastic process

In probability theory, a stochastic process is a random system or phenomena that evolves during the time in a random manner, with enormous outcome as results, due to its randomness. Various outcomes are related to time by *index set* over a given *state space* that could be a n-dimensional Euclidean space or just the integers. In such processes, changes among the index values are considered as different points in time. There are different ways to classify a stochastic process by using some factors such as state space, index set, or by utilizing the dependency of random variables, i.e. the cardinality of the index set and the state space. Modeling the commu-

nications in the network as a stochastic process, found many applications like the transmission of infectious disease, viral marketing campaigns and opinion leader detection problem (Amor et al., 2016; Leibovich, Zuckerman, Pfeffer, & Gal, 2017; Zhao, Erdogdu, He, Rajaraman, & Leskovec, 2015).

3.2.2.1. Markov stability. We start this section with some basic definition. In probability theory and statistics, the memoryless property of a stochastic process is called *Markov property*. In fact, a stochastic process has the *Markov property* if only the present state of a process affects on the conditional probability distribution of a future state. The *Markov chain* is a stochastic process that introduces a discrete index set, often representing time, and the process is represented as a sequence of events in which the probability of each depends only on the accomplished state in the previous event (Gagniuc, 2017). That model of the stochastic process could be applied on graph structure where the *Markov matrix M* is defined as a uniform probability that depends only on the current vertex in the graph as an inverse of its degree for immediate neighbours and zero for other vertices. This way it is possible to compute a probability of random walk through Markov matrix, e.g. *i, j* entry of M^k represents the probability of a length k that the random walk is going from v_j to v_i vertices respectively. Markov chain has application in community detection (Lambiotte, Delvenne, & Barahona, 2014). *Markov stability* reveals clusters of nodes among which the diffusion process flow becomes ambushed in a particular timescale (Lambiotte et al., 2014). In such way, the method implicitly identifies possible opinion leaders, detecting leader behaviour of nodes through communication distribution comparison.

The authors (Amor et al., 2016) proposed a method for community discovery and associated role classification on followers and retweet networks separately. The idea is based on the role difference of users in the spreading of information, that they define by *inflow* and *outflow communication patterns*. The Markov stability is applied to the generated *role-based similarity graph* where it achieves groups of similar *communication patterns*. The role-based similarity graph itself is generated by the cosine measure over *profile vectors* and the *relaxed minimum spanning tree* algorithm (Amor et al., 2016). The *profile vector* (Eq. (3)) for a node v is a $1 \times 2K_{\max}$ vector, where the first entries $1 \times K_{\max}$ describe the number of paths of length 1 to these nodes from v, and the second $1 \times K_{\max}$ entries give the number of paths which end at v (scaled by a tunable constant). $(A^k)_{ij}$ is the number of paths of length k between nodes i and j.

$$X(\alpha) = [\overbrace{\dots(\frac{\alpha}{\lambda_1}A^T)^k 1\dots}^{\text{incoming}}|\overbrace{\dots(\frac{\alpha}{\lambda_1}A)^k 1\dots}^{\text{outgoing}}] \tag{3}$$

where $\alpha \in (0, 1)$ is regulator parameter and λ_1 is the largest eigenvalue obtained from adjacency matrix A.

The experiments of the framework were conducted on retweet network dataset, where they have identified discussion communities with specific topics. This is a role-based unsupervised group revealing technique that could reduce the complexity of future methods for opinion leader detection.

Self-exciting point process. Many of the real world stochastic processes need more complicated mathematical modeling, rather than Markov process, since each arriving entity provokes the process in such a way that the chance of a subsequent arrival probability is increased for some time period after the initial arrival, e.g. trade orders, earthquakes. One of the well known non-Markovian extension of the *Poisson process* is *self-exciting* or *Hawkes process* (Laub, Taimre, & Pollett, 2015), that is often used to model "rich get richer" phenomena assuming that in a process all the previous instances have an impact on the future evolution of it (Eq. (4)), unlike the Poisson process that assumes constant intensity over time.

$$\lambda(t) = \mu(t) + \sum_{T_k<t} r(t - T_k) \tag{4}$$

where T_k is the time of occurrence of the kth event in the process, r defines as a kernel function which expresses the positive impact of past events T_k on the current value of the intensity process $\lambda(t)$, and μ is a non-stationary function representing the deterministic part of the intensity. Such a model is ideal for social network dynamic modelling, e.g. information cascades of post reshares.

Opinion leader can be identified in a social network by the expected number of possible reshares for his/her fresh post. Theoretical framework SEISMIC (Zhao et al., 2015) implements this idea on the basis of self-exciting point process for twitter microblogging platform to predict an expected number of possible reshares of a particular post. This way, such framework finds application for opinion leader detection by identifying breakout tweets. The framework tends to describe the temporal patterns of information propagation and can give a probability of the possible follower in addition to an estimated number of a possible retweet for a particular post. In this framework, different factors of each post such as the quality of the post's content, the current local time, network topology and etc, are combined into a single infectiousness parameter p_t to measure the reshare probability. In other words, they updated Hawkes processes (Eq. (4)) by assuming the process intensity λ_t depends on the stochastic process p_t that allows the infectiousness to change over time. Given the number of reshares R_t of a considered post during a predefined period of time t and the cascade speed of spreading λ_t, that also takes into account a human reaction time, this framework is aiming to predict the final number of reshares R_∞. In Fig. 3, the structure of the information diffusion tree is shown. To simulate the growth of the cascade tree, the authors define random variables Z_k, which denote the number of reshares created by the kth generation descendants. Here, the final reshare counts R_∞, can be computed as $R_t + \sum_{k=1}^{\infty} Z_k$. This model shows a convincing results by achieving 15% and 25% relative error in predicting the number of final retweet in 60 and 10 minutes as time period, respectively. SEISMIC has linear time complexity and low space complexity since it is based on some simple factors, including "the time history of reshares, the degrees of the resharing nodes, and minimal knowledge about the information cascade and the underlying network structure".

3.2.3. Summary

This section discusses the applications of statistical and stochastic models for opinion leader detection. Although the number of papers that use this group of methods to analyze the influential nodes in the social networks is few, the performance of statistical methods is strongly depended on the key assumptions of the underlying data distribution and the predefined confidence interval. Moreover, the nature of the statistical models determines their computational complexity, for example, Gaussian, Poisson, and Multinomial distribution have a linear complexity with the number of data size and attributes. The advantages of stochastic models are that they can provide a theoretical framework to define and describe the temporal patterns of information cascades. Additionally, such methods are flexible since they require minimal knowledge about information cascade and semantics of data, thus they could be generalized and extended by other approaches to get a new method for opinion leader. One of the main limitations of statistical methods in analyzing the high dimensional graphs is the uncertainty about their distribution. In addition, choosing the most proper hypothesis tests is a challenging task, which requires a judge of a professional expert. Hereby, statistical and stochastic

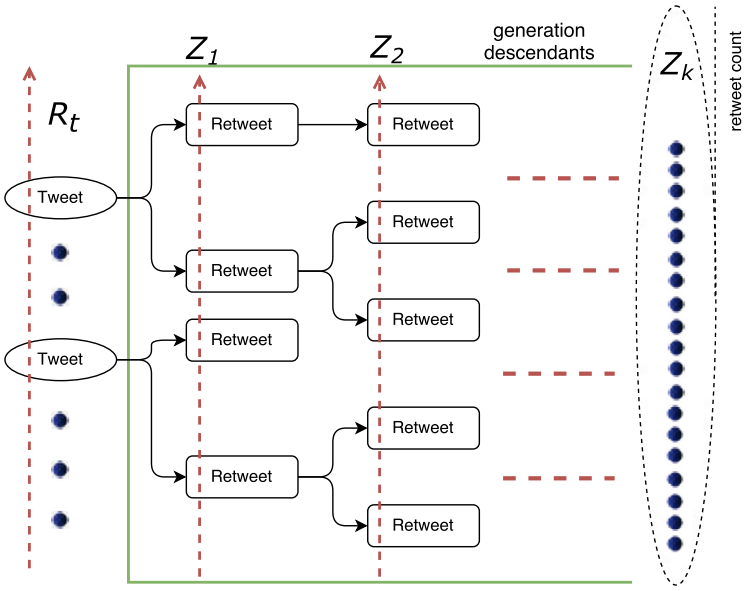


Fig. 3. The information diffusion tree presentation (Zhao et al., 2015).

approaches may result in finding fake opinion leaders, those sharing too many links from others as these methods only consider network links and interaction information while the real content of provided information could be just a replica from a real opinion leader.

3.3. Diffusion process based approaches

In a network of customers, a user may influence some neighbouring users to buy a particular item and thus continue the chain of purchases recursively that leads to increasing profit from sales of a particular product. In this way, the measurement of *expected profit* is the network value. The study of the *diffusion of innovations* throughout social groups takes a big role in devising business strategies for their products because it helps to model customers *network value* (Rogers, 2010). Nowadays, the information overload presented in the WEB, makes the marketing field to be more specific and original to catch the attention of online users. Moreover, the innovations are various products of sophisticated construction and forms (Haron, Johar, & Ramli, 2016). Thus users hardly manage optimal decision making on product choosing among a big number of candidates (Bilici & Saygın, 2017; Duan, Gu, & Whinston, 2009). Therefore, since the 1990s the study about the diffusion of innovation got a rapid growth in *simulation models*. One of the general simulation models for diffusion of innovation is the *influence cascade* (Bikhchandani, Hirshleifer, & Welch, 1992) that reflects a group behaviour of economic agents making decisions not only on the basis of their own information but also taking into account the observed behaviour of other agents. The influence cascade model also considers a personal decision process of an agent as an expense rather than a group decision that actually reflects the usual customer behaviour. Such diffusion model mostly used to test various strategies for gaining initially leading role in the adoption of a product in the network of clients.

The majority of approaches for stochastic influence cascade modeling are based on the *linear threshold* model (LT) (Granovetter, 1978) and *independent cascade* model (IC) (Goldenberg, Libai, & Muller, 2001). Generally, these models are some specific modification of graph percolation process, since a set of initial (given) nodes are assumed to be active being able to spread an influence to its neighbouring nodes. Afterward, at each step an inactive node may become active under assumptions of some specific criteria (heuristics). In the IC model a node activation, or influence propagation, involves *edge activation probability* while the LT model introduces a uniformly chosen threshold for the weighted sum (edge weight) of its active neighbours.

3.3.1. Influence maximization approaches

A social network topology changes with time representing individuals and their actual interactions, such platform plays a vital role in spreading of information, ideas, and influence among its members. By analysing the dynamic exchanges between participants of an interaction network it is possible to reveal the *flow of influence* resulting to valuable insights about the locally or globally *opinion leading* participant of the network (Proskurnikov, Tempo, Cao, & Friedkin, 2017). The flow of information could be affected by strategies for promotion of innovations. *Influence maximization problem* was introduced to find such strategies (Kempe, Kleinberg, & Tardos, 2003). Therefore, the correctly modelled information diffusion processes or flow of influence in a network could be applied for opinion leader detection task. In fact, by having an interaction network, it is formulated as the problem of identifying an initial set of nodes of predefined size k with the biggest influence spread S to the network. The resulting initial seed set of users, from the influence maximization solution, are actually used for spreading an advertisement for a new product. For the problem, there was proposed a discrete optimization version of it, *greedy hill climbing* approach (Kempe et al., 2003), that result in 63% of the optimal solution. The influence maximization is the optimization problem of NP-hard complexity, moreover, a typical collaboration graph from a social network is prohibitively large. Hence, in the case of discrete optimization of the influence cascade model, the influence spread is estimated by the *Monte-Carlo* simulations that result in a non-scalable solution. Therefore, many studies on discrete optimization of the influence maximization problem were aimed to accelerate such models to be efficient and scalable (Chen, Yuan, & Zhang, 2010). For example, Zhao, Li, and Jin (2016a) propose a method by combining influence maximization algorithm and *label propagation*, named (IM-LPA) to identify opinion leaders in social networks with community structure.

The diffusion process was also studied for identification of aspects and features of efficient opinion leaders that could drive

strong innovation promotion. In the study of technological innovation adoption (Cho, Hwang, & Lee, 2012) the authors examine diffusion processes with the assumption of opinion leaders initiate it. They also examine the resulting effect of opinion leaders holding different features, e.g. degree of sociality. Considering the most acquired number of adopters, they suggested that marketers can use *rank-nomination* and *distance* centralities to identify an opinion leader (Cho et al., 2012). In addition, they state that *sociality centrality* has been proved to be the most effective measure to get higher diffusion.

3.3.2. Contagion theory

The *Contagion theory* developed by Le Bon (1897) is a theory of collective behavior which explains the hypnotic impact of the crowd on individuals (Monge & Contractor, 2001). The theory is widely used in the diffusion of a disease through some population, how the loss of one species in an ecosystem has an effect on others or the spread of some information within a group of people. A number of different parameters define how diffusion happens and how long it takes. This theory is mostly about infectiousness metric, that explains how infectious is spreading on the network and how resistant are the nodes to the infection, resulting in a *resistance parameter*. Finally, it reveals if the diffusion is taking place in a strategic manner or randomly. The contagion theory is mostly applied to the social networks in order to study the *adoption diffusion* of a market product, by understanding the adoption strategies for innovative *opinion dynamics* and thoughts of influencers. The authors in Weisbuch, Deffuant, Amblard, and Nadal (2002) believe that in the era of tremendous information exchanges, information contagion theory has more applications in modeling the adoption dynamics than methods of *game theory* (Weisbuch et al., 2002). In the approaches based on game theory, payoffs are directly learned through consequent switching behaviors denoted as the pre- and post-interaction states, but the long-term costs of *switches* are underestimated, while in contagion theory the long-term effect of outcomes tends to be defined with initial parameters given to the system. Among the different models of opinion dynamics, bounded confidence-based models have been studied in many contexts and the models have shown interesting flexibilities, particularly in clustering, polarization and fragmentation of opinions, and influence of extremists (Chen, Glass, & McCartney, 2016a; Dabarera, Premaratne, Murthi, & Sarkar, 2016; Deffuant, Amblard, & Weisbuch, 2004; Kurmyshev & Juárez, 2013; Zhao, Zhang, Tang, & Kou, 2016b). Therefore, such a model in this group hardly detects opinion leader, but it gives a good framework to evaluate its possible impact on a already known network of users (Zhao et al., 2018).

3.3.3. Diffusion centrality

Most of the topological centralities (Section 3.4.1) do not consider more than only network structure to determine the centrality of nodes. In this situation, it is hard to use them for detecting opinion leader in dynamic networks or in diffusion modeling. In order to tackle the limitations of conventional centrality measures, diffusion centrality (DFC) is proposed by Kang, Molinaro, Kraus, Shavitt, and Subrahmanian (2012). The goal of DFC is to detect central nodes with respect to the spread level of a diffusive property p by using a given diffusion model for p (Kang, Kraus, Molinaro, Spezzano, & Subrahmanian, 2016), and the network topology. This way, given a property p and a predefined diffusion model, DFC reveals the high spread nodes with high probability that could be opinion leaders. The DFC metric (Eq. (5)) is computed as the difference of two components, both are the expected number of infected vertices, by the presumption that v has diffusive property p in the first component and otherwise for the second one.

$$DFC(v) = \sum_{v' \in V - \{v\}} p(T_{\oplus p,D})(v') - \sum_{v'' \in V - \{v\}} p(T_{\ominus p,D})(v'') \quad (5)$$

where $T_{\oplus p,\ D}$ and $T_{\ominus p,\ D}$ are predicates of having or not having the property p by given diffusion model D. The experimental analysis of the DFC illustrates the promising results on real YouTube data (Kang et al., 2012). It should be noted that since a diffusion model is used to reveal the top centralities, the implementation of the method based on DFC has higher computational complexity than typical topology-based centrality metrics but faster than betweenness and closeness centralities.

3.3.4. Summary

In the high-tech area, firms face with rapid improvement and evolution of technologies, e.g. block-chain got the reputation of the fastest growing technology with changing concepts and applications. Firms need different marketing strategies to tackle with a dynamic environment and use the force of social network effect and other aspects of their properties. Additionally, to the network dynamic properties, the degrees of trust of opinion followers toward opinion leaders have an important effect on the influence power of opinion leaders (Zhao et al., 2018). Therefore, to design a diffusion model, we need to consider various properties with different weights, which could lead to an effective and real world like opinion leader detection approach. Moreover, for the marketing, the opinion leaders are attractive goals because of their diffusion speed and the accumulative number of customers that may accept the new products (Cho et al., 2012), that is directly used in influence maximization methods to detect the leaders. Actually, the social network or a blogosphere is a dynamic environment where an opinion leader mostly shared his/her thoughts and beliefs with others that influence the whole groups ideas and also its spreading process. The opinion leader detection is very sensitive to the model itself, in other words, the model decides the possible global or local flow of a node.

3.4. Topological measures

The social interactions of individuals in the networks are constructing complex network structures, which have a proper knowledge of their topology is strongly useful to understand and predict the role and importance of influential persons in the network. In fact, topological factors of users or posts network that reflect social indicators can be utilized to identify key influencers in a social network environment. The social network analysis mostly relies on topological metrics such as centrality and community concepts, and many of the terms used to measure these metrics are a reflection of their sociological origin. The heuristics of network topology are used to identify the possible leaders by their position in the network (Li, Li, Van Mieghem, Stanley, & Wang, 2015a), e.g., vertex centrality and community structure. These methods assume that if a node is located in an appropriate location in the graph, it could be an influential node.

3.4.1. Centrality metrics

The centrality metrics have been investigated in various aspects of highlighting the important nodes. Centrality measures vary on different aspects of the network topology (Freeman, 1978), that differently capture characteristics of the influential user or even the opinion leader. In this section, we will consider just *topological centralities* that are ranking of the nodes based on the network structure and position in it. The topological structure contains the inbound and the outbound links that respectively represent influence and prestige effect. This way it is important to separate the case of a non-oriented and oriented graphs since in a non-oriented graph,

each node has only one type of links that could play both roles of the inbound and the outbound links. For the opinion leader detection problem, it is necessary to mention that influence is based on the choices made by a user while prestige of a user comes from the attention of others toward him/her. The combination of both measures is used in the advanced centrality metrics simultaneously, e.g., *PageRank* and *LeaderRank* (Lü, Zhang, Yeung, & Zhou, 2011; Page, Brin, Motwani, & Winograd, 1999).

Degree centrality. The *degree centrality* (DC) is a measure of local centrality that is calculated from the immediate neighborhood links of a vertex. DC is the most intuitive idea of the validity of a user within a group (Freeman, 1978) without considering the global structure of the graph. The idea is to compute the normalized number of links incident to a node as shown in Eq. (6). In graph theory, it is known as the *degree of a node*.

$$DC(v_i) = \frac{1}{N-1}\sum_{j=1}^{N}\alpha_{i,j} \tag{6}$$

where, N is a number of vertices,$\alpha_{i,j} = 1$, if there is a direct link between v_i and v_j s.t. $j \neq i$. In case of oriented graph, it has measures of degree centrality for inbound and outbound by the same equation Eq. (6), respectively.

Risselada, Verhoef, and Bijmolt (2016) providest an empirical investigation of self-reported opinion leadership on smartphones adoption and degree centrality metric among the customers of a mobile telecom operator. The authors show a significant and positive effect of DC on an actual opinion leadership. Nonetheless, the self-reported leaders have a bigger influence on strong tights. The effect of the immediate influence level is not possible to be captured by DC since the spread of adoption behavior and accepting the unknown products need an intense relationship. The structure of the network is important, but also the nature of the relationship between adopter and potential followers should be captured in a more sophisticated centrality measure. This means that the available effort of DC per person would be smaller in a big customers network (Katona, Zubcsek, & Sarvary, 2011).

The degree centrality is an indicator for local opinion leaders, the possible positive application is restricted to small clusters of users like a *forum*. The obvious deduction about a user with high DC is a frequent communicator directly with other users that have the possibility of influencing them (Bodendorf & Kaiser, 2010). Therefore, DC is an acceptable measure of the immediate rate of influence spread from nodes in a short-term perspective (Ilyas, Shafiq, Liu, & Radha, 2013). To detect influential nodes more accurately there was proposed *ClusterRank* (CR) (Chen, Gao, Lü, & Zhou, 2013) that detects spreading the influence of each node v according to its clustering coefficient as presented in (Eq. (7)).

$$CR(v) = 10^{-c_v}\sum_{u\in\mathcal{W}_1(v)}\left(k_v^{out}\right) \tag{7}$$

where c_v is a clustering coefficient and k_v^{out} is out-degree of vertex v. Results presented in Chen et al. (2013) emphasizes that in an undirected network, the ClusterRank centrality measure performs better than degree centrality and *k-core* decomposition.

3.4.1.1. Closeness and Harmonic centrality. The *closeness centrality* (CC) is a global qualitative metric that shows how quickly a node can communicate with other users in the network (Musiał, Kazienko, & Bródka, 2009) calculated by Eq. (8). A user with high CC is an effective spreader of opinions and information (Yang, Qiao, Liu, Ma, & Li, 2016). Nonetheless, in Ilyas et al. (2013) author states that CC metric is unsuitable for the common problems because of its basic assumption that any flow on the network happens once along shortest paths (Chen, Lü, Shang, Zhang, & Zhou, 2012).

$$CC(v_i) = \frac{N-1}{\sum_{j=1}^{N} c(v_i, v_j)} \tag{8}$$

where N is a number of vertices, $c(v_i, v_j)$ is a function defining the distance between nodes v_i and v_j s.t. $i \neq j$ (i.e. *min, max, mean* or *median*).

The CC is based on adding distances, therefore it is very sensitive to a single large distance or missing link (Dekker, 2005). In the extreme case, disconnecting one node sets distance to infinity and hence CC of all nodes to zero. To tackle this problem, there was proposed *Harmonic centrality* (HC) (Dekker, 2005; Rochat, 2009), reversing the sum and reciprocal operations in the definition of closeness centrality (Eq. (9)).

$$HC(v_i) = \sum_{j\neq i}\frac{1}{c(v_j, v_i)} \tag{9}$$

where $1/c(v_j, v_i) = 0$ if there is no path from v_j to v_i. Harmonic centrality can be normalized via dividing by $N-1$. The CC and PageRank have a large intersection set of the opinion leaders identified by them since CC detect closeness of a node to all other nodes in the graph and PageRank also tend to rank higher such nodes which are closer to others.

Betweenness centrality. The *Betweenness centrality* (BC) is a measure (Eq. (10)) that pinpoints how much a node is able to control the flow of information between other nodes in the network (Freeman, 1978). A node with high BC shows a great capacity to facilitate the interaction between the other nodes, it means that the network without a node with high BC would split into subnetworks.

$$BC(v_i) = \frac{\sum_{i\neq k\neq j}\sigma_{i,j}(v_k)}{\sum_{j=1}^{N}\sigma_{i,j}} \tag{10}$$

where N is a number of vertices, $\sigma_{i,j}$ is the total number of geodesic paths connecting v_i and v_j, $\sigma_{j,j}(v_k)$ is the number of geodesic paths including v_k.

For opinion leader detection it is possible to consider that a node with the role of leadership in the network operates as a bridge for the shortest path among all possible pairs of nodes. Actually, BC is difficult to be applied in large-scale networks due to the computational complexity.

Eigenvector centrality. The *eigenvector centrality* (EC) suggests the idea that a node is more central as it is connected to important (central) nodes (Bonacich, 2007). A node with high eigenvector centrality is not surely well connected to others. The centrality of node v_i is proportional to the centrality of nodes which v_i is connected, in other words, EC is proportional to a location near the most significant nodes or communities in a graph. Let $A = (a_{i,j})$ to denote the adjacency matrix of a graph. The eigenvector centrality x_i of a node i is calculated by Eq. (11).

$$x_i = \frac{1}{\lambda}\sum_{k}^{N} a_{k,i}\,x_k \tag{11}$$

where $\lambda \neq 0$ is a constant. In the matrix form, it is presented in Eq. (12).

$$\lambda x = xA \tag{12}$$

Hence the EC vector x is the left-hand eigenvector of A associated with the eigenvalue λ. The better choice for λ is the largest eigenvalue for the matrix of absolute values of A. The EC has important shortcomings on directed networks. It is possible to have zero scores for a node in case of no incoming links Eq. (11), that leads to zero contribution in centrality metric of other nodes. That

is a big issue for acyclic graph, e.g. network of scientific publications where all papers have been assigned zero scores. Additionally, EC is proportional to a location near the most significant nodes that makes the metric to be shifted toward the biggest community structure in the network, especially for large graphs.

Katz centrality. The *Katz centrality* (KC) (Katz, 1953) is a variant of Eigenvector centrality. The measure of a vertex importance considers not only the *1-hop neighbors*, but also the numbers of hops from the vertices of the connected component to the target vertex. Let G be a *loop-free* network and A is the adjacency matrix. The KC of a vertex v_i is calculated by Eq. (13).

$$KC(v_i) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ji} \tag{13}$$

where A^k is the kth power of A and $\alpha < 1$ is a positive damping factor, the role of which is to give the path weight of length varying from 1 to ∞ between vertices i and j, it means that the greater path to a neighbor, the lower influence on the neighbor. The matrix notation of Katz centrality is presented by Eq. (14).

$$KC = \sum_{k=0}^{\infty} \alpha^k A^k = (\mathcal{I} - \alpha A)^{-1} - \mathcal{I} \tag{14}$$

To reach the *matrix convergence,* α should be chosen, s.t. $A(0 \geq \alpha < 1/|\lambda_{\max}|)$, where $\lambda_{\max}$ is the largest eigenvalue of the adjacency matrix Eq. (15).

$$KC(v_i) = \sum_{j=1}^{N} KC_{i,j} \tag{15}$$

KC has a performance bottleneck, in case the inverting of the adjacency matrix that initially induces a complexity of $O(n^3)$ (Landherr, Friedl, & Heidemann, 2010). However, this complexity can be reduced by applying the algorithm of Coppersmith and Winograd (1990) to $O(n^{2.376})$ (Coppersmith & Winograd, 1990).

A node is able to influence the other nodes' centrality just by introducing new links to a particular group of nodes. It happens because the Katz centrality considers the numbers of hops from the vertices of the connected component to the target vertex that's why it suffers from heterogeneous out-degree distribution. Such limitation could be avoided by weight reduction for ingoing links from spendthrift nodes as in the PageRank centrality.

PageRank. The *PageRank* centrality (PR) is a variant of the eigenvector centrality that changes the case when a node with high centrality, is linked to many other nodes, so causes all those nodes get high centrality, too (Page et al., 1999). PageRank algorithm is used to calculate the importance of any node (e.g. Web page) according to the links the node receives. The method considers count and quality of links to a node in order to roughly estimate how important the node is.

The main idea of PageRank is to consider three distinct factors: (i) the number of ingoing links, (ii) the linkers predisposition for linking, and (iii) the centrality of the linkers. For a node, the value of the endorsement decreases proportionally to the number of ingoing links from vertices with high out-degree centrality. In other words, the links coming from greedy nodes are worthier than those with relatively numerous outgoing links. The PageRank considers the network topology by an endogenous component, and it also depends on the exogenous component which is independent of the network structure. Let $A = (a_{i,j})$ to be the adjacency matrix of a directed graph. The PageRank centrality $PR(v_i)$ of node v_i is given by Eq. (16).

$$PR(v_i) = \beta + \alpha \sum_{j \in \mathcal{W}(i)} \frac{a_{k,i}}{k_j^{out}} PR(v_j) \tag{16}$$

where $\mathcal{W}(i)$ neighbors with ingoing links to node i, k_j^{out} is the out-degree of node j if such degree is positive, or $k_j^{out} = 1$ if the out-degree of j is null. To reduce the influence of the others nodes, this total vote is "damped down" by multiplying it by α called *damping factor* constant, and $\beta = 1 - \alpha$ means that if a page has no ingoing links even then it will still get a small PR. The matrix form of PR centrality is given by Eq. (17) (x = PR).

$$x = \alpha x D^{-1} A + \beta \tag{17}$$

where β is a vector whose every element is equal to a given positive constant and D^{-1} is a diagonal matrix with ith diagonal element equal to $1/d_i$.

The PR centrality tends to identify the primary source node that makes the metric to be very indicative of opinion leader detection. Analytical tools for opinion leader detection has direct application of PR in their system (Li, Huang, & Sun, 2015b). Actually, PR has a big amount of variants for ranking purpose on microblogs to find the most influential posts and even to detect an opinion leader on the basis of posts, comment or interaction network (Chen et al., 2014a; Eom & Shepelyansky, 2015; Guha, Kumar, Raghavan, & Tomkins, 2004; Richardson, Agrawal, & Domingos, 2003; Tai, Ching, & Cheung, 2005; Weng, Lim, Jiang, & He, 2010). For opinion leader detection, it also has variations in combination with topic models, e.g. InfluenceRank (Song, Chi, Hino, & Tseng, 2007), OpinionRank (Zhou & Zeng, 2009), Dynamic OpinionRank (Huang, Yu, & Karimi, 2014; Song, Wang, Feng, & Yu, 2011) and others.

LeaderRank. The *LeaderRank* (LR) (Lü et al., 2011) was devised similar to PageRank method except that it proposes a *ground node*, which makes the graph fully connected by introducing links to the rest of nodes in the network. The ground node has the same role to the damping factor in PageRank. In fact, this factor is personalized by the network itself making the LeaderRank be parameter free and eliminates the calibration steps of PageRank. Actually the ground node enforces a degree-dependent random steps probability, that is used instead of random teleportation steps in PageRank.

By considering a network of N nodes and M directed links, the method introduces a *ground node* and connects it to every node of the network through *bidirectional links*. The augmented network thus becomes powerfully connected and consists of $N+1$ nodes and $M+2N$ links. The ranking process is implemented in the way of *random walk* that is described by the *stochastic matrix P* with elements $p_{i,j} = \alpha_{i,j}/k_i^{out}$ representing the probability of the next step of a random walker going from node i to node j, where k_i^{out} denotes out-degree. The random walk step of LR is defined as the score of the node i at time t in the Eq. (18). The initial score equal to one is assigned for all the nodes except the ground node which is initially rated with zero score.

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{\alpha_{i,j}}{k_i^{out}} s_i(t+1) \tag{18}$$

The experiments on a network of people in the "Delicious" bookmarking service shows the LeaderRank outperforms the PageRank in the task of identifying influential users (Lü et al., 2011). Additionally, authors emphasize the robustness property of LR against manipulations and noisy data.

Semilocal and local structural centralities. The authors of Chen et al. (2012) proposed a *semi-local centrality* (SLC) to identify users with the most influential power in an undirected network concerning the nearest and the next nearest neighbors. The authors also show that SLC (Eq. (19)) is computationally efficient metric and it has a strong positive correlation with

closeness centrality and a weak correlation with both degree and betweenness centrality.

$$SLC(v) = \sum_{u \in \mathcal{W}_1(v)} \left(\sum_{w \in \mathcal{W}_1(u)} N_2(w) \right) \tag{19}$$

where $\mathcal{W}_1(v)$ is the set of the nearest neighbors of node v and $N_2(w)$ is the number of the nearest and the next nearest neighbors of node w.

The model performance is evaluated by the *Susceptible-Infected-Recovered* (SIR) model (May & Lloyd, 2001) with different spreading rate and the number of infected nodes. They suggest that the SLC identify influential users better than degree centrality-based and betweenness centrality-based measures (Chen et al., 2012). While the other existing random-walk approaches such as PageRank and LeaderRank are better indicators of influential nodes, but less computationally efficient. The SLC method performed differently depending on the network structure, e.g. having bad results for tree-shaped networks.

Gao, Wei, Hu, Mahadevan, and Deng (2013) proposes *evidential Semi-local centrality* (ESC) combining an extended semi-local centrality for weighted networks and *evidential centrality* (EVC) that is a measure of node influence based on the Dempster-Shafer theory of evidence (Dempster, 1967; Shafer et al., 1976). As a result, the ESC measure is more reasonable than the existing evidential centrality due to the fact that the degree distribution of a complex network is taken into consideration, moreover, the proposed approach can effectively identify influential nodes in weighted networks.

For nodes with the same semi-local centrality, the one with denser connected neighbors is supposed to have stronger spreading ability since denser connected neighbors get more chance to influence each other. The topological connections among neighbors of a node have an impact on its spreading ability. For such case, there was proposed *local structural centrality* (LSC) (Gao, Ma, Chen, Wang, & Xing, 2014), which considers both the number of nearest and the next nearest neighbors and the topological connections among neighbors Eq. (20).

$$LSC(v) = \sum_{u \in \mathcal{W}_1(v)} \left(\alpha N_2(u) + (1-\alpha) \sum_{w \in \mathcal{W}_2(u)} c_w \right) \tag{20}$$

where $\mathcal{W}_2$ set of nearest and next-nearest neighbors, c_w represents the local clustering of node w and α is a tunable balance parameter between 0 and 1. The results presented in Gao et al. (2014) shows algorithm's superiority over SLC to detect influential nodes. SLC and LSC both considers the nodes local information in the computationally efficient way and could be applied over a large-scale networks.

3.4.2. Community measures

The importance of an individual in a network could be measured in various ways presented in the previous section. Additionally, we could use the community structure of the network to make, the more context-oriented level of importance for every user in it. A community structure is introduced by Girvan and Newman (2002) where they define it as an area with a high connection density in the graph. In most scenarios, *clusters* or *communities* are groups of vertices with common properties (Newman, 2004), e.g. common friends, frequent communication with each other, like-minded individuals, etc. For measuring opinion leadership level of a node, the main factor is the social location of the node that is the case of community-driven measures, since most of the metrics of this group consider both the connections inside and outside the community (Katz, 1957).

Embeddedness. The *embeddedness* measure shows a proportion of neighboring nodes belonging to its own community (Lancichinetti, Kivelä, Saramäki, & Fortunato, 2010) as presented in Eq. (21), where the k_{int} is the number of links to other vertices within the community and k is the total degree of the same node.

$$e = k_{int}/k \tag{21}$$

The embeddedness metric is ranging from zero to one, here one represents the case where all the neighbors are in the same community ($k_{int} = k$), on the other hand, when all the neighbors belong to different communities ($k_{int} = 0$) this value will be zero. The main issue with embeddedness measure is that it does not consider the internal degree of the vertex, as it just insist of the ratio to the total degree, that could be misleading. In real networks, most of nodes usually have a low degree inside their own community and no edges outside that lead to a very high embeddedness value near $e = 1$ (Lancichinetti et al., 2010; Orman, Labatut, & Cherifi, 2012). In this kind of case, the embeddedness measure is unable distinguished between important and unimportant vertices for a community, and this way the metric has not a direct application for opinion leader detection. Vertices, located at the border of their communities, have embeddedness value significantly less than one, this way it is possible to say about the likelihood of a vertex to leave its community in the future (Parau, Lemnaru, & Potolea, 2015). This way, the embeddedness measure could be used to modify other measures or create a new one. For example *relative commitment measure* that is used for simple vertex classification and finally could be applied for opinion leader detection.

Relative commitment measure. Relative commitment measure (RCM) (Parau et al., 2015) of a vertex computes the strength of the membership of a vertex to its community in the way of combining embeddedness (Lancichinetti et al., 2010) and significance (Rosvall & Bergstrom, 2010) measures. RCM metric is calculated as the ratio between the internal score and the total score, the result of which is multiplied by the relative internal degree of the vertex (Eq. (23)). Relative internal degree of vertex i is rk_{in_i} that is presented in Eq. (22).

$$rk_{in_i} = \frac{\log(k_{in_i} + 1)}{\log(Max_s(k_{in}) + 1)} \tag{22}$$

where k_{in_i} represents the internal degree (number of edges inside the community) of vertex and $Max_s(k_{in})$ is the maximum internal degree in community s.

$$RCM(t) = rk_{in_t} * \frac{\sum_i rk_{in_i}}{\sum_i rk_{in_i} + \sum_j rk_{in_j}} \tag{23}$$

where the internal and external neighbors are denoted as i and j, respectively.

Parau et al. (2015) states that the relative commitment provides a more accurate estimation of a vertex commitment than embeddedness measure. Moreover, RCM is descriptive measure. Combining embeddedness and significance measures RCM describes why a community exists by detecting the *keeping vertices* from leaving its community, that is of a big interest for detecting opinion leaders of a particular community.

Principal Component Centrality. Principal component centrality (PCC) (Ilyas & Radha, 2011) determines users with influential neighborhoods in a network (Rocha et al., 2016), in other words a user himself could be poorly or well connected but it will have a bigger PCC just if it is tied with well-connected users. The PCC takes into consideration communities and it is more robust to the influence of the biggest community. The PCC is computed as Eigenvector centrality (Eq. (24)) on an adjacent matrix, that results in a centrality measure for each user.

$$PCC = \sqrt{((AX_{N\times P}) \odot (AX_{N\times P}))1_{P\times 1}} \tag{24}$$

The parameter P is a tuning factor to tune the number of eigenvectors included in PCC, s.t. $P \ll N$ where N is number of vertices in the given network, A is adjacency matrix and the $\odot$ is the *Hadamard operator* (Mathias, 1993). A vector of ones of length P is denoted by $1_{P\times 1}$. Let $X = [x_1 x_2 \dots x_N]$ denote the matrix of concatenated eigenvectors of dimension $N \times N$ and $X_{N\times P}$ denote the sub-matrix of X consisting of the first N rows and P columns. Vector of eigenvalues is denoted as $\wedge = [\lambda_1 \lambda_2 \dots \lambda_N]^T$.

Additionally, the PCC could be computed in terms of the eigenvalue and eigenvector matrices $\wedge$ and X (Eq. (25)), of the given adjacency matrix A. The assumption on adjacency matrix A limits application of the PCC to undirected graphs (Ilyas et al., 2013). The social networks topology is a directed network that needs a generalized definition of PCC Eq. (25).

$$PCC = \sqrt{|(X_{N\times P} \odot X_{N\times P})|\ |(\wedge_{P\times 1 P\times} \odot \wedge_{P\times 1})|} \tag{25}$$

To allow interpretation of centrality scores the PCC value is normalized (Ruhnau, 2000), restricting all inputs to the range [0, 1] (e.g. the *Euclidean norm* or *maximum norm* of the centrality vector).

PCC determines influential neighborhoods in a network as the primary basis for the centrality. A user with high PCC is capable of propagating an opinion well since it has other representative neighboring influential users who could propagate the opinion. This way the higher the centrality, the higher the users influence capability.

Community disruption. *Community disruption* measures the centrality role of a vertex in the structure of its community (Parau et al., 2015) considering both the community without and with the vertex in it. The main idea of the metric states that the important node plays a tight role for its community and its removal will cause sufficient changes in the community structure, for example, the migration of vertices to other communities or the subcommunities organized from a community split. The metric calculation consists of three consequent steps: on the first step it removes a vertex from the graph, afterward the second step determines community structure in this graph, finally, as the third step, it computes the difference between the original community structure and the new one, computing the *variation of Information* (Eq. (26)). Variation of information metric shows the difference level of two community structures, considering the number of vertices clustered together in both structures.

$$VI = -\sum_{ST} N(s,t) \log \frac{N(s,t)}{N(t)} - \sum_{ST} N(s,t) \log \frac{N(s,t)}{N(s)} \tag{26}$$

where $N(s, t)$ is the number of vertices, that appear in both S and T communities, divided by the total number of vertices. $N(s)$ and $N(t)$ represent the number of vertices that appear in community S and T respectively, divided by the total number of vertices.

Community disruption is a computationally intensive measure of vertex importance for its community (Parau et al., 2015). The measure could be used to identify group leaders and influential people. The most direct application of community disruption, for opinion leader detection problem, is to apply the metric to a repost network or to a communication network, where a detected important post or a leading communicator would be an opinion leader with high probability.

Kernels detection. *Community kernels* are the main key members of communities (Wang, Lou, Tang, & Hopcroft, 2011b), where the kernel of a community include the influential vertices inside the community. In community structure, the authors of Wang et al. (2011b) distinguish the kernels and an auxiliary community of *nonkernel* members (e.g. readers, fans). The metric describes the idea of two distinguishing types of users in social networks since users exhibit different behavior and influence. The biggest part of social network content is produced by the first type, while the others, nonkernel users, have significantly less influence and, consequently, different behavior (Wang et al., 2011b).

Wang et al. (2011b) state that it is impractical to use cut-based algorithms for kernels searching because the definitive links information from auxiliary nonkernel to kernel members would be neglected, e.g. prefiltration by degree centrality and a particular threshold would preserve just top vertices with respect to centrality values. For community kernels detection, the authors propose two different methods, *GREEDY* and *WEBA* that consider the whole topology information. For GREEDY, let's consider an undirected graph $G(V, E)$, a given kernel size k that defines the final number of vertices in a kernel, and for every new kernel initialize a subset $S \subset V$: S of one random vertex $v_r \in V$. The method then initiates the iterative process of enlarging subset S by a top vertex $v' \in V$ with respect to a number of its connection to S. In order to reduce the effect of the initial random point, the GREEDY subroutine is repeatedly executed up-to steady-state results. WEBA, in its turn, heuristically solves an optimization problem over the weight vectors associated with every vertex, where the vector of a vertex represents the relative importance for each community kernel.

A community kernel represents a particular social group, preserving the *homophily* effect (McPherson, Smith-Lovin, & Cook, 2001), that makes instances likely to link to other individuals with similar attribute values. The kernel of a community consists of only influential vertices inside the community, the number of which is regularized by the kernel size. This way the members of one kernel could be considered as competing opinion leaders of the same community since influential users pay attention to those with the most common similarities.

3.4.3. Summary

The topological approach to opinion leader detection problem, aggregated in the table (Table 3), focuses on various aspects of the topological structure of a graph, by analyzing a social influence and connectivity features. Nonetheless, other important aspects of a node are not considered, e.g. user expertise, post time, content similarities etc. The centralities, listed in this section, emphasize the influence or the importance of a node only by topological factors. This way, as the number of nodes raises, the graph will be too complicated to detect opinion leader just with topological methods. It is not sufficient for a node just to have an appropriate location to be considered as an opinion leader. Most of the centralities are used as a part of complex algorithms, e. g. PageRank has a big number of variations applied in topic models.

All the centralities have their pros and cons, there is no universal metric to surely reveal the most influential user or even an opinion leader. As an example, for detecting the most influential nodes there was proposed *coreness centrality* (k-shell centrality) (KS) (Kitsak et al., 2010), that computes a node centrality in the way of balancing the degree and the *k-shell decomposition* of a spreader node (Carmi, Havlin, Kirkpatrick, Shavitt, & Shir, 2007). From the perspective of the KS centrality, the location of a node is considered more important than the number of its linked neighbours, hence by this way a node with larger coreness value represents the more centrality of a node in the network. Nonetheless, according to results from Liu, Ren, and Guo (2013), the nodes with the same k-shell value have different spreading influences. Researchers still improve centrality metrics to make them more efficient and robust, e.g. improved k-shell was proposed in Liu et al. (2013).

Table 3
Topological measures.

Sub category	Measure	Key references
Centrality measures		
	Degree centrality	(Freeman, 1978), (Risselada et al., 2016), (Katona et al., 2011), (Bodendorf & Kaiser, 2010), (Ilyas et al., 2013)
	Closeness centrality	(Musiał et al., 2009), (Ilyas et al., 2013), (Chen et al., 2012), (Dekker, 2005), (Yang et al., 2016)
	Harmonic centrality	(Rochat, 2009), (Dekker, 2005),
	Betweenness centrality	(Freeman, 1978)
	Eigenvector centrality	(Bonacich, 2007)
	Katz centrality	(Katz, 1953), (Landherr et al., 2010), (Coppersmith & Winograd, 1990)
	PageRank	(Page et al., 1999), (Li et al., 2015b), (Richardson et al., 2003), (Chen et al., 2014a), (Guha et al., 2004), (Weng et al., 2010), (Tai et al., 2005), (Eom & Shepelyansky, 2015)
	LeaderRank	(Lü et al., 2011)
	Semilocal centrality	(Chen et al., 2012)
	Evidential semi-local centrality	(Gao et al., 2013)
	Coreness centrality	(Carmi et al., 2007), (Kitsak et al., 2010)
	ClusterRank	(Chen et al., 2013)
	Local structural centrality	(Gao et al., 2014)
Community importance		
	Embeddedness	(Newman & Girvan, 2004), (Lancichinetti et al., 2010)
	Relative commitment measure	(Parau et al., 2015)
	Principal Component Centrality	Ilyas and Radha (2011), (Ilyas et al., 2013), (Ruhnau, 2000)
	Community disruption	(Parau et al., 2015)
	Kernels detection	(Wang et al., 2011b)

3.5. Data mining and machine learning approaches

The main purpose of the methods in data mining field is to unveil underlying knowledge burried in the massive data, which aims to reveal meaningful and useful information (e.g. patterns, relations) in big volumes of data. This field includes variety of tools from statistics and artificial intelligence. Recent advancement of data mining techniques has broadened its applicability in various fields in order to get an interesting insight from data. The approaches in this category include frequent pattern mining techniques, clustering methods, and some advanced learning models. The data mining proposes solutions for mining the graph-based data and further analysis of these structured data. For opinion leader detection, the pattern discovery approach relies on the flow of information to identify the most influential nodes by a list of actions performed by the node and the social network topology itself. Here, clustering refers to the task of grouping the nodes of a graph into some clusters by taking into this condition that there should be many connections within each cluster and relatively few between the clusters. The most sophisticated approaches rely on models that pretend to have the ability of a *learning machine* to perform accurately on unseen examples after having experience on a training (learning) data. In this section, we would discover and describe existing data mining solutions for opinion leader detection problem.

3.5.1. Pattern mining

Many real-life applications consider the common pattern mining task, including *association rules* and *frequent sequences*. *Pattern mining* provides various approaches in data mining to reveal not obvious and useful patterns in databases, e.g. top or rare patterns with some level of confidence threshold. The common pattern mining algorithms are used over the known and precise content of traditional static transaction databases, strings, graphs, etc.

The pattern mining is used for opinion leader detection in a particular social network with the assumption, that actions performed by a user are visible to their friends network and could affect their actions. The undirected network of a social network is always associated with transactions, s.t. list of actions performed by nodes. This way it is possible to build a directed acyclic *propagation graph* (Goyal, Bonchi, & Lakshmanan, 2008) in order to integrate with *frequent pattern mining* algorithm. The main idea of frequent itemset (pattern) mining is to discover top patterns with a high confidence (Borgelt, 2012). *GuruMine* is a pattern mining system for discovering leaders (Goyal, On, Bonchi, & Lakshmanan, 2009) that is provided with a graphical user interface. The framework implements the method, proposed in Goyal et al. (2008). The algorithm considers only the influence of a user by an action to be repeated over the local friend network. Therefore, leadership level is defined in the way of an accumulated number of reachable nodes, from an initial user, $|\mathcal{M}_{\mathcal{T}}(u,\alpha)|$ that perform the same action α as an initial user u within a time frame $\mathcal{T}$. To describe the opinion leadership quality, they consider *confidence* and *genuineness* measures. More precisely about the metrics for a user $v \in V$, let $P(v)$ denote the set of actions performed by user v and $L(v)$ denote set actions the leader of which is user v. Then the leadership confidence of v is the ratio $conf(v) = |L(v)|/|P(v)|$. User v is said to be a confidence leader if it is a leader and $conf(v)$ is bigger than a given threshold. In order to describe a genuine opinion leader, they use genuine leader metric proposed in Eq. (27). In order to define a genuine leader they provide the genuineness score of v and a threshold for it.

$$gen(v) = \frac{|\{\alpha \in L(v) | \nexists u \in V : u \text{ is leader for } \alpha \wedge v \in \mathcal{M}_{\mathcal{T}}(u,\alpha)\}|}{|L(v)|} \tag{27}$$

In Tsai, Tzeng, Lin, and Chen (2014) authors propose a method to accumulate the influence between users in a probabilistic way in order to detect community leaders in a social network considering that every user has a snipped of a time-related sequence of actions. The probabilistic time-based graph propagation model, that they build initially as the basis for the algorithm,

uses an exponential decay function to model the influence between users. They apply *apriori probabilistic path mining algorithm* (Agrawal, Srikant et al., 1994) to mine *influence chains* by pruning exhaustive information. Finally, community leaders are users whose their initiated number of influence chains are more than the number of influence chains in which they are involved.

3.5.2. Clustering

Clustering is a method of *unsupervised learning* which means the classes are not known in prior to clustering. Cluster analysis is the task of assigning observations into subsets called clusters so that objects inside the same cluster are similar based on some predefined criterion or criteria.

The most widely used clustering algorithm is *k-means* (MacQueen et al., 1967). The K-means algorithm is used to group the entities with the similar properties, e.g. opinions. The property set is represented as a vector and a distance metric is used for similarity measurement, e.g. *Euclidean distance*. The parameter '*k*' is a predefined number of final groups, the formation of which is started from a randomly chosen *centroids*, that is the space central element of a cluster. In each iteration, all the entities are assigned to a group with respect to closest centroid and the centroid is recomputed. The process goes until the convergence of the square error for every cluster, that is computed between entities and its respective centroids. Finally, the algorithm comes up with *k* clusters, where each cluster has data points that are as close as possible, and points in different clusters have much difference as possible.

The clustering methods are used in combination with simple metrics or as a step of a sophisticated framework for opinion leader detection. Cardente propose to use k-means algorithm to detect clusters with similar centrality properties. His hypothesis states that top innovators of collaboration network have abnormal centrality attributes, that would be collected in one or more clusters. The k-means algorithm also was used in faction groups to detect faction opinion leaders (Arvapally, Liu, & Jiang, 2012).

A more sophisticated method was implemented in *TCOL-Miner* framework (Chen et al., 2017a; Chen, Cheng, & Hsu, 2016b), that discovers opinion leaders in three consequently executed components (steps). The first-stage detects *significant communities* on the basis of topological structure till the *modularity gain* metric (Chen, Zhu, Peng, Lee, & Lee, 2014b) of formed communities becomes negative, finally, some small communities are pruned resulting in some average or big node groups. The second stage applies k-mean clustering to build the candidate set for the role of opinion leader, using four parameters retrieved for every resulting user from the first stage. It takes into account the total number of published articles by a user, a probability for a reply to any of the user's article, *expert degree* that is a probability of all the posts of a user to be in one domain, and a probability of the user's reply other users. In the final stage, every cluster from the previous step is scored and the resulting *k* opinion leaders are selected from the high-score clusters. TCOL-Miner effectively solves the influence overlapping problem by first two stages, as a result, it permits the method to detect influential opinion leaders.

The proposed framework in Duan, Zeng, and Luo (2014) for opinion leader detection has an application over stock message boards, where the authors are aiming to detect active opinion leaders that suggest an actual stock price direction in the forum. The main idea states that the number of opinion leaders is small in any social network or forum (Tang, Lou, & Kleinberg, 2012). Hence, the proposed method generates clusters over specific user features from message boards (e.g. the length of a sentence, number of replies and others), where any of traditional clustering algorithm could be applied, for example k-mean or *expectation-maximization-based clustering*. Finally, the framework selects top *k* groups as the candidate opinion leaders from the sorted sequence of clusters w.r.t calculated *index value*. As the last step for opinion leader detection from the candidate list, they provide correlation analysis over the actual stock price movement and the associated *sentiment* of a candidate, resulting in true opinion leaders as a strong correlation holder.

The *longitudinal user centered influence* model (LUCI) (Shafiq, Ilyas, Liu, & Radha, 2013) is a framework for opinion leader detection in a social communicating environment as forums or social network posts, etc. Moreover, the framework is able to classify resulting leader to introvert and extrovert one. The data format of user interaction is represented as triplets: the timestamp of interaction, the senders and receivers identifiers. The method itself relies on a system of equations for every user that is based on a generalization of the *Friedkin-Johnsen* influence model (Friedkin & Johnsen, 1997). Having $t_{\max} - 1$ equations, where $t_{\max}$ is the maximum activity period for each user, it computes a tuple of the *ego coefficient* ρ and the *network coefficient* γ by optimizing for the least square error, that is actually a linear regression. Finally the *kernel k-means* clustering algorithm (Dhillon, Guan, & Kulis, 2004) is used on (ρ, γ) features for each user. The resulting four clusters occupy distinct regions in the $\rho - \gamma$ plane: extrovert and introvert opinion leaders, followers and neutral clusters. The resulting clusters are supported with experiments over Facebook and 'Everything2' data.

3.5.3. Learning models

Learning models are inferred as an ability of a machine to learn from training data and accurately predict the unseen test data. Learning models could be applied to the social network to analyze how a produced content is distributed and shared across a social media, also it is used for sentiment analysis over a topic model, etc. In general, *machine learning* problems based on their goals are considered as classification, prediction or modeling. Based on availability of labels the learning techniques can group as supervised learning that the inputs and associated target values are provided, *unsupervised learning* that does not rely on labeling of the training set and *semi-supervised learning* that in this case, the training set includes both labeled and unlabeled data and the model is trained based on the available labeled class. There are some other learning methods such as *reinforcement learning* which the learning agent try to maximize the notion of cumulative reward by interacting with its environment.

There was proposed an unsupervised learning model QUOTUS (Niculae, Suen, Zhang, Danescu-Niculescu-Mizil, & Leskovec, 2015) to identify patterns in the outlet-to-quote *bipartite graph* which helps to understand the overview of the most shared content and therefore which authors of the content could be considered as the most influential or even as the opinion leader. Their model was trained based on the speeches data collected from 6 years of Barack Obama and how these speeches were covered in the media. In order to reduce the *false positive* rate in the model application for opinion leader identification, there is a community boundary which is used to reduce decision space.

The content of online social review platforms is mostly used to detect the general sentiment direction of a product reviewed. For opinion leader detection problem the typical supervised learning model was extended to *sparse overlapping user lasso* learning model (SOUL) (Lo, Tang, Li, & Yin, 2015) to jointly learn sentiment, keywords and opinion leaders in one process over the social reviews content. Their idea is motivated by such fact that opinion leaders are more knowledgeable and share important contents than the ordinary users. For such extended task, SOUL model has numerous hyper-parameters to tune for solving the objective function. Hence, the authors use *alternating direction method of multipliers* (ADMM) for sparse modeling with multiple hyper-parameters. The main property of ADMM (Goldstein & Osher, 2009) is to break

down the optimization problem into some sub-optimization problems (Bamakan, Wang, & Shi, 2017). In this model, the best result can obtain when reviews as inputs are linked to the same *social objects* (the target of review).

3.5.4. Summary

As mentioned in this section, the main purpose of the methods in data mining field is a knowledge discovery that aims to reveal meaningful and useful information (e.g. patterns, relations) in big volumes of data. For opinion leader detection we explicitly described the way of data mining application. Most of the opinion leader detection frameworks integrate clustering techniques as a filtering, dimension reduction or an outlier detection component before decision-making process. The most sophisticated approaches rely on models that pretend to have the ability of a *learning machine* to predict accurately new test examples after training from data. Nonetheless, most of the learning model application applied to the opinion leader detection problem suffers from lacking labeled data to study, which makes application of these methods inefficient in practical cases.

3.6. Hybrid content mining approaches

The online civic participation has a positive association with social networks, as for example Twitter content is used for opinion leader detection in the usual social environment, e.g. a college student in a particular city or country (Park & Kaye, 2017). In an online social network, written contents, writing styles and the sentiment orientations are the major resources for analyzing the behavior of users (Ho et al., 2016; Li & Du, 2011). Furthermore, the semantic information of document content is used to make topic models. Content mining and topic modeling are the two widely used techniques for various social analyzing tasks since they propose valuable insights about every node in a network. The topic model strategies for opinion leader detection mostly rely on the extracted structures and content information of networks. Moreover, the opinion leadership is considered as topic dependent, that means the contents are used to exploit topic information and explicit user features such as post time or sentiment, etc. Although there is not a comprehensive strategy to identify the opinion leaders by content and topic modeling, these approaches are still useful for detecting *monomorphic opinion leaders* (Richmond, 1980), those with specialty and considerable knowledge within single-topic areas.

3.6.1. Pagerank over a topic model

The ranking models are used in the task of opinion leader detection, e.g. the SPEAR algorithm (Shinde & Girase, 2016) involves ranking of online users with respect to their knowledge and expertise using the *HITS* algorithm (Kleinberg, 1998). The PageRank (Page et al., 1999) is an efficient webpage ranking algorithm which is proposed to evaluate and determine the importance of web pages (Section 3.4.1.1). That is why it has variations in the field of opinion leader detection and especially in combination with topic models, e.g. InfluenceRank (Song et al., 2007), OpinionRank (Zhou & Zeng, 2009), Dynamic OpinionRank (Huang et al., 2014; Song et al., 2011), TopicSimilarRank (Wang, Du, & Tang, 2016) and others.

The blogosphere is an available and informative channel to understand users reaction to events and theirs response to a specific promotion. A blog contains explicit content features and network structure of reposts or replies, which can be utilized to model the opinion leader detection problem. The authors in Song et al. (2007) proposed an algorithm called *InfluenceRank*, to distinguish opinion leaders among blogs. This algorithm considers the blog centrality over blogosphere network structure and its information novelty since these properties are the most definitive for true opinion leader detection. To examine the innovation of a blogs entries and afterward the information novelty of each blog, they use a cosine similarity measure over feature vectors obtained by *Latent Dirichlet Allocation* on blogs content (Blei, Ng, & Jordan, 2003). The novelty score obtioned by InfluenceRank is controlled by a parameter, in such a way that zero reflects the lack of innovation in a blog. Thus the performance of this algorithm will be the same as PageRank. The performance comparison has been done in terms of coverage, diversity, and distortion shows the promising results on a blog dataset (Song et al., 2007). The main shortcoming of InfluenceRank is that it does not take into account the impact of posting time. Algorithms which developed based on PageRank for opinion leader detection mostly differ by the network on which they are applied, e.g. *TwitterRank* (Weng et al., 2010) is applied to follower network with the topic similarity weight formation to detect influential posts. The *TrustRank* (Chen et al., 2014a) is proposed to construct a network based on the results of direct and indirect links labeling by sentiment analysis and social balance theory. Such networks are constructed via four phases (Chen et al., 2014a): first, construct a basic weighted post network by considering explicit and implicit links, where explicit link is a reply or a citation; and implicit link is a semantic similarity of posts or comments. Second the sign of explicit links should be labeled by sentiment orientation, afterward the sign of implicit links are inferred by applying structural balance theory and the last phase is to transform the signed post network into a signed user network.

The concept of *opinion networks* was proposed in Zhou and Zeng (2009) with *OpinionRank* algorithm that makes ranking the nodes in the opinion network. The opinion network is a directed graph, where a vertex is a member of a community and an edge is defined as an opinion orientation from influencer to influence. Here, the set of opinion scores is computed by the edges. As mentioned by Zhou and Zeng (2009), OpinionRank adapted PageRank algorithm to rank the nodes based on their opinion scores. The experimental evaluation reveals that OpinionRank results are nearly similar to PageRank (Huang et al., 2014; Song et al., 2011). It can be inferred that user network and comment network can be modeled by considering information from sentiment orientation analysis, opinion similarity calculation, explicit and implicit link mining. In this case, the implicit link is a similarity of weights equal to *tf-idf* (Salton & Buckley, 1988). Moreover, the comment network considers time interval between comments as a weight between them. By applying PageRank on the comment network, the authors identify opinion leader's comments. Furthermore, the most persuasive user in the network is selected by a clustering algorithm *DBSCAN* (Ester, Kriegel, Sander, Xu et al., 1996; Lin & Han, 2015) regard to internal relations between the opinion leader comments with others users, and the scoring grade of the comments. The results evaluation shows a significant improvement of OpinionRank but still depends a lot on accurate *natural language processing* techniques applied and implicit connections formation methods.

3.6.2. Ontology-based extraction methods

Public tags are given to online items by users, to make it easy refinding those items. Such systems are known as folksonomy (Vander Wal, 2007). In Szomszor, Alani, Cantador, OHara, and Shadbolt (2008), authors developed *user profile ontology* on the basis of a user's distributed tags and Wikipedia. The results obtained by this model showed the ability of discovering 15 new concepts of interest in average by taking into account the user's tag cloud. While folksonomies are created by a user, ontologies are created by experts that make it an enabling technology for the Semantic Web (Gruber, 1993). It would be helpful for people to state what they mean by the terms used in the data they share. In other words, on-

tology is basically a taxonomy that represents entities, ideas, and events, along with their properties and relations, according to a system of categories.

For opinion leader identification in the social blogosphere, there was proposed an ontology-based BARR framework (Li & Du, 2011). The framework builds an ontology for a marketing product to identify opinion leaders by analyzing users' profiles, blog contents, and links between authors of blogs and their readers. For the relationship, they consider both *homophily* (Lazarsfeld, Merton et al., 1954) and tie strength between the sources of information and who are looking for them. The homophily and the tie strength are represents of the degree of similarity of communicators and the closeness of the parties involved, respectively. The authors in Gilly, Graham, Wolfinbarger, and Yale (1998) state that the greater the homophily between communicators may lead to the appearance of the more persuasive communication. In this work, both quality and quantity of blogs and its content are considered. Here, supposed that a blogger with only a few posts cannot be considered as an influential one (Li & Du, 2011). The expertise of an author is computed as the cosine between the *term frequency vectors* obtained for an author, by considering all his blogs, and for the extracted ontology of a marketing target product. Choosing a hot-blog is a multi-attribute decision making problem (Li & Du, 2011), in such a way that the framework creates a complex topic model that consists of several factors such as blog portals, keywords and web sources. The proposed method uses the technique of order preference by similarity to ideal solution (TOPSIS) (Yoon & Hwang, 1981) to finally express the popularity of a blog and the influence of an author.

The framework is relatively comprehensive since it needs to construct the ontology that can be quite difficult and demanding, nonetheless, it neglects "novelty" characteristic as the timestamp that is a crucial feature for detecting opinion leader on time-dependent information environment.

3.6.3. Mixed frameworks

Opinion leaders with considerable knowledge within single-topic areas are quite hard to reveal. Hence, some researchers try to benefit from the advantages of more than one specific model by constructing combined models. For example, authors in Li, Li, and Zhu (2016) proposed *global consistency maximization (GCM)* as a link-based classification model to classify the opposite opinions in the Twitter network. *SuperedgeRank* (Ma & Liu, 2014) is another mixed framework to find the influential users based on *supernetwork theory*, that is composed of network topology analysis and text mining.

Moreover, the authors in Li et al. (2013) developed ranking framework ENIA to automatically identify topic-specific opinion leaders. The score for opinion leadership is computed from four measures include expertise, novelty, influence, and activity. As it is presented on Fig. 4 the framework uses various sources to compute a score for each of these factors, e.g. textual content, temporal information and observed user behavior.

The *users expertise* on a specific topic considers the user's interest which has direct relations to a number of topics or comments for a particular topic. To construct such metrics they use Latent Dirichlet Allocation (Blei et al., 2003) and TOPSIS (Yoon & Hwang, 1981). The next indicator is *novelty*, that is a time-sensitive indicator. First of all, they find the most similar document by cosine over tf-idf space and the temporal order of similar documents with the same topic. Additionally, the framework considers a case when a user share a lot of documents from other users, which can be inferred these documents are not novel. The next indicator is *influence* which *viewing* and *reply influence* are its sub-indicators. Viewing reveals the popularity of a document as a relative number of readers in a community, and the reply-to is regarded as a number of replied users for a document. The *activity* is defined as a number of user's documents and replies. Here, *longevity* that is a temporal factor presents the concentration stability of users on a specific subject and *centrality* are used to show the topic-focus degree of opinion leaders.

3.6.4. Summary

To identify opinion leaders, a number of studies use a wide range of available information from blog content, authors, readers, their relationships, etc. Mostly such heavy frameworks are supported by online learning, considering the global information of the whole forum or others social platforms. The obvious advantage of *content based* approaches is the topic-specific opinion leader search. The methods from that group mostly work with high-dimensional data. The hybrid content mining approaches require massive data in order to achieve high effectiveness of opinion leader search.

4. Public datasets

This section presents a number of available real-world datasets, which have been used to evaluate existing methods. To be noted, preparing a dataset for the task of opinion leader detection or identifying the most influential users/nodes strictly depends on the problem definition, the availability of personalized information, the properties of user friendship networks and so forth. Therefore, the most obvious way of acquiring appropriate dataset is to collect datasets for specific uses (Aleahmad, Karisani, Rahgozar, & Oroumchian, 2016; Chen, Hui, Wu, Liu, & Chen, 2017b; Jiang, Ge, Xiao, & Gao, 2013; Song et al., 2007; Weng et al., 2010). The most commonly used social networks that are used for crawling purpose include Twitter as a microblogging service and news media, Flickr as a photo sharing website, YouTube as a video sharing website and Amazon as an electronic commerce website, etc. Although building a crawler gives more flexibility in acquiring datasets, it is time-consuming and it needs preprocessing even after downloading the datasets. Instead, some prefer to utilize public datasets that are easy to use despite limited availability. In Table 4, we summarize the available public datasets for opinion leader/influential user detection.

5. Discussion and challenges

During the last decade, opinion leader detection has been widely studied and there exist a plenty of methods and techniques with the promising findings that are proposed to deal with this problem from different perspectives and domains. It is not only shown the significant role of opinion leaders in formatting and shaping opinions of others, but also in accelerating, controlling and disseminating of information flow.

In order to provide a global scheme of studies carried out in this field, we built the author-collaboration network as presented in Fig. 5. To come up with graphs presented in Fig. 5 we firstly collected 261 works related to several problems as opinion leader detection, influential users, opinion spam detection and graph measure metrics. After filtration by predefined keywords, s.t. 'opinion', 'leader', 'influential', over the texts from abstracts and titles, we finally got 165 articles. In the Fig. 5a, the collaboration directed network is generated on the basis of 165 articles, the vertices uniquely represent authors and edges are the connections from a first author of an article to collaborator authors of the same article. This way, we could see the most collaborative researchers in the field of opinion leader, that is represented by the size of author name and its node in the figure. As the next step, we clustered this sparse graph to thematic clusters. The clustered graph was generated automatically by a Python script and text mining tools, to

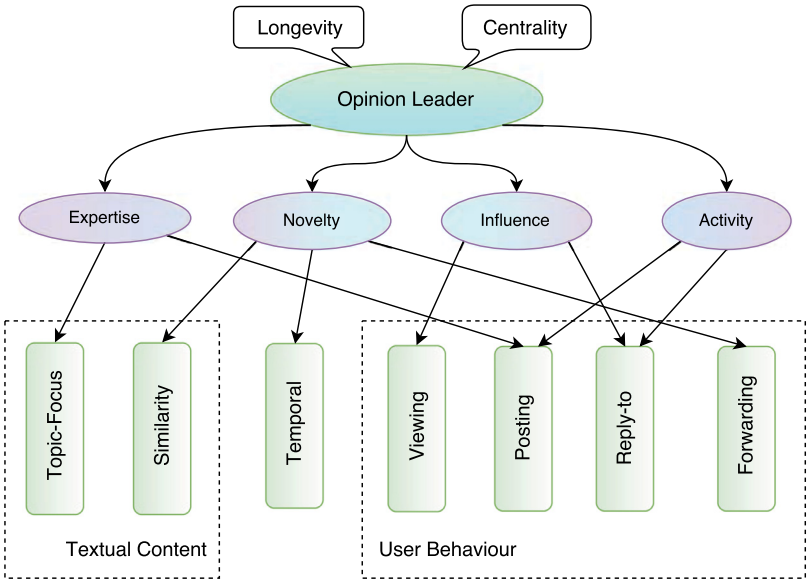


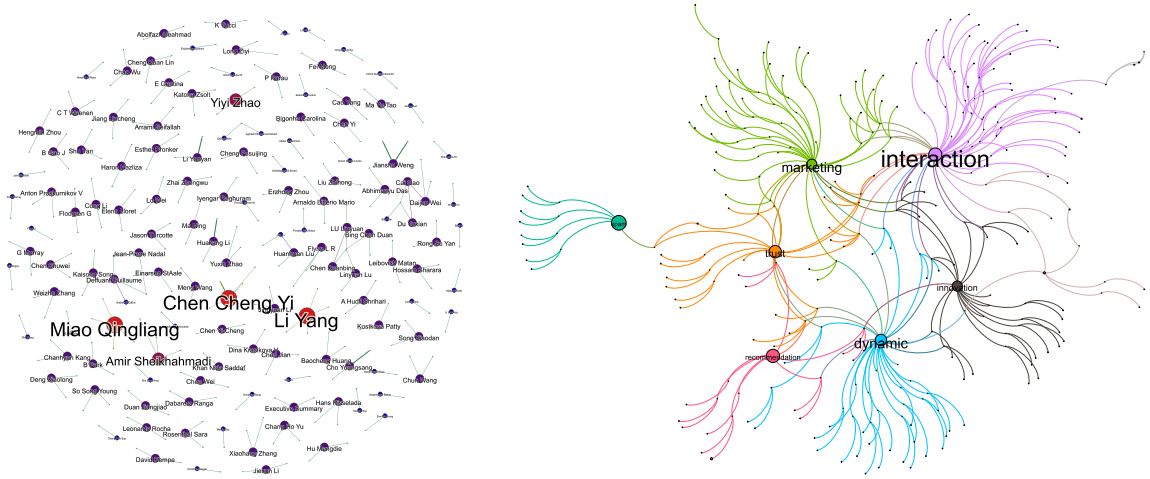
Fig. 4. The ENIA framework for opinion leader detection (Li et al., 2013). Adapted with permission.

Table 4
Public datasets for opinion leader detection .

Dataset	Domain	#Instances	Description	References
Higgs Twitter Dataset	Influential users detection	456,626 nodes	This dataset is a collection of tweets about the announcement of the discovery of a new particle between 1st and 7th July 2012. It contains the follow relation among users and user reaction to posted messages.	(De Domenico, Lima, Mougel, & Musolesi, 2013; Sheikhahmadi, Nematbakhsh, & Zareie, 2017)
ISIS Twitter Data	Influential node detection	17,000 tweets	This is a collection of tweets from 112 users posted between January 2015 and May 2016, used to identify influential nodes talking about ISIS.	(Li, Kailkhura, Thiagarajan, Zhang, & Varshney, 2017)
SNAP: SEISMIC	Information cascade modeling	166,076 tweets	This is a dataset of over 3.2 billion tweets and retweets on Twitter from October 7, to November 7, 2011. It includes the following information; tweet id, posting time, retweet time, and the number of followers of the poster/retweeter.	(Zhao et al., 2015)
RepLab 2013	Automotive, Banking, Universities, Music	142,527 tweets	This dataset contains tweets in regards to the reputation of entities e.g. companies, organizations, celebrities, etc. Each tweet is manually labeled based on Polarity, Centrality and Users authority.	(Amigó et al., 2013; Vilares, Hermo, Alonso, Gómez-Rodríguez, & Vilares, 2014)
Epinions social network	Trust Management	100,000 opinion between	This dataset contains the who-trust-whom relations in an online social network of Epinions.com	(Aghdam & Navimipour, 2016)
SNAP: LIM	Information diffusion	500 million tweets	This dataset is used to evaluate the performance of Linear Influence Model to model influences of nodes and to predict the temporal dynamics of information diffusion.	(Yang & Leskovec, 2010)
Enron emails and co-author publication datasets	Inferring social relationship	255,000 emails, 1 million authors and 80,000 papers	It includes three genres of real-world data sets: Publication (coauthor network of Arnetminer), Email (email network of Enron employees), Mobile (mobile network of Reality Mining Project).	(Jaber, Wood, Papapetrou, & Helmer, 2014; Tang, Zhuang, & Tang, 2011)
OpinRank Dataset	OpinRank Dataset	-	This dataset contains the reviews and ratings of different aspects of cars and hotels available on Edmunds.com and Tripadvisor.com	(Ganesan & Zhai, 2012)

select the main theme from abstracts, and cluster first author vertices by seven categories (Fig. 5b) including: 1) trust and credibility, 2) troll and spam opinions, 3) diffusion of information, 4) e-commerce and marketing, 5) dynamic modeling, 6) recommendation systems, 7) innovation. As a result, we got one big component and some smaller ones, which were removed. Finally, over the biggest component we executed modularity (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to reveal thematic groups consisting of vertices that are more densely connected together than to the rest of the network. By analysing modularity groups we could see that, authors mostly concentrate on agent-based systems or dynamic models to apply them for marketing and social economic modeling. Additionally to it, among researches there is an interest in trust issue, e.g. opinion spam detection and recommendations credibility.

Summing up the results, it is concluded that we have seen a satisfactory growth in the development of opinion leader detection techniques and methodologies with variety of pros and cons. However, deciding which method shows a better performance and in different situations which group of approaches are more suitable, are important issues that need to be discussed.



(a) Authors collaboration network on the basis of their articles. (b) Modularity thematic groups over collaboration network.

Fig. 5. Contribution networks on the topic of opinion leader detection.

In social and life science studies, which mostly the researchers are looking for high influential users in societal problems theme, such as accelerating the behavior changes in a groups, increasing the possibility of adoption of a new norms or help to spread of a specific medical treatment, utilizing the descriptive and statistical methods are more prevent. However, these groups of methods suffer from the scalability problem that to overcome this drawback and simultaneously take the advantages of rich information available in social networks, utilizing the topological measures and learning methods become more applicable. As mentioned, social networks contain the rich and high dimensional data such as textual data, the interactions networks topologies, the position of individuals within a network, the degree of importance of each node or user in a graph, the quantitative measure of followers and following, etc. But the problem with topological approaches is that they do not consider more than only network structure to determine the centrality of nodes. For example, modeling based on the number of user links could result to be easily manipulated by sharing too many links from others and to detect fake opinion leaders. On the other side, lacking the gold-standard datasets is the main limitation in machine learning approaches. To overcome the aforementioned limitation and drawbacks, hybrid content mining approaches consider the topological properties of a network in conjunction with the textual content, writing styles and the sentiment orientations of posted opinions. Therefore, this group of methods has superior performance in dealing with high-dimensional data, by combining text mining techniques and social interactions.

In general, opinion leader detection is a process of some phases, including domain understanding, specify the available features and data collection, data cleaning and transforming, network structuring, modeling and visualizing the data, finally identifying key influential individuals and performance evaluation. In each of these phases, there are important factors and challenges that precisely evaluating them will certainly increase the total effectiveness of models. Factors such as providing an exact definition of opinion leader in a certain context, their intrinsic and extrinsic characteristics and the topological properties of their friendship network, all together make the opinion leader detection as a difficult and challenging task (Potolea, 2016).

Despite of recent advances in opinion leader detection methodologies, several other questions remain to be addressed as challenges and future trends.

It is worth mentioning that scoring the opinion leaders based on the level of efficacy and the scope of influence (local or global) in social networks are an interesting and challenging topics for the researchers in the area of advertising and marketing (Li & Du, 2011). The degree of influence of opinion leaders is varying in different communities in regards to the degrees of their commitment and proficiency. Furthermore, the score of opinion leaders is a dynamic subject based on users prestige and activity, their follower-following network's structure, the flow of influence and the measures used to identify opinion leaders. As discussed in Sections 3.4.1 and 3.4.2 plenty of centrality and community measures are proposed that each of them has their own advantages and drawbacks, so it yields to the different scores for the identified leaders. In order to, further analyze and compare the results of aforementioned measures, researchers take the advantages of multiple-criteria decision-making (MCDM) or multiple-criteria decision analysis (MCDA), to explicitly evaluate multiple conflicting criteria in ranking the key opinion leaders (Du, Gao, Hu, Mahadevan, & Deng, 2014; Jing & Lizhen, 2014; Li et al., 2013). MADM methods are considered as practical decision-making methods, which combine the qualitative analysis with the quantitative assay. These methods are based on the hierarchies of different alternatives and their relative importance. Among the MCDM methods, a technique for order preference by similarity to ideal solution (TOPSIS) has been used by Li and Du (2011) and Gao et al. (2013), to identifying the influential nodes in complex networks and the authors in Jing and Lizhen (2014) used analytic hierarchy process (Saaty, 1988) to detecting opinion leaders in microblog networks. As a new research direction, it would be worthwhile to study in depth the application of MCDM methods in rating and identifying opinion leaders.

Examining the trustworthiness and credibility of opinion leaders is another challenging issue. The level of trust that members have with each other can be considered as a fundamental element of any social community formation. Although, trust is considered as a multidisciplinary concept and a plenty of definition is proposed based on the context of examined fields, generally it is defined as a measure of confidence that members would behave in an expected manner (Sherchan, Nepal, & Paris, 2013). A complete survey of trust in social networks is conducted by Sherchan et al. (2013). Based on Sherchan et al. (2013), there are different facets of trust such as calculative, relational, emotional, cognitive and so on. Among them, the relational aspect of trust is more applicable to the context of social networks, where trust is built based on the number of interactions between the opinion leaders and followers. In fact, reliability and dependability in

the previous interactions are the two important factors, which strengthens the trust between two parties (Sherchan et al., 2013). In online systems two types of trust, including direct trust and recommendation trust are defined, which the former is based on the direct experience of two parties and the latter is based on the experiences of other members in the community with the other party and follows the propagative property of the trust. From our point of view, both of them are applicable in investigating the opinion leader and followers relationship. The level of trust between the influencer and the influencee are tightly related to the bounded confidence concept (Chiregi & Navimipour, 2016), that means as long as the confidence levels of followers in a social networks are sufficiently high, the opinion leaders have a sufficient power to influence and change their opinions and thoughts (Zhao et al., 2018). It should be noted that the drawbacks of centrality and community measures are that they do not take into account the credibility of influential nodes. Despite the importance of analyzing the level of confidence between the opinion leaders and followers, few studies (Chiregi & Navimipour, 2016; Jiang, Wang, & Wu, 2014; Kim & Tran, 2013; Yan, Zheng, Wang, Song, & Zhang, 2015) are conducted that further research in this issue would be of interest.

Another challenge is to distinguish between the real opinion leaders and the trolls in the social networks. By growing spread of social networks uncountable information, reviews, and opinions are available to the users, but there is a critical question on their accuracy and integrity. Unfortunately, the possibility of compromising the integrity of information and disseminating misinformation throughout the social networks is increased by the presence of persons who deliberately post incorrect, destructive, provocative or unreal comments and opinions that are called “trolls” (Chiregi & Navimipour, 2016; Coles & West, 2016; Kumar, Spezzano, & Subrahmanian, 2014). Even though the authors in Al-Oufi, Kim, and El Saddik (2012), presented a method based on the *Advogato trust metric* and the *capacity-first maximum* to identify and rank the trustworthy users in a general case. Their method would be applicable to troll detection by identifying the valid influential users and rank them based on their trustworthiness and restrict the access of those who are going to illicitly gain high reputation in the community. The author in Ortega, Troyano, Cruz, Vallejo, and EnríQuez (2012) proposed a method based on the foundation of PolarityRank named *“PolarityTrust”*. In this method, by taking the advantages of two mechanisms, including *Non-Negative Propagation* and *Action-Reaction Propagation*, they prevented the propagation of bad ideas in the networks and the users with dishonest behavior are penalized in the ranking of trust. The existing challenges in providing a robust trust and reputation system against the possible manipulation and attacks, also the necessary requirements of such systems are discussed in Jøsang and Golbeck (2009).

In addition to the credibility of influential users, the integrity and correctness of posted comments, reviews, and opinions are the crucial factors in social network analysis. As already discussed, social networks provide a platform, that facilitates the interactions of people in the society to easily express their personal experiences, opinions, emotions, and feelings regard to a specific product, service or even in a political sphere and economic issues (Heydari, ali Tavakoli, Salim, & Heydari, 2015). Since, this information is utilized to promote or demote an object consequently yield financial gains or cause monetary loss, respectively. Therefore, dissemination of information in social networks is an attractive target for opinion fraud or spam opinion (Jindal & Liu, 2008). By considering the existing literature, it is obviously felt a research gap between the validity of opinion leaders and the correctness of their opinions (Choo, Yu, & Chi, 2015; Ott, Cardie, & Hancock, 2013; Wang, Xie, Liu, & Philip, 2011a).

At the end, it should be noted that there are some common challenges for the task of opinion leader detection that rooted in the graph mining difficulties. Scalability is one of the main challenges that comes up with analyzing the large-scale networks and it causes high computational complexity and high cost of graph visualization (Pienta, Abello, Kahng, & Chau, 2015). Therefore, as far as we are going to analyze big graphs to find opinion leaders, scalability of applied algorithms should be considered. Graph sampling and clustering are frequently used methods to address scalability issues (Leskovec & Faloutsos, 2006; Wu et al., 2017). Although a wide range of sampling techniques from simple node-based to advanced traversal-based schemes have been proposed, the problem is that none of them can preserve all the topological properties of the original network (Zhang et al., 2016; Pienta et al., 2015; Wu et al., 2017). The authors in Wu et al. (2017), evaluate the performance of the five mainly used sampling strategies and the authors in Zhang et al., 2016 evaluate the measures of cluster quality which can be used as a guideline for preparing a dataset for opinion leader detection.

6. Conclusion and future work

During the past decades, there has been a substantial growth of interest in understanding the opinion formation mechanisms in social networks and the role of opinion leaders in shaping the opinions of others. It was the main purpose of this paper to draw attention to those approaches that contribute to detection of opinion leaders in different ways. At first, the characteristics and roles of opinion leaders from different domains were discussed. Moreover, based on the available data and the properties of approaches, we categorized them into six groups, including descriptive approaches, statistical and stochastic, diffusion process based approaches, topological based methods, learning approaches and finally hybrid content mining. Although many measures and techniques have been proposed to identify opinion leaders, each of them has its own strengths and weaknesses, which were discussed in details at the end of each section. Furthermore, we systematically summarized and compared the contributions of studies with a critical perspective and defined the possible challenges and future research directions.

With the proliferation of various available data and rich features in describing the diffusion of information, finding opinion leaders in societal systems has been becoming challenging for existing studies, especially, in consideration of the complexity and dynamics of such systems. Further efforts are thus needed to be invested to full-fledged the theory, techniques and applications of opinion leader detection. The complexity of system is originated from the availability of high dimensional data resources such as, rich contextual information, the networks topology, available meta-data of users, large-scale graphs, and the stream nature of social media data. Uncertainty on information and behavior diffusion in social networks is another factor that affect the complexity of a system. The definition of uncertainty is mostly considered by understanding the studied fields. Uncertainty in case of opinion leader detection comes from leaders incomplete knowledge about topics or events, the inherently stochastic property in leaders behavior, environmental uncertainties as an external unpredictable variation and so on, which bring the risk and fuzziness to this problem. It should be noted that in real-life applications, social networks are function of time, which bring the concept of dynamicity to the analysis of opinion leaders behavior. It means that an opinion leader might be no longer be a leader by evolving the networks with time. Hence, understanding the dynamic exchanges between participants of an interaction network by taking into account the valuable information driven from spatio-temporal data, would be an insightful direction to identify the local and global opinion leaders.

Acknowledgement

This work was partially supported by the CAS Pioneer Hundred Talents Program, China [grant number Y84402, 2017], and the China Postdoctoral Science Foundation Grant, China [grant number 2018M633186, 2018].

References

Aghdam, S. M., & Navimipour, N. J. (2016). Opinion leaders selection in the social networks based on trust relationships propagation. *Karbala International Journal of Modern Science, 2*(2), 88–97.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, vldb: 1215* (pp. 487–499).

Al-Oufi, S., Kim, H.-N., & El Saddik, A. (2012). A group trust metric for identifying people of trust in online social networks. *Expert Systems with Applications, 39*(18), 13173–13181.

Aleahmad, A., Karisani, P., Rahgozar, M., & Oroumchian, F. (2016). Olfinder: Finding opinion leaders in online social networks. *Journal of Information Science, 42*(5), 659–674.

Amigó, E., De Albornoz, J. C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., et al. (2013). Overview of replab 2013: Evaluating online reputation monitoring systems. In *International conference of the cross-language evaluation forum for european languages* (pp. 333–352). Springer.

Amor, B. R., Vuik, S. I., Callahan, R., Darzi, A., Yaliraki, S. N., & Barahona, M. (2016). Community detection and role identification in directed networks: Understanding the twitter network of the care. data debate. In *Dynamic networks and cyber-security* (pp. 111–136). World Scientific.

Arrami, S., Oueslati, W., & Akaichi, J. (2017). Detection of opinion leaders in social networks: A survey. In *International conference on intelligent interactive multimedia systems and services* (pp. 362–370). Springer.

Arvapally, R. S., Liu, X., & Jiang, W. (2012). Identification of faction groups and leaders in web-based intelligent argumentation system for collaborative decision support. In *Collaboration technologies and systems (cts), 2012 international conference on* (pp. 509–516). IEEE.

Bamakan, S. M. H., Wang, H., & Shi, Y. (2017). Ramp loss k-support vector classification-regression; a robust and sparse multi-class approach to the intrusion detection problem. *Knowledge-Based Systems, 126*, 113–126.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy, 100*(5), 992–1026.

Bilici, E., & Saygın, Y. (2017). Why do people (not) like me?: Mining opinion influencing factors from reviews. *Expert Systems with Applications, 68*, 185–195.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bodendorf, F., & Kaiser, C. (2010). Detecting opinion leaders and trends in online communities. In *Digital society, 2010. icds'10. fourth international conference on* (pp. 124–129). IEEE.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks, 29*(4), 555–564.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 437–456.

Buchler, N., Rajivan, P., Marusich, L. R., Lightner, L., & Gonzalez, C. (2017). Sociometrics and observational assessment of teaming and leadership in a cyber security defense competition. *Computers & Security*.

Cardente, J. Using centrality measures to identify key members of an innovation collaboration network. Available at http://snap.stanford.edu/class/cs224w-2012/projects/cs224w-043-final.pdf

Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., & Shir, E. (2007). A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences, 104*(27), 11150–11154.

Carpenter, C. R., & Sherbino, J. (2010). How does an "opinion leader" influence my practice? *CJEM, 12*(5), 431.

Chan, K. K., & Misra, S. (1990). Characteristics of the opinion leader: A new dimension. *Journal of Advertising, 19*(3), 53–60.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR), 41*(3), 15.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1777–1787.

Chen, D.-B., Gao, H., Lü, L., & Zhou, T. (2013). Identifying influential nodes in large-scale directed networks: The role of clustering. *PloS One, 8*(10), e77455.

Chen, S., Glass, D. H., & McCartney, M. (2016a). Characteristics of successful opinion leaders in a bounded confidence model. *Physica A: Statistical Mechanics and its Applications, 449*, 426–436.

Chen, W., Yuan, Y., & Zhang, L. (2010). Scalable influence maximization in social networks under the linear threshold model. In *Data mining (icdm), 2010 ieee 10th international conference on* (pp. 88–97). IEEE.

Chen, Y., Wang, X., Tang, B., Xu, R., Yuan, B., Xiang, X., & Bu, J. (2014a). Identifying opinion leaders from online comments. In *Chinese national conference on social media processing* (pp. 231–239). Springer.

Chen, Y.-C., Chen, Y.-H., Hsu, C.-H., You, H.-J., Liu, J., & Huang, X. (2017a). Mining opinion leaders in big social network. In *Advanced information networking and applications (aina), 2017 ieee 31st international conference on* (pp. 1012–1018). IEEE.

Chen, Y.-C., Cheng, J.-Y., & Hsu, H. (2016b). A cluster-based opinion leader discovery in social network. In *Technologies and applications of artificial intelligence (taai), 2016 conference on* (pp. 78–83). IEEE.

Chen, Y.-C., Hui, L., Wu, C. I., Liu, H.-Y., & Chen, S.-C. (2017b). Opinion leaders discovery in dynamic social network. In *Ubi-media computing and workshops (ubi–media), 2017 10th international conference on* (pp. 1–6). IEEE.

Chen, Y.-C., Zhu, W.-Y., Peng, W.-C., Lee, W.-C., & Lee, S.-Y. (2014b). Cim: Community-based influence maximization in social networks. *ACM Transactions on Intelligent Systems and Technology (TIST), 5*(2), 25.

Childers, T. L. (1986). Assessment of the psychometric properties of an opinion leadership scale. *Journal of Marketing Research*, 184–188.

Chiregi, M., & Navimipour, N. J. (2016). A new method for trust and reputation evaluation in the cloud environments using the recommendations of opinion leaders' entities and removing the effect of troll entities. *Computers in Human Behavior, 60*, 280–292.

Cho, Y., Hwang, J., & Lee, D. (2012). Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach. *Technological Forecasting and Social Change, 79*(1), 97–106.

Choo, E., Yu, T., & Chi, M. (2015). Detecting opinion spammer groups through community discovery and sentiment analysis. In *Ifip annual conference on data and applications security and privacy* (pp. 170–187). Springer.

Chowdhry, K., & Newcomb, T. M. (1952). The relative abilities of leaders and non-leaders to estimate opinions of their own groups.. *The Journal of Abnormal and Social Psychology, 47*(1), 51.

Cialdini, R. B. (2001). *Influence: Science and practice*. Allyn & Bacon.

Coles, B. A., & West, M. (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior, 60*, 233–244.

Coppersmith, D., & Winograd, S. (1990). Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation, 9*(3), 251–280.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks, 25*(4), 283–307.

Dabarera, R., Premaratne, K., Murthi, M. N., & Sarkar, D. (2016). Consensus in the presence of multiple opinion leaders: Effect of bounded confidence. *IEEE Transactions on Signal and Information Processing over Networks, 2*(3), 336–349.

Das, A., Gollapudi, S., & Munagala, K. (2014). Modeling opinion dynamics in social networks. In *Proceedings of the 7th acm international conference on web search and data mining* (pp. 403–412). ACM.

De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific Reports, 3*, 2980.

Deffuant, G., Amblard, F., & Weisbuch, G. (2004). Modelling group opinion shift to extreme: The smooth bounded confidence model. online arXiv paper for open access, https://arxiv.org/abs/cond-mat/0410199.

Dekker, A. (2005). Conceptual distance in social network analysis. *Journal of Social Structure (JOSS), 6*.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 325–339.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel *k*-means: Spectral clustering and normalized cuts. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 551–556). ACM.

Du, Y., Gao, C., Hu, Y., Mahadevan, S., & Deng, Y. (2014). A new method of identifying influential nodes in complex networks based on topsis. *Physica A: Statistical Mechanics and its Applications, 399*, 57–69.

Duan, J., Zeng, J., & Luo, B. (2014). Identification of opinion leaders based on user clustering and sentiment analysis. In *Proceedings of the 2014 ieee/wic/acm international joint conferences on web intelligence (wi) and intelligent agent technologies (iat)-volume 01* (pp. 377–383). IEEE Computer Society.

Duan, W., Gu, B., & Whinston, A. B. (2009). Informational cascades and software adoption on the internet: An empirical investigation. *Mis Quarterly*, 23–48.

Earp, J. A., Eng, E., O'malley, M. S., Altpeter, M., Rauscher, G., Mayne, L., et al. (2002). Increasing use of mammography among older, rural african american women: Results from a community trial. *American Journal of Public Health, 92*(4), 646–654.

Einarsen, S., Aasland, M. S., & Skogstad, A. (2007). Destructive leadership behaviour: A definition and conceptual model. *The Leadership Quarterly, 18*(3), 207–216.

Eom, Y.-H., & Shepelyansky, D. L. (2015). Opinion formation driven by pagerank node influence on directed networks. *Physica A: Statistical Mechanics and its Applications, 436*, 707–715.

Erdogan, B. Z., Baker, M. J., & Tagg, S. (2001). Selecting celebrity endorsers: The practitioner's perspective. *Journal of advertising research, 41*(3), 39–48.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd: 96* (pp. 226–231).

Flodgren, G., Parmelli, E., Doumit, G., Gattellari, M., O'Brien, M. A., Grimshaw, J., & Eccles, M. P. (2011). Local opinion leaders: Effects on professional practice and health care outcomes. *The Cochrane Library*.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.

Friedkin, N. E., & Johnsen, E. C. (1997). Social positions in influence networks. *Social Networks, 19*(3), 209–222.

Gagniuc, P. A. (2017). *Markov chains: From theory to implementation and experimentation*. John Wiley & Sons.

Ganesan, K., & Zhai, C. (2012). Opinion-based entity ranking. *Information Retrieval, 15*(2), 116–150.

Gao, C., Wei, D., Hu, Y., Mahadevan, S., & Deng, Y. (2013). A modified evidential methodology of identifying influential nodes in weighted networks. *Physica A: Statistical Mechanics and its Applications, 392*(21), 5490–5500.

Gao, S., Ma, J., Chen, Z., Wang, G., & Xing, C. (2014). Ranking the spreading ability of nodes in complex networks based on local structure. *Physica A: Statistical Mechanics and its Applications, 403*, 130–147.

Gates, G., & Kennedy, S. (1989). Peer educators reach college students with nutrition information. *Journal of American College Health, 38*(2), 95–96.

Geijsel, F., Meijers, F., & Wardekker, W. (2007). Leading the process of reculturing: Roles and actions of school leaders. *The Australian Educational Researcher, 34*(3), 135–161.

Gentina, E., Kilic, D., & Dancoine, P.-F. (2017). Distinctive role of opinion leaders in the social networks of school adolescents: An investigation of e-cigarette use. *Public Health, 144*, 109–116.

Gilly, M. C., Graham, J. L., Wolfinbarger, M. F., & Yale, L. J. (1998). A dyadic study of interpersonal information search. *Journal of the Academy of Marketing Science, 26*(2), 83–100.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters, 12*(3), 211–223.

Goldstein, T., & Osher, S. (2009). The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences, 2*(2), 323–343.

Gotecha, M. R., & Patwardhan, M. S. (2016). Identification of key opinion leaders in healthcare domain using weighted social network analysis. In *Computing communication control and automation (iccubea), 2016 international conference on* (pp. 1–6). IEEE.

Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2008). Discovering leaders from community actions. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 499–508). ACM.

Goyal, A., On, B.-W., Bonchi, F., & Lakshmanan, L. V. (2009). Gurumine: A pattern mining system for discovering leaders and tribes. In *Data engineering, 2009. icde'09. ieee 25th international conference on* (pp. 1471–1474). IEEE.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420–1443.

Grissa, L. (2016). What makes opinion leaders share brand content on professional networking sites (eg linkedin, viadeo, xing, skilledafricans...). In *Digital economy (icdec), international conference on* (pp. 8–15). IEEE.

Gruber, T. (1993). What is an ontology. WWW Site http://www-ksl.stanford.edu/kst/whatis-an-ontology.html (accessed on 07-09-2004).

Guha, R., Kumar, R., Raghavan, P., & Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on world wide web* (pp. 403–412). ACM.

Guimera, R., & Amaral, L. A. N. (2005). Cartography of complex networks: Modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment, 2005*(02), P02001.

Guldbrandsson, K., Nordvik, M. K., & Bremberg, S. (2012). Identification of potential opinion leaders in child health promotion in sweden using network analysis. *BMC Research Notes, 5*(1), 424.

Haron, H., Johar, E. H., & Ramli, Z. F. (2016). Online opinion leaders and their influence on purchase intentions. In *e-learning, e-management and e-services (ic3e), 2016 ieee conference on* (pp. 162–165). IEEE.

Harrison, C., & Killion, J. (2007). Ten roles for teacher leaders. *Educational Leadership, 65*(1), 74.

Heydari, A., ali Tavakoli, M., Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications, 42*(7), 3634–3642.

Ho, Y.-C., Liu, H.-M., Hsu, H.-H., Lin, C.-H., Ho, Y.-H., & Chen, L.-J. (2016). Automatic opinion leader recognition in group discussions. In *Technologies and applications of artificial intelligence (taai), 2016 conference on* (pp. 138–145). IEEE.

Horio, B. M., & Shedd, J. R. (2016). Agent-based exploration of the political influence of community leaders on population opinion dynamics. In *Proceedings of the 2016 winter simulation conference* (pp. 3488–3499). IEEE Press.

Huang, B., Yu, G., & Karimi, H. R. (2014). The finding and dynamic detection of opinion leaders in social network. *Mathematical Problems in Engineering, 2014*.

Ilyas, M. U., & Radha, H. (2011). Identifying influential nodes in online social networks using principal component centrality. In *Communications (icc), 2011 ieee international conference on* (pp. 1–5). IEEE.

Ilyas, M. U., Shafiq, M. Z., Liu, A. X., & Radha, H. (2013). A distributed algorithm for identifying information hubs in social networks. *IEEE Journal on Selected Areas in Communications, 31*(9), 629–640.

Iyengar, R., Van den Bulte, C., Eichert, J., & West, B. (2011). How social network and opinion leaders affect the adoption of new products. *GfK Marketing Intelligence Review, 3*(1), 16–25.

Jaber, M., Wood, P. T., Papapetrou, P., & Helmer, S. (2014). Inferring offline hierarchical ties from online social networks. In *Proceedings of the 23rd international conference on world wide web* (pp. 1261–1266). ACM.

Jiang, L., Ge, B., Xiao, W., & Gao, M. (2013). Bbs opinion leader mining based on an improved pagerank algorithm using mapreduce. In *Chinese automation congress (cac), 2013* (pp. 392–396). IEEE.

Jiang, W., Wang, G., & Wu, J. (2014). Generating trusted graphs for trust evaluation in online social networks. *Future Generation Computer Systems, 31*, 48–58.

Jin, S. V. (2017). "Celebrity 2.0 and beyond!" effects of facebook profile sources on social networking advertising. *Computers in Human Behavior*.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230). ACM.

Jing, L., & Lizhen, X. (2014). Identification of microblog opinion leader based on user feature and interaction network. In *Web information system and application conference (wisa), 2014 11th* (pp. 125–130). IEEE.

Jøsang, A., & Golbeck, J. (2009). Challenges for robust trust and reputation systems. In *Proceedings of the 5th international workshop on security and trust management (smt 2009), saint malo, france* (p. 52).

Kang, C., Kraus, S., Molinaro, C., Spezzano, F., & Subrahmanian, V. (2016). Diffusion centrality: A paradigm to maximize spread in social networks. *Artificial Intelligence, 239*, 70–96.

Kang, C., Molinaro, C., Kraus, S., Shavitt, Y., & Subrahmanian, V. (2012). Diffusion centrality in social networks. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (asonam 2012)* (pp. 558–564). IEEE Computer Society.

Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research, 48*(3), 425–443.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly, 21*(1), 61–78.

Katz, E. (2015). Where are opinion leaders leading us? *International Journal of Communication, 9*, 1023.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). 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.

Kim, Y. S., & Tran, V. L. (2013). Assessing the ripple effects of online opinion leaders with trust and distrust metrics. *Expert Systems with applications, 40*(9), 3500–3511.

Kingdon, J. W. (1970). Opinion leaders in the electorate. *The Public Opinion Quarterly, 34*(2), 256–261.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics, 6*(11), 888–893.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. *In proceedings of the acm-siam symposium on discrete algorithms*. Citeseer.

Koeslag-Kreunen, M. G., Van der Klink, M. R., Van den Bossche, P., & Gijselaers, W. H. (2017). Leadership for team learning: The case of university teacher teams. *Higher Education*, 1–17.

Kostkova, P., Mano, V., Larson, H. J., & Schulz, W. S. (2017). Who is spreading rumours about vaccines?: Influential user impact modelling in social networks. In *Proceedings of the 2017 international conference on digital health* (pp. 48–52). ACM.

Krasikova, D. V., Green, S. G., & LeBreton, J. M. (2013). Destructive leadership: A theoretical review, integration, and future research agenda. *Journal of Management, 39*(5), 1308–1338.

Kumar, D. (2015). Identifying key opinion leaders using social network analysis. *Insights*, 20.

Kumar, S., Spezzano, F., & Subrahmanian, V. (2014). Accurately detecting trolls in slashdot zoo via decluttering. In *Advances in social networks analysis and mining (asonam), 2014 ieee/acm international conference on* (pp. 188–195). IEEE.

Kurmyshev, E., & Juárez, H. A. (2013). What is a leader of opinion formation in bounded confidence models? arXiv:https://arxiv.org/abs/1305.4677.

Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2014). Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering, 1*(2), 76–90.

Lancichinetti, A., Kivelä, M., Saramäki, J., & Fortunato, S. (2010). Characterizing the community structure of complex networks. *PloS One, 5*(8), e11976.

Landherr, A., Friedl, B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering, 2*(6), 371–385.

Laub, P. J., Taimre, T., & Pollett, P. K. (2015). Hawkes processes. arXiv:https://arxiv.org/abs/1507.02822.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice*. Oxford, England: Duell, Sloan & Pearce.

Lazarsfeld, P. F., Merton, R. K., et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and Control in Modern Society, 18*(1), 18–66.

Le Bon, G. (1897). *The crowd: A study of the popular mind*. Fischer.

Lees-Marshment, J. (2012). Political marketing and opinion leadership: Comparative perspectives and findings. In *Comparative political leadership* (pp. 165–185). Springer.

Leibovich, M., Zuckerman, I., Pfeffer, A., & Gal, Y. (2017). Decision-making and opinion formation in simple networks. *Knowledge and Information Systems, 51*(2), 691–718.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 631–636). ACM.

Li, C., Li, Q., Van Mieghem, P., Stanley, H. E., & Wang, H. (2015a). Correlation between centrality metrics and their application to the opinion model. *The European Physical Journal B, 88*(3), 65.

Li, F., & Du, T. C. (2011). Who is talking? an ontology-based opinion leader identifi-

cation framework for word-of-mouth marketing in online social blogs. *Decision Support Systems, 51*(1), 190–197.
Li, H., Huang, S., & Sun, G. (2015b). An opinion leader perceptual model based on pagerank algorithm. In *Behavioral, economic and socio-cultural computing (besc), 2015 international conference on* (pp. 150–155). IEEE.
Li, J., Li, X., & Zhu, B. (2016). User opinion classification in social media: A global consistency maximization approach. *Information & Management, 53*(8), 987–996.
Li, Q., Kailkhura, B., Thiagarajan, J., Zhang, Z., & Varshney, P. (2017). Influential node detection in implicit social networks using multi-task gaussian copula models. In *Nips 2016 time series workshop* (pp. 27–37).
Li, Y., Ma, S., Zhang, Y., Huang, R., et al. (2013). An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems, 43*, 43–51.
Lin, X., & Han, W. (2015). Opinion leaders discovering in social networks based on complex network and dbscan cluster. In *Distributed computing and applications for business engineering and science (dcabes), 2015 14th international symposium on* (pp. 292–295). IEEE.
Liu, J.-G., Ren, Z.-M., & Guo, Q. (2013). Ranking the spreading influence in complex networks. *Physica A: Statistical Mechanics and its Applications, 392*(18), 4154–4159.
Lo, W., Tang, Y., Li, Y., & Yin, J. (2015). Jointly learning sentiment, keyword and opinion leader in social reviews. In *Collaboration and internet computing (cic), 2015 ieee conference on* (pp. 70–79). IEEE.
Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS One, 6*(6), e21202.
Ma, N., & Liu, Y. (2014). Superedgerank algorithm and its application in identifying opinion leader of online public opinion supernetwork. *Expert Systems with Applications, 41*(4), 1357–1368.
MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability: 1* (pp. 281–297). Oakland, CA, USA..
Mathias, R. (1993). The hadamard operator norm of a circulant and applications. *SIAM Journal on Matrix Analysis and Applications, 14*(4), 1152–1167.
May, R. M., & Lloyd, A. L. (2001). Infection dynamics on scale-free networks. *Physical Review E, 64*(6), 066112.
McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.
Mohammadi, S. A., & Andalib, A. (2017). Using the opinion leaders in social networks to improve the cold start challenge in recommender systems. In *Web research (icwr), 2017 3th international conference on* (pp. 62–66). IEEE.
Monge, P. R., & Contractor, N. S. (2001). Emergence of communication networks. *The new handbook of organizational communication: Advances in theory, research, and methods*, 440–502.
Musiał, K., Kazienko, P., & Bródka, P. (2009). User position measures in social networks. In *Proceedings of the 3rd workshop on social network mining and analysis* (p. 6). ACM.
Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 38*(2), 321–330.
Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.
Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th international conference on world wide web* (pp. 798–808). International World Wide Web Conferences Steering Committee.
Van den Nieuwboer, M., Van De Burgwal, L., & Claassen, E. (2016). A quantitative key-opinion-leader analysis of innovation barriers in probiotic research and development: Valorisation and improving the tech transfer cycle. *PharmaNutrition, 4*(1), 9–18.
Orman, G. K., Labatut, V., & Cherifi, H. (2012). Comparative evaluation of community detection algorithms: A topological approach. *Journal of Statistical Mechanics: Theory and Experiment, 2012*(08), P08001.
Ortega, F. J., Troyano, J. A., Cruz, F. L., Vallejo, C. G., & EnríQuez, F. (2012). Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks, 56*(12), 2884–2895.
Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam.. In *Hlt–naacl* (pp. 497–501).
Padilla, A., Hogan, R., & Kaiser, R. B. (2007). The toxic triangle: Destructive leaders, susceptible followers, and conducive environments. *The Leadership Quarterly, 18*(3), 176–194.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web.. *Technical Report*. Stanford InfoLab.
Parau, P., Lemnaru, C., & Potolea, R. (2015). Assessing vertex relevance based on community detection. In *Knowledge discovery, knowledge engineering and knowledge management (ic3k), 2015 7th international joint conference on: 1* (pp. 46–56). IEEE.
Park, C. S., & Kaye, B. K. (2017). The tweet goes on: Interconnection of twitter opinion leadership, network size, and civic engagement. *Computers in Human Behavior, 69*, 174–180.
Pienta, R., Abello, J., Kahng, M., & Chau, D. H. (2015). Scalable graph exploration and visualization: Sensemaking challenges and opportunities. In *Big data and smart computing (bigcomp), 2015 international conference on* (pp. 271–278). IEEE.
Potolea, P. P. C. L. M. D. R. (2016). *Sentiment analysis in social networks* (pp. 243–255)). Morgan Kaufmann.
Pronker, E., Weenen, T., Commandeur, H., Claassen, E., & Osterhaus, A. (2015). Scratching the surface: Exploratory analysis of key opinion leaders on rate limiting factors in novel adjuvanted-vaccine development. *Technological Forecasting and Social Change, 90*, 420–432.
Proskurnikov, A. V., Tempo, R., Cao, M., & Friedkin, N. E. (2017). Opinion evolution in time-varying social influence networks with prejudiced agents. *IFAC-PapersOnLine, 50*(1), 11896–11901.
Richardson, M., Agrawal, R., & Domingos, P. (2003). Trust management for the semantic web. *The Semantic Web-ISWC 2003*, 351–368.
Richmond, V. P. (1980). Monomorphic and polymorphic opinion leadership within a relatively closed communication system. *Human Communication Research, 6*(2), 111–116.
Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2016). Indicators of opinion leadership in customer networks: Self-reports and degree centrality. *Marketing Letters, 27*(3), 449–460.
Rocha, L., Mourão, F., Vieira, R., Neves, A., Carvalho, D., Bandyopadhyay, B., et al. (2016). Connecting opinions to opinion-leaders: A case study on brazilian political protests. In *Data science and advanced analytics (dsaa), 2016 ieee international conference on* (pp. 716–725). IEEE.
Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. *Asna*.
Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.
Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovations: A cross-cultural approach*. Free Press.
Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PloS One, 5*(1), e8694.
Ruhnau, B. (2000). Eigenvector-centrality node-centrality? *Social Networks, 22*(4), 357–365.
Saaty, T. L. (1988). What is the analytic hierarchy process? In *Mathematical models for decision support* (pp. 109–121). Springer.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.
Schneider, J. A., Zhou, A. N., & Laumann, E. O. (2015). A new hiv prevention network approach: Sociometric peer change agent selection. *Social Science & Medicine, 125*, 192–202.
Shafer, G., et al. (1976). *A mathematical theory of evidence*: vol. 1. Princeton University Press Princeton.
Shafiq, M. Z., Ilyas, M. U., Liu, A. X., & Radha, H. (2013). Identifying leaders and followers in online social networks. *IEEE Journal on Selected Areas in Communications, 31*(9), 618–628.
Sheikhahmadi, A., Nematbakhsh, M. A., & Zareie, A. (2017). Identification of influential users by neighbors in online social networks. *Physica A: Statistical Mechanics and its Applications*.
Sherchan, W., Nepal, S., & Paris, C. (2013). A survey of trust in social networks. *ACM Computing Surveys (CSUR), 45*(4), 47.
Shinde, M., & Girase, S. (2016). Identification of topic-specific opinion leader using spear algorithm in online knowledge communities. In *Computing, analytics and security trends (cast), international conference on* (pp. 144–149). IEEE.
Song, K., Wang, D., Feng, S., & Yu, G. (2011). Detecting opinion leader dynamically in chinese news comments. In *International conference on web-age information management* (pp. 197–209). Springer.
Song, S. Y., Cho, E., & Kim, Y.-K. (2017). Personality factors and flow affecting opinion leadership in social media. *Personality and Individual Differences, 114*, 16–23.
Song, X., Chi, Y., Hino, K., & Tseng, B. (2007). Identifying opinion leaders in the blogosphere. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 971–974). ACM.
Szomszor, M., Alani, H., Cantador, I., OHara, K., & Shadbolt, N. (2008). Semantic modelling of user interests based on cross-folksonomy analysis. *The Semantic Web-ISWC 2008*, 632–648.
Tai, A., Ching, W.-K., & Cheung, W.-S. (2005). On computing prestige in a network with negative relations. *International Journal of Applied Mathematical Sciences, 2*, 56–64.
Tang, J., Lou, T., & Kleinberg, J. (2012). Inferring social ties across heterogenous networks. In *Proceedings of the fifth acm international conference on web search and data mining* (pp. 743–752). ACM.
Tang, W., Zhuang, H., & Tang, J. (2011). Learning to infer social ties in large networks. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 381–397). Springer.
Tejavibulya, P., & Eiamkanchanalai, S. (2011). The impacts of opinion leaders towards purchase decision engineering under different types of product involvement. *Systems Engineering Procedia, 2*, 12–22.
Tsai, M.-F., Tzeng, C.-W., Lin, Z.-L., & Chen, A. L. (2014). Discovering leaders from social network by action cascade. *Social Network Analysis and Mining, 4*(1), 165.
Tucci, K., González-Avella, J., & Cosenza, M. (2016). Rise of an alternative majority against opinion leaders. *Physica A: Statistical Mechanics and its Applications, 446*, 75–81.
Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. *Health Education & Behavior, 34*(6), 881–896.
Vander Wal, T. (2007). *Folksonomy*.
Vilares, D., Hermo, M., Alonso, M. A., Gómez-Rodríguez, C., & Vilares, J. (2014). Lys at clef replab 2014: Creating the state of the art in author influence ranking and reputation classification on twitter.. In *Clef (working notes)* (pp. 1468–1478).
Wang, C., Du, Y. J., & Tang, M. W. (2016). Opinion leader mining algorithm in microblog platform based on topic similarity. In *Computer and communications (iccc), 2016 2nd ieee international conference on* (pp. 160–165). IEEE.

Wang, G., Xie, S., Liu, B., & Philip, S. Y. (2011a). Review graph based online store review spammer detection. In *Data mining (icdm), 2011 ieee 11th international conference on* (pp. 1242–1247). IEEE.

Wang, H., & Zhai, C. (2017). Generative models for sentiment analysis and opinion mining. In *A practical guide to sentiment analysis* (pp. 107–134). Springer.

Wang, L., Lou, T., Tang, J., & Hopcroft, J. E. (2011b). Detecting community kernels in large social networks. In *Data mining (icdm), 2011 ieee 11th international conference on* (pp. 784–793). IEEE.

Weenen, T., Pronker, E., Commandeur, H., & Claassen, E. (2013). Barriers to innovation in the medical nutrition industry: A quantitative key opinion leader analysis. *PharmaNutrition, 1*(3), 79–85.

Weisbuch, G., Deffuant, G., Amblard, F., & Nadal, J.-P. (2002). Meet, discuss, and segregate!. *Complexity, 7*(3), 55–63.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270). ACM.

Wu, Y., Cao, N., Archambault, D., Shen, Q., Qu, H., & Cui, W. (2017). Evaluation of graph sampling: A visualization perspective. *IEEE Transactions on Visualization and Computer Graphics, 23*(1), 401–410.

Yan, S.-R., Zheng, X.-L., Wang, Y., Song, W. W., & Zhang, W.-Y. (2015). A graph-based comprehensive reputation model: exploiting the social context of opinions to enhance trust in social commerce. *Information Sciences, 318*, 51–72.

Yang, J., & Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *Data mining (icdm), 2010 ieee 10th international conference on* (pp. 599–608). IEEE.

Yang, L., Qiao, Y., Liu, Z., Ma, J., & Li, X. (2016). Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm. *Soft Computing*, 1–12.

Yoon, K., & Hwang, C.-L. (1981). *Multiple attribute decision making: Methods and applications*. Springer-Verlag BERLIN AN.

Zhang, J., Pei, Y., Fletcher, G. H. L., & Pechenizkiy, M. (2016). Structural measures of clustering quality on graph samples. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 345–348). San Francisco, CA.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., & Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1513–1522). ACM.

Zhao, Y., Kou, G., Peng, Y., & Chen, Y. (2018). Understanding influence power of opinion leaders in e-commerce networks: An opinion dynamics theory perspective. *Information Sciences, 426*, 131–147.

Zhao, Y., Li, S., & Jin, F. (2016a). Identification of influential nodes in social networks with community structure based on label propagation. *Neurocomputing, 210*, 34–44.

Zhao, Y., Zhang, L., Tang, M., & Kou, G. (2016b). Bounded confidence opinion dynamics with opinion leaders and environmental noises. *Computers & Operations Research, 74*, 205–213.

Zhou, H., & Zeng, D. (2009). Finding leaders from opinion networks. In *Intelligence and security informatics, 2009. isi'09. ieee international conference on* (pp. 266–268). IEEE.