CHAPTER

OPINION LEADER DETECTION 10

P. Parau, C. Lemnaru, M. Dinsoreanu, R. Potolea
Technical University of Cluj-Napoca, Cluj-Napoca, Romania

1 INTRODUCTION

Analyzing social networks allows one to gain valuable insights into the structure, dynamics, and flow of information within a group of people. This claim is especially true considering the number of interactions the Internet facilitates as a social communication framework. A task in the field of social network analysis is identifying relevant or out of the ordinary individuals. Among these, opinion leaders are individuals who can influence and shape the opinions of others. Identifying opinion leaders is useful in domains such as health care, for raising awareness on important issues, in advertising and marketing, or in the political domain. Therefore it is important to have methods and measures to objectively and accurately detect opinion leaders. This chapter presents approaches that can be used for detecting opinion leaders and discusses the advantages and drawbacks of different types of methods in different contexts. We also present general methods for identifying leaders or influencers, on the basis of the idea that an influential or important individual in a social network, or more generally a communication network, is likely to be an opinion leader as well.

The chapter is structured as follows: Section 2 contains the problem definition and application examples. Section 3 contains a description of various approaches for identifying opinion leaders, classified in categories based on the type of data analyzed. Section 4 contains a high-level critical discussion of the approaches presented. The chapter ends with concluding remarks in Section 5.

2 PROBLEM DEFINITION

Opinion leaders are individuals who exert a significant amount of influence within their network and who can affect the opinions of connected individuals. Opinion leaders play an important role within the two-step flow of communication model, where information is transferred from the mass media to the public in two steps: first, from the media to opinion leaders and then from opinion leaders to the larger audience [1]. The two-step flow is therefore relevant to the process of influencing and changing people's opinions [2]. According to Katz [1], the following three factors impact the status of opinion leadership: (1) personification of certain values, (2) personal competence, and (3) strategic social location. The

Sentiment Analysis in Social Networks. http://dx.doi.org/10.1016/B978-0-12-804412-4.00010-3

status of opinion leadership might change with time and different individuals may be opinion leaders in different domains.

There are many domains in which information can be disseminated more effectively when opinion leaders are targeted. For instance, in the medical field, detecting opinion leaders has been used to raise awareness and promote new and effective treatments [3], as in HIV prevention or child health promotion [4]. Climate change awareness campaigns have also relied on opinion leaders to effectively disseminate information [5]. Two such examples are campaigns led by Al Gore in the mid-to-late years of the first decade of this century that involved selecting opinion leaders to give presentations and inform the larger public on climate change issues. Opinion leader identification is also of interest in advertising, marketing, and product adoption [3], where companies are interested in attracting potential clients to their products as effectively as possible. Political campaigns can also benefit from identifying and targeting opinion leaders, as evidenced by the 2004 US presidential election campaign of George W. Bush, where opinion leaders were selected to promote the campaign [5]. Knowing who the opinion leaders are can yield benefits in a variety of domains, which makes their correct and efficient identification a very important task.

Traditionally, sociological approaches rely on manual or explicit collection of leader information via questionnaires or interviews [1]. A survey covering 191 articles in the field of sociology found that the main strategies for leader identification are distributed as follows: 19% use sociometric methods, 13% use self-selection, 12% use positional approaches, 11.5% use judges' ratings methods, and the rest do not reveal the identification mechanism [6]. Such approaches possess several drawbacks: On one hand, self-reported information is subjective by nature and self-claimed influence is likely to be a reflection of self-confidence [3]. Having an objective assessment of opinion leadership would lead to more reliable methods of finding leaders. On the other hand, sociological studies are limited in scope by the resources needed to undertake them; having scalable algorithms that can accurately identify opinion leaders allows us to take advantage of the ever-growing quantities of social data generated on the Internet. Thus methods that rely on network science and data mining techniques to identify opinion leaders have been developed, and such approaches are the focus of this chapter.

3 APPROACHES

This section presents the most important approaches for opinion leader identification, grouped, according to the characteristics of the available data, into four categories: methods that rely on static measures extracted from the network topology, either at global level or at community level; methods that additionally employ interaction information, which are suitable in contexts in which such information can be observed or approximated; a third category encompasses approaches that extract—totally or partially—network data from the content generated by users; finally several approaches employ both interaction and content information to determine influential users within the network. For each category we present relevant approaches, highlighting the methods and metrics employed and some evaluation results. We also present several approaches for related challenges, whose solutions can be applied to the leader detection problem.

3.1 MEASURES BASED ON NETWORK STRUCTURE

This section presents measures for assessing leadership on the basis of the structure of the network, considering topological factors as indicators for influential nodes. Heuristics based on vertex degree,

network paths, or the community structure are used to identify potential leaders by their position within the network, considering that social location is an important factor in opinion leadership [1].

3.1.1 Centrality measures

A straightforward method of assessing the relevance of a vertex is to compute its centrality. The centrality of a vertex is a measure of its importance in the network in a given perspective [7] and can be expressed in various ways; thus there are multiple types of centrality measures. Which measure is suitable in a given context depends on which aspect of the network topology the centrality measure captures.

The simplest centrality measure is the degree of the vertex, also called *degree centrality* [7]. This measure suggests that the better connected a vertex is, the more important it is in the network. From the perspective of opinion leaders, it seems intuitive that the potential for communication and influence is greater for individuals with more connections. *Eigenvector centrality* attempts to capture the qualitative aspect of the connections of a vertex. On the basis of the premise that connections to more influential vertices are more important than connections to less influential vertices, the measure also takes the centrality of the neighbors into account [7]. Similarly to eigenvector centrality is *PageRank*, which assesses the importance of webpages in a search engine [8]. PageRank models the probability that a "random" web surfer, who starts at a random webpage and continues following links, will visit a webpage (or vertex in the webpage network). The measure also introduces a damping factor that expresses the probability at each step that the surfer will not continue with a link but will jump to a random webpage. Other centrality measures explicitly consider the indirect connectivity to other vertices in the network. *Closeness centrality* is such a measure, and it represents the average distance, or average shortest path, to all other vertices in the network [7]. The basic idea here is that a central vertex will be closer on average than other, less central vertices. This also makes sense from an influence perspective: a person who can easily reach other people will be more effective in spreading influence. *Betweenness centrality*, on the other hand, attempts to capture the ability of a vertex to control the flow of communication: it indicates how many times a vertex is located on the shortest path between two other vertices [9]. A vertex situated on a large number of such paths has an increased power to control communication since any information passing through those paths will pass through the vertex.

As stated earlier, different centrality measures are suitable for identifying the most important vertices in different contexts, contexts that match the assumptions implied by the specific formulation of the measure. Whether a centrality measure truly reflects importance in a network depends on how information flows in that particular network [10]. Next we look at several findings on centrality measures as indicators of leadership. The findings of a study on opinion leadership in smartphone adoption suggest that degree centrality does indeed indicate opinion leadership [11]. In the same study it was found that, as expected, social influence is greater over stronger connections or ties, but no significant interaction between the connection strength and degree centrality was found. The findings in [12] suggest that degree centrality is an indicator of local opinion leadership, since a high degree centrality means many connections in the direct environment of a vertex. To assess global opinion leadership, the closeness and betweenness centrality measures are better suited. The reason for this, Bodendorf and Kaiser [12] argue, is that individuals who are easily reachable by others have a higher chance of being noticed (closeness centrality), while individuals who lie on communication paths can exert a greater influence on that communication (betweenness centrality). Shafiq et al. [13], in experiments on data collected from Facebook, found that leaders tend to have significantly higher degree centrality on average than other types of individuals, which supports the idea that being well

connected and having a high degree centrality is beneficial to the opinion leader status. However, they also noted that a large number of leaders do not have a high degree centrality, which suggests that degree centrality does not capture all aspects of leadership. The best closeness centrality is found among leaders, especially among introvert leaders. In their experiments, eigenvector centrality could not differentiate between leaders and followers but was significantly lower for neutral nodes (neither leaders nor followers). PageRank, on the other hand, was higher for leaders than for followers. Despite these differences in PageRank and degree centrality between leaders and nonleaders, Shafiq et al. do not recommend these measures for detecting leaders in interaction data.

Lu et al. [14] found the PageRank measure to be less effective in networks of people than in webpage networks, so they propose LeaderRank as a more effective way for ranking people in a network. LeaderRank is similar to PageRank but is parameter free: it introduces a ground node, which is connected to all other nodes in the network, and it has a role similar to that of the damping factor in PageRank. This factor is thus absent, which eliminates the need for calibration. Their experiments on data from Delicious, a social bookmarking service, suggest that LeaderRank outperforms PageRank in the task of identifying influential users. They also found that this algorithm has better performance than just taking into account the number of followers a user has (i.e., degree centrality).

Chen et al. [15] proposed a new semilocal centrality measure that relies on the nearest and the next nearest neighbors to identify the most influential nodes in an undirected network. The proposed approach seems to identify influential nodes better than degree centrality–based and betweenness centrality–based methods. However, other existing random-walk approaches such as PageRank and LeaderRank seem to be better indicators of influential nodes but less computationally efficient. The method was tested on four different real-life networks: bloggers on the MSN Spaces (Windows Live Spaces) website, coauthors in network science a router-level topology, and e-mails of the members of a university. The method performed differently depending on the network structure and is recommended for heterogeneous networks and less for tree-shaped networks. Chen et al. also analyzed the relation between the centrality measures considered, showing that, in general, the proposed local centrality has a strong positive correlation with closeness centrality, and a weak correlation with betweenness centrality and degree centrality.

Cho et al. [16] conducted a systematic study using a stochastic cellular automata model associated with the small-world assumption to analyze the effect of different centrality metrics and network properties on the diffusion speed and maximum cumulative number of adopters in the context of the diffusion of innovation. The experimental results reveal that distance centrality and rank-nomination centrality are the best for maximizing the cumulative number of adopters, whereas sociality provides the best speed of diffusion.

3.1.2 Community-driven measures

There are a variety of measures for assessing the importance of an individual in a network, but many of them—for instance, the classic centrality measures mentioned in the previous section—do not take the community structure of the network into account. A common definition for a community is an area of the graph that has a higher connection density [17], and is usually formed by vertices that share common properties [18]: individuals who are friends, who frequently communicate with each other, or groups of like-minded individuals. From the perspective of social location as a factor for opinion leadership, both the connections to people inside the group and also external connections are

important [1]. The community structure is an important characteristic of networks, and it seems natural to take the position of individuals within it into account when one is assessing their importance.

One of the simplest measures that can be used to discover relevant vertices in the community structure of a network is *embeddedness* [19–21]. It is defined as the ratio between the internal degree (the number of connections to other vertices within the community) and the total degree (all connections, including ones with vertices outside the community). This measure quantifies how strongly a vertex belongs to its community, however, in real-world networks many vertices have no connections to other communities and they have high embeddedness values [19,20,22]. In such cases the ability of embeddedness to distinguish between important and unimportant vertices is diminished. Parau et al. [22] proposed the *relative commitment* measure, which is simple to compute yet captures more information than embeddedness. On one hand, relative commitment takes the magnitude of the internal degree of a vertex into account and, on the other hand, it represents a weighted measure: connections to highly committed vertices will have a greater impact on the score than connections to less committed vertices. This can be viewed as a force acting on the vertex, pulling it towards the community of the connected vertex with a force that depends on the commitment of the latter. Rosvall and Bergstrom [23] presented a method for quantifying the *statistical significance* of vertices in a community structure and identifying the significance clusters of vertices within communities. Their method consists in generating a number of bootstrap networks in which the connections are modified to a certain extent compared with the original network. The community structure of all of these networks is compared and the largest subset of vertices in a community that appear together in the same community in at least 95% of the bootstrap networks is the significance cluster of that community. The significance of a vertex, defined as the percentage of bootstrap networks in which the vertex maintained its original community membership, can thus be viewed as a measure of the membership strength of that vertex, and the findings in [22] suggest that it is a superior measure compared with embeddedness and relative commitment. A disadvantage of significance is that it is computationally intensive. Guimera and Amaral [24] proposed two measures that are jointly used to classify vertices in a community structure in seven roles. The *z score* measures how well connected a vertex is within its own community, and the *participation coefficient* quantifies the degree to which the connections of a vertex are distributed among the communities of the network. On the basis of the z score, vertices are classified as hubs and nonhubs, which are then further classified on the basis of the participation coefficient.

The previously mentioned community-driven measures are classified in [22] as measures of commitment to a community. Parau et al. argue that there is also another dimension of the community-driven relevance of a vertex—community importance—and propose a categorization of vertices based on these two dimensions. Commitment is an indicator of the membership strength of a vertex toward its community, while community importance shows how prominent a role the vertex plays for the structure of the community. Parau et al. proposed a community importance measure called *community disruption*, which assesses the importance of a vertex by examining the effects on the community of removing the vertex from the network. The measure is based on the idea that the removal of an important vertex will cause greater changes in the community (e.g., the community splitting into subcommunities, vertices migrating to other communities) than unimportant vertices. A disadvantage of this measure is that it is computationally intensive.

A category among the community-driven approaches to detect relevant vertices is represented by methods that aim to identify community kernels: the core members of a community. In [25] the kernel

of a community consists of influential vertices inside that community: each community is composed of such a kernel and an auxiliary community of nonkernel members. Wang et al. [25] argue that kernel members can be detected with use of a centrality measure such as degree centrality or PageRank followed by a community detection method on these kernels only, but that doing so ignores the connections to auxiliary members. Thus they propose two algorithms to find the kernels and conduct experiments on coauthorship networks and on a Twitter dataset. An extension of these algorithms is presented in [26], where the attributes of vertices (eg, gender, age) are also considered while the community kernels are found. Du et al. [27] found community kernels by using overlapping maximal cliques. They also performed a step of community naming in which they identify the central entities of each community through a central entity resolution algorithm and build a community profile based on them.

3.1.3 Summary

The findings presented in this section suggest that while measures based solely on the topological properties of a network capture important aspects of leadership, there are other factors that can influence the status of a vertex that are not captured by these measures. In the following sections, we examine approaches that consider additional factors to identify opinion leaders.

3.2 METHODS BASED ON INTERACTION

Network topology information captures the relationships between individuals at a certain moment in time. By contrast, interaction information captures dynamic exchanges between individuals, usually also considering the temporal dimension of these exchanges. Analyzing user interactions can help establish the flow of influence in a network, which can provide significant insights into the leading individuals within the network. The flow of influence in a network is closely related to information diffusion processes, in the sense that the global (or local) influence of a node can be generally inferred if the diffusion process is known, or modeled correctly. All approaches in this section rely on interaction information and most consider structural network information also; in terms of the algorithms employed, the approaches range from greedy search strategies to frequent pattern mining, or clustering.

A general framework for modeling the diffusion process in a dynamic network is presented in [28]. The diffusion is modeled as a discrete network of continuous, conditionally independent, temporal processes that may occur at different rates. Rodriguez et al. [28] start by considering three well-known functions to model the conditional transmission likelihood between two nodes (exponential, power-law and Rayleigh). On the basis of these, survival and hazard functions are defined, which leads to the definition of the likelihood of a cascade. The proposed solution managed to track near perfect performance in the evaluations performed, for both continuous and discontinuous transmission rates. Although it does not explicitly identify leaders, the approach is relevant because it models the information flow in the network, and can be used to infer the influential nodes.

In the domain of advertising in marketing, analyzing the purchasing habits of users can be used not only to predict purchases but also to identify how user influence acts on the decision to buy a product. Richardson and Domingos [29,30] proposed probabilistic models for networks extracted from knowledge-sharing sites to estimate customer network value, so as to devise the best viral marketing

plan. The motivation for their approach is the fact that many markets possess strong network effects (i.e., an individual's decision to purchase is influenced both by an intrinsic motivation and by the influence exerted by people whose opinion he/she values). Therefore each potential customer possesses an intrinsic value and a network value, which captures the strength of the influence the particular customer has on the probability of a purchase by other customers.

Kempe et al. [31] proposed a greedy hill climbing approach for the problem of detecting the k most influential individuals, and then proved that it provides a solution within 63% of the optimal solution, for several diffusion models. They explicitly consider two classes of diffusion models—the independent cascade model and the linear thresholds model—and use the theory of submodular functions to prove the quality of the greedy approach. Comparisons performed with centrality-based heuristics show that the greedy approach achieves a greater number of adopters than the other heuristics, which confirms that approaches that consider network dynamics perform better than ones that rely on the structural properties of the graph alone.

Another approach that considers user interactions is presented in [32]. It relies on frequent pattern mining and integrates an undirected social network with a log of actions performed by the nodes to build a directed acyclic propagation graph. Leadership is defined in terms of the number of nodes that perform the same action as an initial user, after the initial user but within a time frame, and are reachable from that initial user via social links. On the basis of the propagation graph, user influence subgraphs are generated, from which leaders and tribe leaders are identified. Additional measures are considered, such as confidence and genuineness. The approach is also related to the community detection problem if a correlation between tribes and communities is considered. The evaluation data were obtained by the integration two real data subsets: one from Yahoo! Movies ratings and one from Yahoo! Messenger.

The *longitudinal user centered influence* model [13] relies on directed user interactions and computes two coefficients by means of linear regression: the ego coefficient and the network coefficient, correlating the user past interaction and the past inward interaction, respectively, with the future outward one. The approach relies on a generalization of the Friedkin-Johnsen influence model. A kernel k-means algorithm is employed to cluster individuals into introvert leaders, extrovert leaders, followers, and neutrals. Shafiq et al. [13] also examined the characteristics of the identified classes on a dataset extracted from Facebook, considering properties such as centrality measures, the number of triangles, and the clustering coefficient. The fact that different classes of users have different average values and distributions for these properties suggests that the selected classes are relevant, but also that some of these measures can be used to discriminate between members of different classes.

Diffusion processes can also be represented within the community structure of the network. Amor et al. [33] presented a method for community detection and role identification that relies on the analysis of a Markov process in the graph and defines the quality of the partition in terms of trapping the flow of the diffusion process. The users have different roles in the propagation of information, defined by similar in and out flow patterns. The solution employs the role-based similarity (cosine) measure to generate a role-based similarity graph by using the relaxed minimum spanning tree algorithm. The Markov stability function was used on the similarity graph to identify communities of users having similar in and out flow patterns (roles). Although this method does not explicitly identify leaders, nodes exhibiting leader behavior can be easily identified. The solution was evaluated on topic-specific Twitter data (the care.data debate). The community detection solution was applied on two networks—the follower and the retweet networks—yielding the interest communities and the conversation communities respectively. Roles were identified in the interest communities.

The approaches reviewed in this section consider the flow of information so as to identify the most influential nodes. The solutions generally combine centrality-based metrics with various models of information propagation such as random walk, Markov processes, and probabilistic models. They include methods for analyzing the models obtained via classification, clustering, or simple threshold-based functions to determine the most influential nodes. The most prominent application field for this category of methods is marketing.

3.3 METHODS BASED ON CONTENT MINING

This category of strategies relies on network structure and information extracted from the content various nodes in the network are exposing and sharing. The content is used to extract topic information (since opinion leadership is topic dependent), explicit user features (eg, expertise, geographical location, post time), or sentiment. Although not all strategies are specifically designed for leader identification, the approaches are applicable to such a context as well.

Song et al. [34] proposed the *InfluenceRank* algorithm to identify opinion leaders among blogs. The algorithm considers not only how central a blog is compared with others but also how novel the information posted on that blog is, and Song et al. argue that both properties are important. They claim that information posted on a blog may originate from other linked blogs or from a source not present in the blog network, or that it may be original. The latter sources of information can be viewed as a hidden node connected to that blog from which information novelty originates. They used latent Dirichlet allocation as a topic model and the cosine similarity measure to assess the novelty of a blog's entries and subsequently the information novelty of the blog. The InfluenceRank measure contains a parameter that controls the extent to which novelty affects the score, and if that parameter is 0, the measure reduces to PageRank.

An original strategy (*BARR—blog content, author, reader properties, relationship*) for opinion leader identification in social blogs was presented in [35]. The method relies on the observation that the influence of a blogger can be estimated as the weighted sum of the quality and the quantity of blogs created. The method combines several data sources (keywords, blog portals, and web sources) and applies an ontology-based extraction method to collect values for 11 parameters, such as the number of visits per blog, the author's expertise, author blog preference, reader expertise, homophily, and tie strength. The method uses the technique for order preference by similarity to the ideal solution (TOPSIS) to compute the related similarity of each blog to the ideal solution, and compute the hot blog of an author. Ultimately, it expresses the influence of an author as the weighted sum of the quality (the normalized, largest related similarity) and the quantity of his/her blogs.

Huang et al. [36] presented the *positive opinion leader detection* method, which starts from a graph-based three-level structure representing the topic, comments, and users. The approach performs sentiment analysis on the comments, classifying them as positive, negative, or neutral. The weights of the edges are computed on the basis of the sentiment similarities of the comments. The impact of time (i.e., the older a comment, the weaker its influence) is also considered as a function used to reduce the probability of the selection of a comment. On the basis of the time-weighted, normalized weights, an improved finite Markov chain is defined.

Opinion leadership can be reflected in the way the content individuals produce is distributed and shared across social media or even more traditional forms of media. For instance, Niculae et al. [37] presented an unsupervised prediction setting to quantify media bias, by means of identifying the quotes

an outlet will broadcast. *QUOTUS* was developed on the basis of data represented by 6 years of speeches by Barack Obama and the way they were reflected in the media. The solution requires the identification of patterns in the outlet-to-quote bipartite graph. While the authors do not explicitly refer to detection of opinion leaders, the principles presented may be applied to identify leaders of a community: if, instead of speeches we consider content produced by individuals and instead of quotes by the media we consider sharing and distributing the said content or parts thereof in different forms of media, we obtain a picture of what content is shared the most and which individuals produce the most shared content.

As mentioned previously, there is value in taking into account the community structure of a network when one is attempting to identify opinion leaders. Considering the observations in [1]—that different individuals can be opinion leaders in different domains—communities in a network can be viewed as an environment in which local opinion leaders act (depending on how they are constructed). Taking into account the characteristics of people when one is building the communities may help one in finding groups of similar people and subsequently finding leaders within them. In this sense, McAuley and Leskovec [38] presented a strategy to identify circles (i.e., communities) in ego networks on the basis of both profile similarity and network structure, which also allows for overlapping circles (relying on different, but not necessary disjoint dimensions of the profile). The evaluations were performed on data harvested from Facebook, Google+, and Twitter. Parau et al. [39] built networks based on the opinions of individuals, both based on the opinion target and aggregated across multiple opinion targets. These networks and their community structure could also be useful for detecting opinion leaders.

3.4 METHODS BASED ON CONTENT AND INTERACTION

The approaches falling into this category rely on both interaction information and content in addition to the network topology to identify opinion leaders. Temporal cues such as novelty, feedback flow, and explicit temporal data are exploited by the methods, together with specific content extracted, such as topic identification and user-related features (expertise, opinions, etc.). Some approaches in this category consider network topology information, while others rely solely on content and temporal information.

Li et al. [40] presented *ENIA*—a framework for ranking opinion leaders in online learning communities on the basis of four indicators: expertise, novelty, influence, and activity. They are extracted from three sources of information—textual content, observed user behavior, and temporal information—to compute a score for each of the four indicators. These scores are then multiplied to obtain the user score. The solution quality is evaluated with use of metrics to estimate longevity and centrality. ENIA was compared with PageRank and HITS on two datasets for online learning communities, showing superior correlation between the estimated ratings and the human-produced ratings compared with the other two approaches.

An opinion leader can also be viewed as an active member of a community whose produced content and feedback is important to other individuals. The findings in [41] suggest that received feedback impacts future user behavior asymmetrically on the basis of the feedback polarity. The approach employs binomial regression, considering only textual features from post content to predict the quality of the post. The outcomes of the evaluations are relevant in the context of the opinion leader by estimating the quality of the posts and measuring the feedback they receive, which would allow the detection of users who are likeliest to polarize interest.

The experiments reported in [42] provide understanding of how information spreads online and are relevant for predicting cascade growth. Moreover, the study identifies that the breadth of the initial

growth in a cascade is an indicator of larger cascades and the depth is not. The features engineered for learning are grouped into four types: content-related features, original poster/resharer features, structural features, and temporal features. The experiments were conducted on Facebook photos uploaded within 1 month and reshares within 28 days of the initial upload. The findings of the study show that temporal and structural features are the most predictive, while the set of temporal features in isolation outperforms all other individual feature sets. Relevant for opinion leader detection are the observations that the star configuration tends to grow into the largest yet slowest cascades, that a median resharer has 35 fewer friends than someone who is active on the site nearly every day, and that cascades with initial fast reshares are likelier to grow significantly. Moreover, the finding that temporal and structural features are key predictors of cascade size allows efficient identification of relevant features for the community growth.

A good strategy to identify an opinion leader would be by predicting the number of reshares of a given post. On the basis of the theory of a self-exciting point, Zhao et al. [43] presented a theoretical framework to explain the temporal patterns of information cascades. The main advantage of the strategy is that it requires only the time history of reshares together, the degrees of the resharing nodes, and minimal knowledge about the information cascade and the underlying network structure. The method needs no feature engineering, hence it is not context dependent, and scales linearly (on the number of reshares of a given post). The mechanism was compared against four methods, and it is better and faster than each of them. *SEISMIC* can make an initial prediction of leading tweets in the first 5 minutes after the original post (with the time to learn parameter and time for prediction of just 0.02 seconds), with 25% error rate after 10 minutes and 15% error rate after 1 hour of tweet surveillance; moreover, just knowing the delay (of a retweet) allows one to accurately model the speed of a cascade spreading. The leader identification can be seen as identifying breakout tweets, applicable in contexts such as trend forecasting and rumor detection.

4 DISCUSSION

The first challenge for identifying opinion leaders is to determine how a leader is defined in different contexts. If we consider opinion leaders as individuals who have the ability to significantly influence other individuals, there are multiple perspectives from which leadership can be viewed. Firstly, a leader must have the ability to transmit influence, and this can be intuitively seen as being reflected in the structural properties of the social network. The leader concept is also commonly defined as the starting point of an action that propagates farther and faster than other actions. Therefore the leader detection problem is closely related to the dynamic propagation problem in networks, process modeling and process mining in networks, and influence modeling. A holistic solution can be achieved by integration of the structural properties, influence flow, and diffusion processes in networks. An important factor for identifying opinion leaders is who they are, their intrinsic properties, which, together with other factors, leads to their status. One can indirectly study these intrinsic properties by looking at personal characteristics, and also by examining the content produced by leaders, whether we talk about posts on social media, blogposts, comments on news articles, or even endorsements of other's content through sharing. In reality, all these elements factor into the opinion leader status, and ideally methods for detecting them would consider all these aspects. However, designing generally applicable methods that consider all these factors is a difficult task. As such, which method is best applicable to a specific problem depends naturally on the context. In some contexts a certain factor may outweigh the others

significantly, such that a method specifically tailored to measure it (eg, degree centrality as a structural measure) can yield good results. In others the leader status is determined significantly by multiple factors, so the best results would be gained by the use of multiple methods, or hybrid methods.

Precisely defining what a leader is in a certain context is the first step in choosing the most appropriate methods or metrics to detect leadership. A further challenge is to identify the leader features in the available data. Even if the characteristics to be measured have been defined, they may not be clearly expressed in the available data, or they may not even be present. So the decision as to what approach to use in identifying opinion leaders is dependent on two factors: defining the concept of opinion leader in a particular context and identifying the defined features in the available data.

In the remaining paragraphs of this section we critically analyze the approach categories presented in Section 3. The advantage of using centrality measures to identify important individuals is that they rely only on the topology of a graph to assess the importance of its vertices. All such approaches are based on examination of the connectivity between vertices, which makes them relatively easy to use as long as the available data can be represented as a graph. This does not mean that they are universally suitable as leadership metrics, but it does represent an advantage over more specific measures: centrality measures generally capture fundamental network properties, which lends them intrinsic value. Whether they are used as is to assess importance or as input for a more complex measure, they can be useful in a variety of contexts.

The community structure of a network is an important property, especially in social contexts, where it can reflect the social structure present in such networks. The community topology shapes the flow of communication in a network and implicitly the flow of influence. Communities may form around important individuals who play an important role in their structure. Individuals may be leaders just in their own community, or they may influence multiple communities, and they can have varying degrees of commitment to different groups. Community-driven measures rely on the topology of the community structure, and as such the advantages and drawbacks mentioned in the previous paragraph apply here as well. On the basis of the idea that leaders are the most important or central individuals in a network, algorithms that search for the group of individuals who form the kernel of a community can also be useful to identify leaders or to select a set of candidates and thus limit the search space of the problem.

The generality of centrality or community-driven measures is simultaneously an advantage and a drawback: the information captured by them may be limited. Their usefulness is tightly coupled with how the network is built and what relationships between individuals it models. In cases where leadership is a combination of more factors than can be easily and effectively represented as network properties, more advanced centrality measures or other methods or algorithms may be needed to fully capture the concept of leadership.

Structural approaches consider only static properties of the network. To take advantage of the dynamic processes occurring in a network, one must consider the flow of influence, the communication processes, and the information diffusion model. Influential individuals can be viewed as hubs, playing important roles in the flow of influence in a network. Interactions in the network can offer important information on how influence flows between different categories of individuals: influence may flow from opinion leaders toward followers, followers may further contribute to the diffusion of influence by sharing or passing on information from the leaders they follow, etc.

To assess not only how individuals are communicating but also what they are communicating, one must examine the content they produce. Analyzing the topics, determining the sentiment, and assessing the novelty and originality of what someone posts are all factors that allow us to understand the characteristics of individuals. In social media especially, the type and quality of the content an

individual produces is as important as his or her ability to distribute it effectively. Analyzing the content is perhaps a more difficult task since it addresses the question of why a certain type of content or certain opinions appeal to a larger audience and other types do not. We believe that while analyzing content can add great value to the task of identifying opinion leaders, content analysis techniques must be accompanied by methods that also assess an individual's ability to communicate: an individual can post the most interesting and original ideas, but he or she will never be an opinion leader if no one reads them.

Trends and directions in this domain can be observed from three main perspectives: approaches, evaluation, and applicability. The first important thing to note is that there is no universal approach. Even if solutions based on network structure and those that consider diffusion models possess a higher degree of generality, many recent methods have shifted toward trying to estimate the intrinsic features of individuals, by considering the content produced by them in conjunction with their observable relations (modeled as a social network) and (possibly) behavior. A potentially promising development would be to correlate the models with external events that might influence the behavior of the individuals, or to use several sources of data to extract information on the same individual (eg, intersocial networks). Analyzing multimedia content in addition to text could also provide a better assessment of an individual's characteristics.

Given the heterogeneity of existing solutions, it is natural that there are no established metrics and evaluation scenarios. This makes it difficult to compare methods. However, several measures that might be used to compare the methods have emerged recently, and it is important to note the evolution from human evaluation toward quantitative metrics.

While initially triggered by a pure economic necessity, with important marketing applications even nowadays, opinion leader detection has gained interest in other fields, such as online learning, disease awareness campaigns, and politics. There is an increasing interest in analyzing/predicting viral content, thus producing a focus shift from the individual toward the information an individual produces. This is strongly related to the emergence of solutions that exploit content.

5 CONCLUSIONS

The aim of this chapter was to present an overview of opinion leader detection approaches. To that end we described the problem of identifying opinion leaders and then reviewed various solutions categorized on the basis of the type of features they measure. We examined structural methods, including centrality and community-driven measures, approaches that focus on interaction data and influence diffusion processes within a network, and approaches that also consider the content of information posted by individuals.

The approaches, models, and processing flows, however, are highly dependent on the problem domain and the specific data available; therefore general-purpose solutions are extremely difficult to devise. However, practically speaking, the objective is to identify leaders on specific topics and in specific contexts (available data, domain); therefore the best specifically tuned methods have to be employed. The approaches presented in this chapter reflect these observations and allow the reader, we believe, to appropriately choose a method or an approach suited to a particular context.

REFERENCES

[1] E. Katz, The two-step flow of communication: an up-to-date report on an hypothesis, Public Opin. Quart. 21 (1) (1957) 61–78.

[2] V.C. Troldahl, A field test of a modified "two-step flow of communication" model, Public Opin. Quart. 30 (4) (1966) 609–623.

[3] R. Iyengar, C. Van den Bulte, J. Eichert, B. West, T.W. Valente, How social networks and opinion leaders affect the adoption of new products, GfK Market. Intell. Rev. 3 (1) (2011) 16–25.

[4] K. Guldbrandsson, M.K. Nordvik, S. Bremberg, Identification of potential opinion leaders in child health promotion in Sweden using network analysis, BMC Res. Notes 5 (1) (2012) 1–4.

[5] M.C. Nisbet, J.E. Kotcher, A two-step flow of influence? Opinion-leader campaigns on climate change, Sci. Commun. 30 (3) (2009) 328–354.

[6] T.W. Valente, P. Pumpuang, Identifying opinion leaders to promote behavior change, Health Educ. Behav. 34 (6) (2007) 881–896.

[7] M.E.J. Newman, The mathematics of networks, in: L.E. Blume, S.N. Durlauf (Eds.), The New Palgrave Encyclopedia of Economics, Palgrave Macmillan, Basingstoke, second ed., 2008, pp. 1–12.

[8] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.

[9] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1) (1977) 35–41.

[10] S.P. Borgatti, Centrality and network flow, Soc. Netw. 27 (1) (2005) 55–71.

[11] H. Risselada, P.C. Verhoef, T.H.A. Bijmolt, Indicators of opinion leadership in customer networks: self-reports and degree centrality, Market. Lett. (2015) 1–12, http://link.springer.com/article/10.1007/s11002-015-9369-7.

[12] F. Bodendorf, C. Kaiser, Detecting opinion leaders and trends in online communities, in: ICDS, 2010, pp. 124–129.

[13] M.Z. Shafiq, M.U. Ilyas, A.X. Liu, H. Radha, Identifying leaders and followers in online social networks, IEEE J. Sel. Areas Commun. 31 (9) (2013) 618–628.

[14] L. Lu, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS ONE 6 (6) (2011) e21202.

[15] D. Chen, L. Lu, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (4) (2012) 1777–1787.

[16] Y. Cho, J. Hwang, D. Lee, Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach, Technol. Forecast. Soc. Change 79 (1) (2012) 97–106.

[17] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.

[18] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3-5) (2010) 75–174.

[19] A. Lancichinetti, M. Kivela, J. Saramaki, S. Fortunato, Characterizing the community structure of complex networks, PLoS ONE 5 (8) (2010) e11976.

[20] G.K. Orman, V. Labatut, H. Cherifi, Comparative evaluation of community detection algorithms: a topological approach, J. Stat. Mech. vol. 2012 (08) (2012) P08001.

[21] G. Palla, A.L. Barabasi, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664–667.

[22] P. Parau, C. Lemnaru, R. Potolea, Assessing vertex relevance based on community detection, in: IC3K, 2015, pp. 46–56.

[23] M. Rosvall, C.T. Bergstrom, Mapping change in large networks, PLoS ONE 5 (1) (2010) e8694.

[24] R. Guimera, L.A.N. Amaral, Cartography of complex networks: modules and universal roles, J. Stat. Mech. 2005 (2) (2005) P02001.

[25] L. Wang, T. Lou, J. Tang, J.E. Hopcroft, Detecting community kernels in large social networks, in: ICDM, 2011, pp. 784–793.

[26] D. Maccagnola, E. Fersini, R. Djennadi, E. Messina, Overlapping kernel-based community detection with node attributes, in: IC3K, 2015, pp. 517–524.
[27] N. Du, B. Wu, X. Pei, B. Wang, L. Xu, Community detection in large-scale social networks, in: WebKDD/SNA-KDD, 2007, pp. 16–25.
[28] M.G. Rodriguez, J. Leskovec, D. Balduzzi, B. Scholkopf, Uncovering the structure and temporal dynamics of information propagation, Netw. Sci. 2 (1) (2014) 26–65.
[29] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: KDD, 2002, pp. 61–70.
[30] P. Domingos, M. Richardson, Mining the network value of customers, in: KDD, 2001, pp. 57–66.
[31] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: KDD, 2003, pp. 137–146.
[32] A. Goyal, F. Bonchi, L.V. Lakshmanan, Discovering leaders from community actions, in: CIKM, 2008, pp. 499–508.
[33] B. Amor, S. Vuik, R. Callahan, A. Darzi, S.N. Yaliraki, M. Barahona, Community detection and role identification in directed networks: understanding the Twitter network of the care.data debate (2015), arXiv:1508.03165.
[34] X. Song, Y. Chi, K. Hino, B. Tseng, Identifying opinion leaders in the blogosphere, in: CIKM, 2007, pp. 971–974.
[35] F. Li, T.C. Du, Who is talking? An ontology-based opinion leader identification framework for word-of–mouth marketing in online social blogs, Decis. Supp. Syst. 51 (1) (2011) 190–197.
[36] B. Huang, G. Yu, H.R. Karimi, The finding and dynamic detection of opinion leaders in social network, Math. Prob. Eng. vol. 2014 (2014), Article ID 32840.
[37] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, J. Leskovec, QUOTUS: the structure of political media coverage as revealed by quoting patterns, in: WWW, 2015, pp. 798–808.
[38] J. McAuley, J. Leskovec, Discovering social circles in ego networks, ACM Trans. Knowl. Discov. Data 8 (1) (2014) 4:1–4:28.
[39] P. Parau, A. Stef, C. Lemnaru, M. Dinsoreanu, R. Potolea, Using community detection for sentiment analysis, in: ICCP, 2013, pp. 51–54.
[40] Y. Li, S. Ma, Y. Zhang, R. Huang, Kinshuk, An improved mix framework for opinion leader identification in online learning communities, Knowl. Based Syst. 43 (2013) 43–51.
[41] J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, How community feedback shapes user behavior, in: ICWSM, 2014.
[42] J. Cheng, L. Adamic, P.A. Dow, J.M. Kleinberg, J. Leskovec, Can cascades be predicted? in: WWW, 2014, pp. 925–936.
[43] Q. Zhao, M.A. Erdogdu, H.Y. He, A. Rajaraman, J. Leskovec, SEISMIC: a self-exciting point process model for predicting tweet popularity, in: KDD, 2015, pp. 1513–1522.