Modeling Influence with Semantics in Social Networks: A Survey

GERASIMOS RAZIS and IOANNIS ANAGNOSTOPOULOS, Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
SHERALI ZEADALLY, College of Communication and Information, University of Kentucky, Lexington, Kentucky, USA

The discovery of influential entities in all kinds of networks (e.g., social, digital, or computer) has always been an important field of study. In recent years, Online Social Networks (OSNs) have been established as a basic means of communication and often influencers and opinion makers promote politics, events, brands, or products through viral content. In this work, we present a systematic review across (i) online social influence metrics, properties, and applications and (ii) the role of semantic in modeling OSNs information. We found that both areas can jointly provide useful insights towards the qualitative assessment of viral user-generated content, as well as for modeling the dynamic properties of influential content and its flow dynamics.

CCS Concepts: • General and reference → Surveys and overviews; • Networks → Online social networks; • Human-centered computing → Social network analysis; • Software and its engineering → Semantics;

Additional Key Words and Phrases: Information quality, online social influence, social networks, social semantics

ACM Reference format:
Gerasimos Razis, Ioannis Anagnostopoulos, and Sherali Zeadally. 2020. Modeling Influence with Semantics in Social Networks: A Survey. ACM Comput. Surv. 53, 1, Article 7 (February 2020), 38 pages.
https://doi.org/10.1145/3369780

1 INTRODUCTION

Nowadays, hundreds of millions of messages are shared on a daily basis among the users of Online Social Networks (OSNs). There is a diversity of users ranging from civilians to politicians and from news channels to big corporations. In this “ocean” of information, a challenging task is the discovery of the important actors who are able to influence others and produce messages of high social quality, importance, and recognition. Those influential users are also called opinion leaders [15], domain experts [16], influencers [111], innovators [81], prestigious [61], or authoritative actors [39]. Often, their degree of influence is also complemented or affected by various quality measurements
that are based on their social semantics. The latter can either be related to the content of the messages (e.g., keywords, hashtags) or to the metadata of the user (e.g., activity, relationship details).

In this article, we study two major aspects of OSNs, namely, the online social influence (Section 3) and the role of social semantics (Section 4) in OSNs, towards the qualitative assessment of viral user-generated content (Section 5). Specifically, we examine how influence can be measured or predicted and what kinds of methodologies are used to measure influence (e.g., based on topology, diffusion, or social authority), and what are the application domains. Regarding the role of semantics in OSNs, we analyze related works based on Semantic Web technologies along with network theory and graph properties for topic identification, detection of similar users and communities, as well as user personalization (e.g., interests, suggestions).

To perform a more detailed analysis and to adequately cover all perspectives of the aforementioned two aspects, we analyzed the reviewed related research works according to the hierarchical classification scheme depicted in Figure 1. In most of the cases, a referred work does not fall within the scope of only one topic, thus demonstrating that related research efforts in these fields are complementary.

More specifically, with respect to the online social influence, we classify the related works according to the following four topics:

- **Topic 1 - Influence Metrics**: This topic includes works proposing methodologies that define online social influence and how to measure it. Thus, this topic is further divided in three subtopics, namely, (a) Direct social information-based metrics, (b) Hyperlink-based metrics, and (c) Metrics based on machine learning techniques.

- **Topic 2 - Information Flow and Influence**: This topic examines the impact of influence/influential users with respect to viral properties of information as well as information propagation and information diffusion. Although there is no clear distinction between “propagation” and “diffusion” in the literature covering the OSNs and often these terms are used interchangeably, in this survey, we explicitly examine separately the impact of influence in information propagation and information diffusion. Diffusion is about the spread of information from a starting node toward the rest of the network, while propagation takes into consideration the intermediate nodes as well, which receive, process, and further decide whether to re-transmit, re-direct, or block the information. Thus, in this work, we divide information flow and influence topic into two subtopics, namely, propagation-oriented and diffusion-oriented.

- **Topic 3 - Network/Graph Properties**: This category contains works that utilize the topology of a network or its structure to measure influence. Usually only a fraction of the whole network is used due to hardware or complexity limitations.

- **Topic 4 - Applications**: This topic presents the usage of the above metrics, mainly in applications that provide solutions for opinion makers, data analysts, and information scientists. This topic is further divided into two subtopics, namely, (a) Ranking and (b) Recommendation.

As for the role of social semantics in the provision of a qualitative assessment of viral user-generated content, we classify the related works we have reviewed into three topics:

- **Topic 1 - Social Modeling**: This topic contains approaches that adopt semantics for modeling the logical topology and structure of online social networks and media as well as the disseminated information.

- **Topic 2 - Social Matching**: The studies presented on this topic exploit the use of social semantics for identifying similar properties and activities with respect to user-generated content,
Fig. 1. The hierarchical classification scheme followed in this work.
description of real-life events, as well as revealing user interests and behavioral patterns across different online social media users. Thus, we divide this topic into two subtopics, namely, (a) User-oriented (e.g., similar user recommendation, user preferences, and so on) and (b) Topic and Event-oriented (e.g., topic profiling and user interest, event detection, product marketing).

- **Topic 3 - Community Detection**: This category covers works that use social semantics for the detection of communities in OSNs.

This survey aims to help both researchers and data scientists to better understand how viral content is propagated, the role and effect of influential nodes in its diffusion, how can we measure influence of social network users, and the reasons why the proper use of semantics for users and their generated contents can provide useful insights and qualitative conclusions for numerous domains such as marketing, information retrieval, recommendation systems, community and/or event detection, query expansion, thematic categorization, homophily tendency, and sentiment analysis. We summarize the goals of this survey as follows: (a) factors affecting the influence of OSN users, (b) factors affecting the spread of social information, and (c) how semantics can help towards the qualitative assessment of viral user-generated content.

To conduct our literature review, we collected 125 studies strongly related to the aforementioned issues. Initially, we used a specific set of related keywords as input for the discovery of relevant publications by submitting them through the academic digital library and search engine Application Programming Interfaces (APIs). Specifically, we utilized open access repositories (e.g., Google Scholar, arXiv, SSRN) and digital libraries that request subscription (e.g., ACM Digital Library, IEEE Xplore, Elsevier, others). A few indicative keywords that we used to search for appropriate publications are the following: “Influence maximization,” “Influence propagation,” “Information propagation,” “Social network semantics,” “User interest,” “Social semantic modeling,” “Context-dependent influence,” “Content-driven approach,” “Diffusion,” “Sentiment-based influence,” “Similarity,” “Tweet quality,” “Event detection,” “Information quality,” “Query expansion,” “Social information retrieval,” and “Social recommendation.” Many of them were used combined with the “AND” Boolean operator in conjunction with terms such as “OSNs,” “online social networks,” “social media,” and so on, as a case-insensitive search.

Next, we performed a review on the selected studies to highlight the most relevant topics and subtopics related to the influence in OSNs and social semantics. In the final step, we have further filtered the selected works based on their date of publications, thus keeping the most recent. However, to not exclude the older but significant related works (with high citation counts), we described their impact in the newer works that have cited them. In this way, we kept our selected works quite up-to-date, including the most recent works with respect to this work. Finally, the selected publications consist of three types, namely, peer-reviewed journals, international conferences and workshops, as well as white papers in acknowledged academic repositories and archives. Figure 2 shows the distribution of the selected works in terms of their publication year and type. More than half of the publications have been published recently (55% after 2014). The distribution according to their publication type is 37% (journals), 60% (conferences), and 3% (white papers), respectively.

---

1https://scholar.google.com.
2https://arxiv.org/.
3https://www.ssrn.com/.
4http://dl.acm.org/.
5http://ieeexplore.ieee.org/Xplore/home.jsp.
6https://www.elsevier.com/catalog?producttype=journals.

ACM Computing Surveys, Vol. 53, No. 1, Article 7. Publication date: February 2020.
The remainder of the article is organized as follows: The next section presents the related literature, stressing the differentiation and the added value of this review along with our contributions. Section 3 defines online social influence and its effects in user generated content. In Section 4, we analyze the role of semantics and why they are significant to receive valuable and tangible insights among users and their social communities. Then, in Section 5, we highlight the qualitative assessments and positive impact of modeling the dynamic properties of influential content through semantics in OSNs. Finally, the last section concludes our review.

2 RELATED WORK

In this section, we present other survey papers from the related literature that tackle similar issues in terms of modeling online social influence with semantics. Then, we highlight the differentiation and the added value of this review, as well as our contributions across our classification scheme.

2.1 Similar Surveys

As already mentioned, the reviewed works were classified into the hierarchical scheme depicted in Figure 1, resulting in 20 hierarchically structured categories. For the purposes of this extensive review, we also considered other survey papers that tackle similar aspects in terms of the impact of influence in OSNs and the role of semantics [15, 85, 98, 106, 107].

More specifically, the authors in Reference [15] focused mainly on the classification of current diverse measurements aimed at discovering influential users in Twitter. Their range varies from those derived from naive metrics retrieved from the Twitter API to the adoption of PageRank algorithm and its variations. Other important factors include the content of the messages, their quality in terms of likeability by others, as well as the activity and popularity of the users. Thus, the authors of Reference [15] covered four aspects of our suggested scheme, namely, “Influence Metrics,” “Network/Graph Properties,” “Social Matching: Topic and Event-oriented,” and “Qualitative Assessment.”

In Reference [106], the authors analyzed a variety of OSN-based measurements and examined factors capable of affecting user influence. Those metrics were grouped under various criteria derived from:

- Neighborhood attributes, including number of influencers, exposure to direct and indirect influence.
• Structural diversity metrics that quantify the activity of the communities.
• Influence of locality and decay.
• Temporal measures including time delay until the reposting of a message.
• Cascade-based criteria, including its size and path length of messages.
• Metadata existence, including the presence of links, mentions, or hashtags.

Moreover, experiments were performed to predict user influence by using machine learning algorithms, with the aforementioned measurements as features. Based on our classification of this work, the survey described in Reference [106] covers the “Applications: Ranking” and “Network/Graph Properties” categories.

The work in Reference [85] presents an overview of studies regarding Adaptive Seeding (AS) methodologies to solve the Influence Maximization (IM) problem. IM is the process of discovering and activating a set of seed influential nodes-users to initiate the diffusion process so the largest number of nodes is reached or influenced. Often, that set of users is restricted to those who are engaged with the topic of interest, and due to structural dependencies of the network can be placed in lower ranking positions. As both IM and AS methodologies include the activation of nodes that in turn propagate the received information and activate others, the work described in Reference [85] covers only the “Information Flow and Influence: Propagation-oriented” category as described in our survey.

The authors in Reference [98] have reviewed approaches that enable Information Retrieval (IR) tasks in OSNs, which exploit content and structural social information. The research works the authors have reviewed have been classified into three categories according to the use of social information. Specifically, the “social web search” category includes techniques where social content is used to improve classic IR processes such as re-ranking of retrieved documents, query reformulation, expansion or reduction, and user profiling. The second category, called “social search,” includes methodologies on information discovery based on users’ generated content, interactions, and relationships. Finally, “social recommendation” aims at predicting users’ interests and is based on content-based and collaborative filtering approaches. Hence, the survey in Reference [98] covers the aspects of “Social Matching: User-oriented,” “Network/Graph Properties,” and “Applications: Recommendation,” as described by this work.

Finally, the authors of Reference [107] provide an overview on various user classification methodologies in OSNs. More specifically, they describe the most common frameworks based on machine (i.e., Bayesian, Decision Tree, Logistics, SVM, and KNN) and non-machine (concept of entropy and based on user similarity) learning techniques. The aim of these methodologies is to classify users into certain categories according to their explicit or implicit features, such as behavioral attributes, profile information, interests, viral content, and interactivity. As a result it covers only the “Social Matching: User-oriented” category as we have presented in this survey.

### 2.2 Review Differentiation and Extension

Table 1 provides comparative insights of this work with respect to the surveys described in References [15, 85, 98, 106], and [107]. It consists of three columns. For every single survey, the first two represent the category and sub-category according to our classification scheme (Figure 1), as well as the respective section where we analyze it. The third column depicts the respective reference number. The mark “✗” is placed in case where—according to the best of our knowledge—there is no other similar survey that covers this category.

Thus, comparing to the surveys presented in Section 2.1, this review aims at covering and analyzing four additional aspects of OSNs, namely, “Information Flow and Influence” (further categorized in “Propagation-oriented” and “Diffusion-oriented”), as well as “Social Modeling” and
Table 1. Classification of Referenced Surveys

| Category/Subcategory | Section | Ref |
|----------------------|---------|-----|
| Influence Metrics    | 3.1.1   | 15  |
|                      | 3.1.2   | 15  |
|                      | 3.1.3   | 15  |
| Information Flow and Influence | 3.2.1 | 85  |
|                      | 3.2.2   | ✗   |
| Network / Graph Properties | 3.3    | 106, 15, 98 |
| Applications         | 3.4.1   | 106 |
|                      | 3.4.2   | 98  |
| Social Modeling      | 4.1     | ✗   |
| Social Matching      | 4.2.1   | 107, 98 |
|                      | 4.2.2   | 15  |
| Community Detection  | 4.3     | ✗   |
| Qualitative Assessment | 5      | 15  |

“Community Detection” (Table 1). Moreover, the differentiation and extra issues covered in this work can be summarized in the following points:

- Information Flow and Influence: In contrast to many research works where the terms “diffusion” and “propagation” are used interchangeably, we tried to explicitly differentiate them by providing a clear distinction between their impact and role in the disseminated information.
- Social Modeling: We include studies aiming at the transformation of unstructured social data into Linked Data, by (i) relating entities to knowledge bases (e.g., Google Knowledge Graph, DBpedia) and (ii) representing them as concepts extracted from ontologies using semantic vocabularies.
- Community Detection: We consider approaches that also employ social semantics and ontologies. Such approaches are not only useful for the analysis of OSNs, but also for gaining deep insights into the structure and characteristics of complex networks.

2.3 Our Contributions

In this work, we also place our approaches and proposals across the review classification scheme, and moreover with respect to:

- **Influence Metrics**: In References [111] and [112], we present a publicly available service\(^7\) aiming at calculating and ranking the importance and influence of Twitter accounts. Specifically, we define “Influence Metric,” which value depends on three parameters, namely, (i) the activity of a Twitter account (e.g., tweets, re-tweets, replies, mentions), (ii) its social degree (e.g., followers, following), and (iii) its impact in Twitter (e.g., content diffusion, social acknowledgment).

- **Information Flow and Influence**: Propagation-oriented and Diffusion-oriented: In Reference [110], we examined the propagation features and patterns of the Reddit social network. Specifically, the study is focused on the virality, lifespan, flows of information, and the speed of diffusion of user generated content across other OSNs.

\(^7\)http://www.influencetracker.com/.
• **Qualitative Assessment**: By extending the work described in Reference [111], we introduced a new qualitative factor based on the established $h$-index metric ([112]). Its aim is to reflect other users’ actions (e.g., retweets) and preferences (e.g., favorites, mentions) over the content and properties of viral posts, thus enhancing the “Influence Metric” of real influencers in Twitter.

• **Social Modeling**: In References [112] and [114], we presented an ontology for social analytics. Specifically, the “InfluenceTracker Ontology”\(^8\) is capable of modeling structural aspects of Twitter accounts, including information of their owners, the used entities (i.e., user mentions, replies, hashtags, photo and content URLs), as well as their online social relationships. To provide a five-star data model, according to Tim Berners-Lee’s Linked Open Data (LOD) rating system [25], in References [114] and [93], we extended our ontological schema by incorporating properties from DBpedia, foaf\(^9\) and yago\(^10\) ontologies. Since the latest update of the LOD\(^11\) cloud, on 20/02/2017, the InfluenceTracker dataset is officially part of this interlinked and interdependent ecosystem of data. Finally, in Reference [93], we present an approach for annotating OSN accounts employing DBpedia thematic labels.

• **Social Matching**: In Reference [113], we proposed a methodology towards the discovery and suggestion of similar Twitter accounts. We consider term matching of user-generated content for all possible Twitter entities that may be used (mentions, replies, hashtags, URLs) according to References [112] and [114]. Finally, we contributed in the field of users’ query expansion in References [115] and [105]. Specifically, we propose an algorithmic approach that expands a user’s query by creating a suggestion set of the most viral and up-to-date Twitter entities (e.g., hashtags, user mentions, URLs).

### 3 ONLINE SOCIAL INFLUENCE

In this section, we describe one major aspect in OSNs, namely, online social influence. We focus on how influence can be measured or predicted and the methodologies (e.g., based on topology, diffusion, or social authority) that can be used to measure influence along with the respective application domains.

#### 3.1 Influence Metric

The topic of measuring user influence on OSNs, along with the identification of opinion leaders on them, is not recent. It includes multiple scientific areas that span from social sciences to viral marketing and from daily communication to OSN platforms [113]. In past publications, there is no clear definition of the “influential user,” thus, the term “influence” has multiple interpretations. While it is rather expected that having many followers could be a very good approximation of being influencer (as described in References [116] and [117]), most recent studies do share a common result, which is that the most active or popular users are not necessarily the most impactful ones. In fact, the authors of Reference [123] define “three central forces,” namely, social influence, entity similarity, and online structural equivalence, that can be related to social contagion and network diffusion [124, 125]. Moreover, they explicitly define social influence as the ratio of social strength interactions from the influencer to the affected node divided by all social strength interactions the affected node receives from the whole community. The ratio can be between 0 and 1, with values close to 1 define higher influencers and vice versa.

\(^8\)http://www.influencetracker.com/ontology.
\(^9\)http://xmlns.com/foaf/spec/.
\(^10\)https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444.
\(^11\)http://lod-cloud.net/versions/2017-02-20/lod.svg.
The works presented in this section discuss issues related to the discovery of influence, and we classify these issues into three categories according to the way they (a) exploit direct social information (number of followers, followees, social content, and so on), (b) incorporate PageRank and related hyperlink-based algorithms, and (c) employ machine learning techniques.

3.1.1 Direct Social Information Metrics. The study in Reference [2] proposes the "Social Networking Potential" as a quantitative measurement for discovering influential users in Twitter, and suggests that having a large number of followers does not guarantee high influence. The methodology utilizes the number of tweets, replies, retweets, and mentions of an account.

The authors in Reference [12] introduce three types of influence for Twitter users, namely, "In-degree" (how many followers the users have), "Retweet" (how many users’ posts get retweeted), and "Mention" influence (how many times a user gets mentioned in posts of other users). For calculating the aforementioned metrics, the users should post at least 10 messages. According to the authors, the "Retweet" and "Mention" measurements inter-correlate well enough apart from the "In-degree." Consequently, the study concludes that the most followed users are not necessarily influential.

Influence in terms of activity or passivity for Twitter users is studied in Reference [22]. To conduct this study, a corpus of tweets is utilized including at least one URL, their creators, and their followers. The derived measurement is based on the "Follower-Following" relationship of the users, in addition to retweeting patterns. As in most studies in this area, it is stated that "the number of followers a user has is a relatively weak predictor of the maximum number of views a URL can achieve."

Another metric for measuring the importance of a Twitter account, called "Influence Metric," is proposed in Reference [111]. In this case, a social function is presented that incorporates the activity (or passivity) of an account, along with its numbers of followers and followees. Moreover, the same authors in Reference [112] improve the aforementioned metric by introducing a new qualitative factor according to the established $h$-index factor, thus enhancing the influence of quality users. H-index is also used in Reference [22] as a predictor of high achievements for retweets containing URLs.

In Reference [3] the "h-index" metric is proposed, aiming at calculating the importance of users on a particular topic. It also relies on the "$h$-index" factor and indicates how many times the message of a user on a certain topic has been reposted. The work found that high influence on one topic does not necessarily mean the same on other topics.

A framework exploiting influence for evaluating and enhancing communication issues between governmental agencies and citizens in OSNs is proposed in Reference [89]. The aim here is to evaluate the quality of the agencies’ responses with respect to the citizens’ requests, to analyze the citizens’ sentimental attitude and their subsequent behaviors, and to suggest influential users to the agencies to obtain new audience. To achieve these, several components are incorporated into the framework, which detect the demographics of the followers, their locations, topics of interest, and sentiments.

The authors in Reference [91] propose a different kind of influence called the "susceptibility to influence." Its metric estimates how easily a Twitter user can get influenced. The proposed metric utilizes the user’s social interactions that depend on three factors, namely, activity, sociability, and retweeting habit. Activity reflects users’ tendency to interact with friends and consequently the chance to become influenced by them, while "sociability" corresponds to the users’ social degree among their activities, implying that interactions with more friends result in a wider diversity of topics and interests.

Finally, the study in Reference [97] presents a methodology for measuring social influence in mobile OSNs by incorporating its entropy. Specifically, the friend and interaction frequency
entropies are introduced to highlight the impact of complexity and uncertainty when measuring influence. A weighted network is constructed based on users’ interactions, upon which three types of influence are introduced, namely, (a) direct influence among related users, (b) indirect influence among unrelated users, and (c) global influence that covers the whole network.

3.1.2 Hyperlink-based Metrics. The studies in this sub-section describe influence metrics through hyperlink-based algorithms (e.g., PageRank), and therefore influence is strongly related to the structure and the topology created by the OSN itself.

The authors in Reference [77] propose an influence model based on two aspects: user relationship and activity. They used three factors, namely, Influence Diffusion Model (IDM), PageRank, and usage behavior. IDM focuses on tweets and their reply chain, while providing the influence of propagation based on word occurrence. The PageRank algorithm is employed for calculating users’ significance based on their relationships, while the user behavior factor affects a user’s influence score as it is based on the number of posts, mentions, follower, and retweets. The core ideas of these models are extracted and are integrated into the proposed influence model.

An influence-ranking method is proposed in Reference [19] based on the fact that user influence is determined by the followers’ influence contribution, which in turn highly depends on their interactions. A user can influence another if the former disseminates more messages related to those of the second user. The proposed measurement is a variation of the PageRank algorithm, is based on the eigenvector centrality, and depends on the similarity between tweets over a graph consisting of users (following and mentions) and messages (retweets and replies).

In Reference [8], the authors proposed a framework for identifying Twitter users’ asserting positive or negative influence. Towards this end, three factors are considered, namely, the users’ structural position, the sentiment, and the textual quality of their posts. The first factor combines the betweenness, the eigenvector, as well as the in-degree centrality measures of the users’ follow-up relationship and interaction (mentions, replies and retweets) networks. The authors employ sentiment analysis techniques to classify the polarity of the tweets as positive, negative or neutral. Finally, the quality of the messages positively affects the degree of influence, and it depends on how well-written and comprehensive those messages are.

In Reference [18], the proposed “Influence Rank” metric implements a modified version of the PageRank algorithm based on the eigenvector centrality, which considers the structure and topology of the network. Specifically, it consists of user relationships (follower/following), retweets, mentions, and favorites for the detection of “opinion leaders” who are able to influence others. The authors conclude that to be influential, a user should have influential followers.

In Reference [21], the authors present a variation of the PageRank for introducing two metrics based on the eigenvector centrality, called “InfRank” and “LeadRank,” which are based on following, retweeting, and mentioning relationships among users. “InfRank” calculates the user impact in terms of his/her capability to disseminate information as well as to get retweeted by other influencers. “LeadRank” measures the leadership of a user with respect to reproducing content and stimulating social actions from other users.

Finally, in Reference [26], the authors present the “MISNIS” framework whose goal is to discover influential Twitter users on a given topic. The framework does so by applying the PageRank algorithm on a graph representing users’ mentions found in Portuguese tweets. Moreover, sentiment analysis is performed, classifying the messages into three categories, namely, positive, neutral, and negative. This work differentiates itself from others in this field in the way that the topics are detected. Instead of performing naive string matching based on the characters of a hashtag, a fuzzy word similarity algorithm is applied utilizing all the contents of a message. Consequently, more relevant tweets for a topic are retrieved despite not containing the exact hashtags or other user-indicated keywords.
3.1.3 Metrics Based on Machine-learning Techniques. In Reference [87], social influence is measured by applying the “InfluenceRank” framework. The authors of the work employ Twitter account profile features (i.e., number of tweets, followers, following, member of lists) and tweets over a two-month period. Their approach makes use of regression-based machine learning model, having “InfluenceRank” calculate the impact of influence across the extracted account features. Although the work seems promising, the authors claim that, due to the small number of instances in the training set, the model is not accurate enough.

Another machine learning framework for discovering popular persuasive users is presented in Reference [99]. The authors’ persuasiveness metric is pair-wise and is based on three factors: influence, entity similarity, and structural equivalence. Influence depends on the strength of social interactions among users, entity similarity measures how close two profiles are, while structural equivalence measures the structural similarity of two entities according to a distance function. Each of these factors is assigned a probability that denotes the likelihood of persuasion.

The work presented in Reference [7] proposes a framework for predicting user influence by combining textual and non-textual attributes. More specifically, the authors employ the user’s basic social information metrics (e.g., the number of followers, followees, mentions, and replies), and then by utilizing statistics over the textual data of the tweets, as well as non-linear learning methods and machine learning techniques, a strong prediction performance metric is derived.

Finally, the authors in Reference [28] propose a machine learning methodology for investigating the impact of profile information towards the increase of Twitter accounts’ popularity, in terms of their followers’ count. Based on the assumption that given names and English words affect the discoverability, profiles were analyzed and categorized into three groups according to the lexical content of the accounts’ name field. The framework consists of three stages to evaluate the popularity dynamics in terms of: (a) the content of the name field, (b) the profile features, and (c) the incorporation of those features in a classifier that identifies the accounts that are likely to increase their popularity. Classification models for each group are different, as they are adapted according to their respective features. The results showed that the existence of known terms in the name field and the provision of other profile information (e.g., description, profile image, URLs, location) have a strong impact on the number of followers.

3.2 Information Flow and Influence

Information flow is vital in all kinds of networks (e.g., social, digital, or computer), and can be affected by the actions or properties of their actors and the sets of dyadic relationships between them. Influential users determine the virality of information and specifically how such information is propagated or diffused. As already mentioned, although propagation and diffusion are often used interchangeably, in this survey, we examine them separately. Diffusion defines the spread of information from a starting node towards the rest of the network, while propagation takes into consideration the intermediate nodes as well, which receive, process, and further decide how to handle information. Thus, this topic is divided into two subtopics, namely, propagation-oriented and diffusion-oriented. The propagation-oriented approach considers works that employ the propagation of information in OSNs to discover and calculate the impact of influential users, whereas the diffusion-oriented approach provides the insights into the identification of influential users being able to boost the diffusion of information in OSNs.

3.2.1 Propagation-oriented Approaches. The authors in Reference [17] propose an extension of PageRank for measuring influence. They applied their extended PageRank approach based on the eigenvector centrality, on a graph of retweets, and user relationships and consider social diversity of users and transmission probabilities of the messages based on the hypothesis that users inherit
influence from their followers. The aim is to explore whether individual characteristics and social actions as well as influence propagation patterns are factors capable of influencing other users.

Similarly, as described in the previous subsection, in Reference [21] the measurement of social influence is based on the eigenvector centrality, by using a variation of PageRank. Specifically, the authors measured the propagation of user influence into the network based on the users’ ability to stimulate social actions of others, such as retweets and mentions.

In several cases, the influence metric derived correlates the information propagation with the user’s retweeting behavior. Such a study is described in Reference [22], where influence is used for measuring the activity or passivity of Twitter users.

The authors in Reference [76] propose a methodology to identify influencers in OSNs with the help of online communities that are discovered by applying propagation-based modes. In this case, the structural features (shortest path, closeness, eccentricity, betweenness, and degree) of each node are extracted, while their weighted representation is computed by considering all the features across the network. By using principal component analysis, the most influential nodes are discovered. By applying maximum flow algorithms, communities are detected, implying positive attitude towards the influencers.

Social influence and propagation can be used as input in recommendation systems. The proposed framework in Reference [14] calculates the influence that social relationships have on users’ rating behaviors and incorporates it into recommendation proposals. Two social influence related attributes are considered: users’ susceptibility, which is the willingness to be influenced, and friends with high influence.

While the above studies consider the propagation of information towards the discovery of influential users, there are many other works [27, 42, 75, 95] that describe frameworks for discovering the propagation of influence in Twitter and its impact on other users.

A methodology for modeling the dissemination of influence in social networks is developed in Reference [27]. The authors characterize influential users as those generating posts with high probability of being propagated, i.e., retweeted, and simultaneously having a large number of followers. Based on past information cascades, the communities of influential users are discovered along with the analysis of their social activities and relations.

In several works, an individual’s influence is calculated based on the volume of information spread over a network. For each influenced node, an algorithm calculates the number of other nodes that are affected. It is based on the hypothesis that the previously influenced nodes determine the number of new ones. The study described in Reference [42] found that the propagation of information is affected by the atomic impact of the nodes. Similar to the previous study [27], the proposed models are considered as stochastic processes in which, according to probabilistic rules, information is spread from a node to its neighbors. In a similar way, the study described in Reference [43] aims at identifying influential nodes by calculating the expected number of influenced ones.

A study analyzing the persuasion-driven social influence based on some topic of interest is presented in Reference [95]. Several influence measurements incorporate users’ social persuasiveness in terms of influence propagation for quantifying user-to-user influence probability. Based on proposed metrics, the framework exploits the topical information, the users’ authority, and the characteristics of relationships between individuals.

A multi-topic influence propagation model is proposed in Reference [75]. It is based on user relationships, posts, and social actions. The influence score consists of direct and indirect influence, where the former considers information propagation from retweets by the direct followers, while the latter takes into account the retweets from non-followers. Both of them are related to different topics. The distribution of users’ topics of interest is discovered according to the collected tweets.
Then, a topic-dependent algorithm is applied and a multi-topical network is created to identify multi-topic influential users.

A model for demonstrating how social influence can impact the evolution of OSNs by simulating influence propagation and activation processes is proposed in Reference [90]. In this model, two types of influence, namely, locality and popularity, are considered, since they have different impact on the network dynamics. Locality affects the information spread through social ties, while popularity has global impact on individuals, since it does not rely on network topology.

Influence Maximization (IM) problem is the process of discovering and activating a set of seed nodes to initiate the diffusion process so the largest number of nodes is reached or influenced. The authors in Reference [79] investigate the IM problem and propose a probability-based methodology that enables greedy algorithms that recursively estimate the influence spread to perform efficiently in large-scale social networks in terms of memory and computing costs. In Reference [96], the authors aim to maximize influence propagation by selecting the most influential intermediate nodes. Therefore, a new optimization problem is formulated that explores the idea of routing multi-hop social influence from the source to a specific target with some time constraint. To achieve this, the topology of the network, the users’ influence, and the responding probability in a specific time frame are taken into consideration. The authors in Reference [92] propose a content-centered model of flow analysis to investigate the IM problem. Moreover, the analysis is not based on the users’ relationships but on the content of the transmitted messages. The authors apply an algorithm to discover the information flow patterns using content propagation patterns. Then, the influencers are discovered by exploiting those patterns, their position, and the number of flow paths they participate in. A different approach on the IM problem is proposed in Reference [109] and is described as “boost set selection.” The authors claim that it is possible to improve the diffusion process of a subset of the initial seed nodes by using additional resources such as by giving out free samples of a product, engaging in gamification, or other marketing strategies to become more influential.

An extension of the IM problem, described as “Influential Node Tracking,” is defined in Reference [88], where the authors focus on the dynamic identification of influential nodes that contribute to the maximization of influence spread at any time. Due to the dynamic nature of the networks, their structure and influence degree of the edges constantly alter. Consequently, the source nodes that maximize the influence spread should also be continuously adjusted. To achieve their goal, the authors compare consecutive snapshots of the network based on the fact that it is unlikely to drastically change, thereby resulting in great structural similarity.

Finally, there are other works [40, 44, 110] that investigated the discovery of information propagation flows in OSNs. In Reference [40], a dissemination network is created based on user mentions, with constraints on topical similarities in the respective posts. The authors claim that the frequency of mentioning a user is among the stronger predictors of information propagation. Similarly, the authors in Reference [44] examine information propagation on Facebook that provides insights into information shared by friends. They found that the individuals being aware of those insights are more willing to re-share the information faster, compared to those who are not. Although these strong relationships are more influential separately, the weaker bonds, exceeding in number, are responsible for the propagation of information.

3.2.2 Diffusion-oriented Approaches. In Reference [37], the authors investigated diffusion issues with an improved version of the K-core method [83]. The authors incorporate a linking and weighting method based on the hypothesis that users’ interactions, namely, retweets and mentions, are significant factors for quantifying their spreading capability in a network. In Reference [104] the authors propose the “SIRank” metric for measuring users’ spread ability and identifying
influential ones. Initially the users’ spread influence is measured by analyzing the information cascade structure. As each user’s influence is directly related to his/her interaction influence with others, pair-wise metrics are calculated by measuring retweeting contributions, users’ interests and closeness, activity frequency, and retweeting intervals. The authors measure users’ spread influence by classifying cascade and user interaction patterns across the diffused information.

Similarly, the main objective in Reference [101] is to investigate the diffusion of messages and users’ influence, based on the retweet cascade size and its attenuation patterns. The proposed influence measurement depends on the number of users who could potentially get a message either directly or via retweets. The latter affects the proposed cascade size metric and sets the upper limit of users who could potentially see that message. The study concludes that the user who has the most followers produces the largest cascades, while most cascades are eliminated after two or three frequency peaks.

The “retweet” functionality and the retweet counter can be considered as a factor for measuring the “interestingness” of a user’s tweets [33]. Based on that, the resulting spread of information is examined in Reference [11]. The authors propose that the number of retweets is an indicator of popularity of both the messages and their authors. According to the authors, as soon as a post is retweeted, it will nearly immediately get diffused up to four levels beyond its source, thus resulting in rapid spread since the first retweet. Three measures of influence, namely, the number of followers, PageRank, and number of retweets, were further compared and evaluated. The results indicated that, in contrast to the third measurement, the first two provide similar rankings of influential users, meaning that influence is not explicitly related to the number of followers and the popularity of tweets. Similar to the results of Reference [101], the average number of additional recipients is not affected by the number of followers of the tweet source. Thus, the tweet is likely to reach a certain number of audiences via retweets.

A different interpretation of the term “influence” is given in Reference [9], where the authors relate the user’s posting activity (and thus influence) with the diffusion of the URLs included in posts through retweets. The influence score with respect to a posted URL is calculated by tracking the diffusion of the URL from its source node until the diffusion event is terminated. The work is similar to the one described in Reference [77], where the influence measurement is related with the Influence diffusion model, which provides the influence of topical spread. However, it differs from Reference [22] in that the diffused influence is studied in terms of activity or passivity of Twitter users by solely analyzing the retweeting behavior of the latter.

In addition to the point of views discussed above, the following studies involve methodologies for analyzing information diffusion and factors that affect it in OSNs. As already described in the previous subsection, the authors in Reference [40] claim that, despite the fact that some properties of the tweets predict great information propagation, the users’ mention rate is the strongest predictor. Information diffusion in Twitter and Digg is studied in Reference [41]. According to the study, information flow and spread are affected by the structure of these networks. Information in networks with sparse and poorly interrelated structure (e.g., Twitter) reaches nodes slower in comparison to networks with dense structure (e.g., Digg), in which information spreads faster. Due to the structure of Twitter, information may spread at a slower pace, but it maintains its diffusion at the same rate as time passes, thus penetrating the network further.

In Reference [44], information spread that is related to exposing insights about information shared by friends is explored on Facebook. The study concludes that social ties greatly affect users’ behavior on re-spreading information in the network. In another work on Facebook, the authors studied diffusion trees of fan pages. The results indicated that there is no solid evidence that a node’s maximum diffusion chain length can be predicted [10].
The ways in which widely used hashtags spread through interactions among Twitter users are analyzed in Reference [45]. Variations of entities (e.g., hashtags) and topics are due to the differences in the spread probability, and differences in the extent to which repeated exposures to hashtags continue to affect their diffusion into the network by other users. The authors in Reference [108] extend their previous work [24] to identify the initial set of users who are able to maximize information diffusion. Initially, the users’ diffusion patterns are recognized by exploiting their posting activities and history. The proposed algorithm combines them with propagation heuristics to achieve the diffusion coverage in the network.

Finally, the authors in Reference [110] studied the lifespan and information flows of another social network (Reddit) based on user-generated content. They were particularly interested in the virality of information and its speed of diffusion in other OSNs.

### 3.3 Network/Graph Properties

The studies presented herein utilize the topology and the structure of the OSNs to measure influence or to discover other social dynamics. Usually, in this domain, only a fraction of the whole network is used due to hardware (e.g., RAM, Hard Disk Drive) or complexity limitations.

The framework proposed in Reference [6] aims to automatically identify influential users in topic-based communities according to the follower/following network, as well as user mentions and replies. The authors consider network edges’ directionality as well as a proposed metric for measuring a so-called “external importance.” As already mentioned in References [21] and [77], influencers are discovered by applying PageRank and newly proposed link-analysis algorithms that are exploiting the topology and properties of the network, including posting, retweeting, and mentioning relationships among users. In Reference [78], influence is measured by applying a hybrid framework that integrates both users’ structural location and attributes. A user’s location is found by applying well-known graph analysis algorithms such as the in-degree, the weighted in-degree, the eigenvector, and PageRank, while the attributes (i.e., activeness) are measured by adapting the contribution measurement, which is used by Flickr.

The authors in Reference [76] propose a methodology for the identification of influencers by exploiting structural features. Specifically, the shortest path, closeness, eccentricity, betweenness centrality, and degree of each node are extracted, and their weighted representation is computed by considering all the features across the network. The most influential nodes are discovered by using principal component analysis. Moreover, by applying maximum flow algorithms, communities are detected. The identified communities imply a positive attitude towards the influencers.

The study in Reference [118] describes five graph centrality metrics used in the literature to identify influential entities in OSNs based on their topological position. The behavior of these metrics, namely, the degree, closeness, betweenness, eigenvector, and Katz centrality, is compared against three generic statements when new relationships are introduced. The results showed that only one of the centralities satisfied all statements, while the rest of them only partially.

In the work described in Reference [119] the authors conducted a randomized field experiment to identify the best candidate nodes for spreading information in OSNs. This is achieved by using the proposed “diffusion centrality,” a network centrality metric capable of predicting information diffusion by an individual. This metric is based on the degree, eigenvector, and Katz-Bonacich centralities.

In Reference [90], the authors proposed a framework to demonstrate how social influence can impact the evolution of OSNs by simulating influence propagation and activation processes. In this framework, two types of influence are introduced that have different effects on the network dynamics. The first type is “locality,” which affects information diffusion through social ties, while
the second is “popularity,” which does not rely on network topology but has global impact on individuals.

All above studies try to identify influencers according to the information derived in particular periods of time, similar to a compilation of different and static sequences. Below, we analyze other works where related issues are considered under properties and concepts that belong to dynamically evolved and complex networks.

Moreover, the authors in the work described in Reference [80] consider influence analysis on dynamic networks that evolve over time. The indexing method is able to recognize and incorporate all graph updates to efficiently answer the queries on influence estimation and maximization according to the latest graph version. The optimized techniques proposed aim at reducing time and space requirements.

In Reference [88], the “Influential Node Tracking” problem is defined as an extension of the “Influence Maximization” one dynamically evolving networks. Due to their nature, the structure and influence degree of the edges continuously change. Therefore, the authors deal with the dynamic network as a set of static ones and compare consecutive snapshots under the assumption that it is not likely to drastically change its structure.

Another work in this area is presented in Reference [100]. A network is modeled and updated dynamically as a continuous flow of information about edges and weights. Under the assumptions of the linear threshold model, two aspects are considered: the discovery of nodes having influence greater than a predefined threshold, and the discovery of the most influential nodes. The sample random paths are gradually updated when the network changes by considering efficiency in terms of both space and time usage.

Apart from discovering influential users, the topological and structural attributes of the social networks can be used towards context-based identification of users’ interests and similarities. For example, a community detection approach is proposed in Reference [50] using node similarity techniques. Towards this end, a virtual network is created. In accordance with node similarity, whose value is derived from the Jaccard Measure in the original network, virtual edges are added. Then the proposed methodology is applied on the resulting network.

Similarly, the author in Reference [68] proposes a semantic followee recommender system in Twitter that exploits users’ tweets to build their interest profiles. An interest graph is created by using specific semantic knowledge graphs that contain a variety of topics. These topics are then mapped and suggested to the users. User interest metrics are calculated using graph theory algorithms such as the Steiner Tree and the “InterSim” (Interest Similarity) one.

Another context-oriented approach is presented in Reference [71], where the context of Twitter posts is retrieved using the DBpedia knowledge base and graph-based centrality theory. In particular, the proposed centrality factor considers both the closeness centrality of a node and its shortest paths to all of its successors. An entity graph of contextualized and weighted content from Twitter is constructed, and two types of similarity metrics are introduced. The “local” similarity measures the contextual proximity of the entities being compared, while the “global” similarity is calculated based on the user’s request and the available tweets.

3.4 Applications

In this section, we consider the influence metrics presented in Section 3.1 to present research efforts that provide solutions for opinion makers, data analysts, and information scientists as services or applications. This topic is further divided into two subtopics, namely, (a) rank-oriented and (b) recommendation-oriented.

3.4.1 Ranking. To rank OSN users according to specific social attributes, the work described in Reference [5] presents a qualitative measurement of tweets that affects their authors’ influence
to propose a topical tweet-oriented user ranking. The degree of quality depends on the topic’s similarity, retweeting patterns, and influence degree for that particular topic of the users who retweeted a post. In Reference [19], the authors propose an influence-ranking method under the assumption that the user influence is based on the followers’ influence and their interactions. The authors found that user A can be more influential with respect to user B if user A posts messages that are strongly relevant to user B. The proposed measurement is a variation of the PageRank based on the eigenvector centrality and depends on the similarity between tweets over a graph of users (i.e., following and mentions to other users) and messages (i.e., retweets and replies).

The authors in Reference [70] propose a framework for discovering topic-specific experts in Twitter by employing two distinct metrics. First, the users’ global topic authority (influence) is calculated offline by exploiting three types of relations, namely, “follower,” “user-list,” and “list-list.” Second, the similarity between the users’ generated tweets and that topic is computed online. By leveraging the users’ topical influence and similarity, those who have the highest-ranking scores are regarded as experts in that domain.

The problem of topic-sensitive opinion leaders’ identification in online communities is also investigated in Reference [84], where a two-staged approach is presented. Initially, the opinion leaders’ expertise and interests are derived from their tags found at the description of the products. Then, a computational approach measures the leaders’ influence and ranks them according to not only the link structure of customer networks, but also according to their expertise and interests. The influence depends on the topical similarity between reviewers on a specific topic.

The authors in Reference [86] created “NavigTweet,” an influence-based visualization platform capable of revealing relationships of followers in Twitter and for identifying the top-k influencers. Those top influencers are detected according to their followers, followees, as well as their generated content (tweets, hashtags, URLs, retweets, favorites, and user mentions). Then, based on the above, a technique called “Analytical Hierarchy Process” ranks Twitter users.

An influence learning–based recommender is presented in Reference [102] for making suggestions to informative users whose posts are highly associated with those of target users. Ranking learning techniques analyze user behavior and model user preferences based on their social interactions (e.g., replies, likes). In another application described also in Section 3.2.2 the authors propose the “SIRank” metric for measuring users’ spread ability and identifying the influential ones [104].

Finally, “InfluenceTracker” is a publicly available service12 capable of calculating and ranking the influence and impact of a Twitter account [111, 112]. The authors introduce a social function, which incorporates the activity (or passivity) of a user, his/her popularity (e.g., number of followers and followees) and his/her social acknowledgment by other users.

3.4.2 Recommendation. The various studies presented in this section describe approaches on recommendation systems that utilize the available information in OSNs for proposing social content or accounts based on the users’ profiles. An interesting problem in the area of social network recommendation systems is to define a set of similar users to follow.

The friend recommendation problem in Flickr is studied in Reference [63], mainly from the viewpoint of network correlation. The authors assume the hypothesis that each user has many different social roles in OSNs. For each role, different social sub-networks are formed, which are aligned for the correlations among them to be found through weighted tag feature selection. When recommendations are made, the similarities of the tag features, among the new and the existing

---

12http://www.influencetracker.com/.
users, are calculated. The more similar the tags are, the more users we have who are similar in terms of those tags.

A semantic followee recommender system is proposed in Reference [65]. This system integrates content-based filtering approaches in Twitter, popularity identification among users using collaborative-filtering over the friendship network, along with publicly available knowledge resources (i.e., Wikipedia, WordNet, Google corpus). The aim is to classify the tweets into six classes and to label the users as a recommendation service. The application of the Kalman filter enables noise removal and the prediction of future tweet patterns, leading to the new multi-labeling of the users.

Similarly, the work described in Reference [68] exploits the users’ tweets for building their interest profiles and for producing recommendation over a semantic knowledge graph that contains a variety of topics. Using graph theory algorithms (as explained in Section 3.3), the authors can recommend similar users. A ranking-based followee recommendation scheme in microblogging systems according to the latent model is proposed in Reference [73]. To model user preferences, both tweet content (original posts and retweets) and social relation information (followers, followees) are taken into consideration. Another followee recommendation methodology that builds interest profiles is proposed in Reference [35]. These profiles are built by exploiting not only the users’ generated content but also of their directly related ones (followers, followees).

A framework for discovering similar accounts in Twitter based only on the “List” feature is proposed in Reference [64]. This functionality allows the users to create their own lists by adding any account they wish. The authors claim that this feature is considered a form of crowdsourcing. The hypothesis of the methodology is that when two accounts are contained in the same list, they should be similar or related to each other. Therefore, the proposed measurement relies on the number of lists that a specified account and a potentially similar one are listed together.

In Reference [38], a matrix factorization framework with social regularization is proposed for improving the accuracy of recommender systems. Social regularization includes two models for representing social constraints and are based on users’-friends’ similarity at an individual and average level. Each social link is then weighted in accordance with the similarity among the users, allowing the exploitation of friends based on the rating similarity.

As already described in the previous subsection, ranking learning techniques are designed to provide recommendation based on the analysis of user behavior, preferences, and social interactions [102]. In addition, as mentioned earlier in Section 3.2.1, we can use related attributes being propagated through the social network, because the effects of friends who have strong influence or are subject to be influenced by a user are highly related with recommendation processes [14].

The recommendation system proposed in Reference [103] is based on the users’ personal interests. In fact, explicit social features such as users’ topic-level influence, topic information, and relations are incorporated into a framework for improving recommendation results. Two kinds of influence are introduced: direct, which is identified by studying the communication records between users, and indirect, which is identified by applying social status theory for the discovery of latent relationships. In both cases, positive and negative influences are also identified. Moreover, topic information is added into the structural analysis of indirect influence. A distributed learning supervised algorithm is applied that considers the aforementioned influence measurements and provides the users’ forwarding behaviors, which can be leveraged to provide improved recommendations.

Considerable efforts have also been devoted to recommendation systems for suggesting personalized streams of information [34, 36, 75, 82].

“Buzzer” [34] is such kind of a service for proposing news articles to Twitter users. To achieve this, terms from both the users’ and their friends’ timelines are mined. These terms act as ratings
for promoting and filtering news content. The methodology described in Reference [66] is based on
the same principles but it also incorporates additional factors affecting the interest of a user on
a tweet, such as its quality, number of retweets, and the importance of its publisher.

URLs as a recommendation parameter in Twitter are examined in Reference [36]. The scope is
to direct the users’ attention through personalized suggestion mechanisms in Twitter posts. There
are three main pillars that should be considered in such kind of recommender, namely, the sources
of the URLs, the users’ area of interest, and social information.

The authors in Reference [75] propose a multi-topic influence diffusion model based on user re-
lationships, posts, and social actions. The influence score consists of direct and indirect influence.
The first is determined by information propagation (retweets) by the direct followers. The latter
depends on the retweets from non-followers. Both of them are related to different topics. Based
on the users’ collected tweets, their distributions of topics of interest are found along with their
generation probability. Finally, a multi-topical network is created to which a topic-dependent al-
gorithm is applied to identify the multi-topic influential users, while the most influential user will
be used during the recommendation process.

Finally, recommenders can also be used for suggesting items on users. The work described in
Reference [82] is based on the observation that users’ purchase behavior is influenced by both
global and local influential nodes that in turn define implicit and explicit social relationships, re-
spectively. Therefore, a dual social influence framework formulates the global and local influence
scores as regularization terms and incorporates them into a matrix factorization–based recom-
mendation model.

3.5 Comparison of Related Works
To provide comparative insights from the above reviewed articles that refer to online social influ-
ence, we provide Table 2. For each reviewed article, the first column denotes its category according
to our classification scheme (see Figure 1). It should be noted that in many cases, a study does not
fall within the scope of only one topic, thus demonstrating that research efforts are strongly related
to each other. In the rest of the columns, we place the reference number of the articles or a mark
of “No” (✗) in cases where the studies employ or propose metrics and characteristics based on:

(a) Relationship: followers (Fs) and followees (Fing);
(b) Behavioral/Conversational activities: posts (P), re-posts (RP), favorites/likes (FL), men-
tions (M), replies (R); and
(c) Domain/Content analysis: works that are applied on specific topics (T) or works that re-
quire content analysis (CA).

4 ONLINE SOCIAL SEMANTICS
In this section, we study the semantics and their role as the second major aspect of OSNs. Specif-
ically, we analyze related works based on Semantic Web technologies along with network theory
and graph properties for transforming unstructured data into Linked Data, topic identification,
detection of similar users and communities, as well as user personalization (e.g., interests, sugges-
tions, and so on).

4.1 Social Modeling
As the adoption of semantics and Linked Data increases, more works have emerged covering as-
pects of semantic modeling in OSNs. In this section, we present approaches that adopt semantics
for modeling the logical topology and structure of social networks and media as well as the inform-
ation they disseminate.
Table 2. Classification of Referenced Works

| Category                              | Relations | Activities | Context |
|---------------------------------------|-----------|------------|---------|
|                                       | Fs        | Fing       | P       | RP | FL | M | R | T | CA |
| Direct social information-based       | 2, 12, 22, 89, 111, 112 | 111, 112 | 2, 12, 22, 89, 91, 111, 112 | 2, 3, 12, 22, 91, 111, 112 | 91, 111, 112 | 2, 12, 89, 91, 97 | 12, 91 | 89, 12 |
| Hyperlink-based                       | 8, 18, 19, 21, 77 | 8, 18, 19, 77 | 8, 18, 19, 21, 77 | 18 | 8, 18, 19, 26, 77 | 8, 77 | 26 | 8, 19, 26, 77 |
| Machine Learning techniques-based     | 7, 87, 99 | 7, 87, 99 | 7, 12, 28, 87, 99 | ✗ | 12, 28 | 7 | 7 | ✗ | ✗ |
| Propagation-oriented Approaches       | 14, 17, 21, 22, 27, 43, 44, 75, 76, 79, 88, 90, 92, 92, 96, 109 | 14, 44 | 14, 22, 27, 43, 75, 79, 96 | 17, 21, 22, 27, 40, 42, 43, 44, 75, 90, 92, 92, 96 | ✗ | 21, 40 | 40, 44 | 40, 75, 95 | 75, 92 |
| Diffusion-oriented Approaches         | 9, 10, 11, 22, 33, 37, 41, 44, 77, 83, 101, 104 | 10, 41, 44, 77 | 10, 11, 22, 24, 41, 45, 77, 108, 110 | 9, 11, 22, 24, 33, 37, 40, 41, 44, 77, 83, 101, 104, 108 | ✗ | 37, 40, 45, 83 | 40, 44, 77, 104 | 40, 45 | 33, 77 |
| Network / Graph Properties            | 21, 50, 76, 77, 78, 80, 88, 90, 100, 118 | 76, 77, 78, 80, 100 | 50, 68, 71, 77, 78 | 21, 77, 90 | ✗ | 6, 21 | 6, 77 | 6 | 50, 68, 71, 77 |
| Ranking                               | 19, 70, 84, 86, 102, 111, 112 | 19, 86, 102, 111, 112 | 5, 19, 70, 84, 86, 102, 111, 112 | 5, 19, 86, 104, 111, 112 | 86, 102, 111, 112 | 19, 86, 104, 102, 104 | 5, 70, 84 | 19, 84 |
| Recommendation                        | 14, 34, 35, 38, 63, 65, 66, 73, 75, 82, 102, 103 | 14, 34, 35, 38, 63, 65, 73, 75, 82, 102, 103 | 14, 34, 35, 36, 38, 65, 66, 68, 73, 75, 82, 102 | 66, 73, 75 | 102 | 66 | 102 | ✗ | ✗ |

Fs and Fing refer to follow-up relationships, P to posts, RP to reposts, FL to favorite or liked posts, M to mentions, R to replies, T to topics, and CA to content analysis.

One of the first studies in this domain is Reference [60], where the use of a specific syntax is proposed for creating a common knowledge representation by incorporating RDF-like syntaxes into Twitter posts. The use of such statements enables users to freely define relations such as hierarchical or equality relations among hashtags.

The authors in Reference [29] propose a methodology for enhancing Twitter posts with semantics relationships introduced among persons, products, and events. These relationships are exploited towards the provision of query suggestions to the users.

Another work on the enrichment of Twitter messages with semantics is described in Reference [30]. Specifically, the authors attempt to create user profiles by exploiting Twitter posts by using Semantic Web technologies. To capture the users’ interests, the URLs of news articles found in tweets are utilized. Lexical analysis is applied on their content so the relationships between the entities in news articles (representing the interests) can be discovered, which are then semantically related to those tweets.

Social semantics can be exploited in the development of semantic recommender systems. Specifically, the studies in References [65] and [68], which were analytically presented in Section 3.4.2, propose two semantic followee recommender systems for Twitter. Their aim is to build user
interest profiles by exploiting the users’ posted messages [65, 68], friendship networks [65], and publicly available knowledge bases (i.e., Wikipedia, WordNet, Google) [65], which are then used during the recommendation process.

A framework for inferring user interests in Twitter is also proposed in Reference [69]. In contrast to the ones described above, it is based on the users’ followees and the content they consume, rather than their original posts. The proposal is based on the hypothesis that famous people maintain accounts that are being followed by a large number of users. The Wikipedia articles of the former are discovered, linking to a higher level of categories and hierarchies, which become an implicit expression of the users’ interests.

The methodology presented in Reference [48] generates weighted semantic networks created by comments from a Chinese social network. The methodology focuses on the “giant component” of the resulting network by reducing the computational complexity to identify larger communities.

The work described in Reference [55] associates tweets with a given event by utilizing structured information found in them. The initial pool of terms for the retrieval of the messages is manually provided. The final associations take place by applying query expansion techniques and by utilizing the relationships derived by the semantified data.

The authors in Reference [74] create graphs of hashtags found in tweets and utilize their relational information to discover latent word semantic connections in cases where words do not co-occur within a specific tweet. Noise and sparseness in the messages are handled by utilizing two types of hashtag relationships: (i) explicit ones that refer to hashtags that are found in a post and (ii) potential ones that refer to hashtags that are not contained in a post yet co-occurring with others. Finally, the hashtags and words that are most likely to appear on a specific topic are discovered.

The following studies employ Semantic Web technologies, ontologies, and the DBpedia knowledge base, which is a semantified version of Wikipedia, to achieve their goals.

The study in Reference [31] proposes a semantic data aggregator in Twitter, which utilizes Semantic Web vocabularies to transform social data into structured microblog content. The framework focuses on the provision of Twitter messages as user-driven Linked Data, and more specifically metadata associated with the authors and content of those social posts.

The framework proposed in Reference [32] also employs Semantic Web technologies, ontologies, vocabularies, and Linked Data towards the extraction of information about scientific events from Twitter. In this work, the authors introduce a methodology for identifying similar users and organizations according to geospatial and topic entities.

The authors in Reference [62] propose an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users. The entities found at the posts are mapped to classes of DBpedia ontology, using the DBpedia Spotlight tool, and are used for the labeling of clusters. Moreover, the topical similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters without predefining their target number.

Another work using the DBpedia knowledge base is Reference [71], where a framework is proposed for retrieving the context of posts in Twitter by applying graph-based centrality theory. Entities from tweets, in terms of words, are extracted and related to DBpedia URLs for semantic concepts to be discovered. An entity-weighted network for each tweet is then constructed according to graph centrality metrics.

The works in References [114] and [93] also attempt to semantically relate Twitter entities to DBpedia URLs. Contrary to the previous work, the entities under investigation are the authors of the posts, and the aim here is to provide a data model of five stars according to Tim
Berners-Lee’s Linked Open Data rating system [25]. Furthermore, in Reference [93] a framework is proposed to enable the labeling of Twitter accounts with DBpedia thematic categories. Similar to Reference [71], these categories are a result of the semantic concepts of the Twitter-DBpedia interrelation.

In References [112] and [114], an ontology aimed at the semantification provision of Twitter social analytics is presented. Specifically, the “InfluenceTracker Ontology” is capable of modeling structural aspects of OSNs, including the accounts, the users owning them, their disseminated entities, as well as their friendship relations. Moreover, the representation of qualitative metrics, such as the influence of the accounts or various likability and impact metrics, is also possible.

4.2 Social Matching

The studies presented in this section exploit the use of social semantics for identifying similar properties and activities with respect to user-generated content, description of real-life events, as well as revealing user interests and behavioral patterns across different online social media users. Thus, we divide this topic into two subtopics, namely, (a) User-oriented (e.g., similar user recommendation, user preferences) and (b) Topic and Event-oriented (e.g., topic profiling and user interest, event detection, product marketing, and others).

4.2.1 User-oriented Matching. Despite the fact that the set of social semantics of each account in an OSN is unique as they depend on personal social activities, common patterns among them can be recognized. These can be exploited to enable the discovery of users’ social behavior and preferences.

The study presented in Reference [1] describes a framework using supervised learning for distinguishing users in OSNs according to their influence and revealed the communities they belong to. The authors do not propose another influence measurement, but they introduce predictive properties associated with the users’ activity level and involvement in those communities. The supervised learning is built on social relationships (e.g., follower/following), interactions (e.g., mentions, replies), the topology and activity of the network, the users’ centrality (in-degree, closeness, betweenness, eigenvector), and the quality of the messages posted.

The aim in Reference [4] is also to identify influential users according to their behaviors over their own generated content on a specific topic. Toward this end, a graph model representing the relationships of the posts is created, which is then transformed into a user graph. The latter is used for the discovery of influential users according to the properties and measures found. Similarly, as described in Section 3.1.2, the authors in Reference [26] follow the same approach and apply the PageRank algorithm on that graph for the detection of influencers. The posts belonging to a specific topic are discovered through a fuzzy word similarity algorithm that utilized all the contents of the messages.

The work described in Reference [5] was presented in Section 3.4.1. Influencers are regarded as those generating tweets of high quality. Their quality is evaluated according to a set of parameters such as the topical similarity, the reposting patterns, and the influence degree on that particular topic of the users who retweeted those messages.

The framework proposed in Reference [6], as presented in Section 3.3, aims to identify influential users in topic-based communities. A measure of alpha centrality is employed on a graph derived from direct communications over the directions of network connections and a proposed metric of “external importance.”

Influential users are discovered in Reference [13] by applying an eigenvector centrality–based algorithm as an extension of PageRank, which considers both the users’ topical similarities and follow-up relationships. It is suggested that “homophily”—that is, the tendency of people to create
social relations with other individuals who have almost identical interests—is responsible for the majority of "Follower-Following" relationships. Moreover, the study claims that user activity does not necessarily imply high influence score.

In Section 3.4.1, we described a study [84] where the problem of topic-sensitive opinion leaders’ identification in online review communities is investigated. Toward this end, a two-staged approach is presented. Initially, the opinion leaders’ expertise and interests are derived from their tags found at the description of the products. During the next stage, a computational approach measures the leaders’ influence and ranks them according to not only the link structure of customer networks but also to their expertise and interests. The influence depends on the topical similarity among reviewers on a specific topic.

The task of the topic experts’ identification, namely, influential users on specific domains, is also presented in Reference [16]. A post-feature-based approach is proposed that utilizes nine kinds of features reflecting how the users interact. Their aggregation results in the production of three different kinds of influence measurement.

The authors in Reference [20] claim that a user’s influential level can be detected by considering the writing style and behavior within the OSNs. Therefore, they introduced 23 user profile features. Some of them are “presence of hashtags,” “URLs,” “self-mentions,” and “number of followers and tweets.” Similarly, they introduced 9 features of a tweet, such as “extension,” “frequency,” “quality,” “number of retweets,” and others. By applying machine learning algorithms, the most influential users are identified.

A framework-exploiting machine learning technique for discovering top persuasive users in OSNs is described in Reference [99]. The proposed persuasiveness metric is pair-wise and is based on three factors: influence, entity similarity, and structural equivalence. Influence derives from the degree of social interactions between users. Entity similarity measures how close two profiles are. Structural equivalence measures the structural similarity of two entities according to a distance function. Each of these factors is assigned a probability that denotes the likelihood of persuasion. A machine learning algorithm achieves the prediction of these probabilities.

The rebroadcasting behavior of users in OSNs is studied in Reference [94]. Specifically, a model is proposed that examines three aspects, namely, “role of content,” “content-user fit,” and “social influence.” The “content-user fit” measures the relation between the content of the message with respect to the user interests. As in Reference [91], influence measures the susceptibility of users for identifying those whose posts affect the reposting behavior of others. To discover the users’ interests, the well-known Latent Dirichlet Allocation (LDA) [59] methodology is applied on each message. The study concludes that the rebroadcasting of messages does not depend only on its content but also on its relevance with a user.

The LDA topic modeling approach is also applied in Reference [67], where a user centric topic discovery framework is proposed. The users’ tweets are analyzed for identifying their interests and for creating personalized topic profiles. Toward this end, a Part-Of-Speech (POS) tagger extracts the nouns of the tweets, which are provided to a search engine to retrieve the top documents based on their relevance. Using LDA on the content on those web pages, the final topics are provided.

Another framework, which was also described in another section (Section 4.1) of this survey, is employed for inferring user interests in Twitter [69]. Contrary to the previous frameworks, it is based on the users’ followees and the content they consume, rather than their original posts. The proposal is based on the hypothesis that famous people maintain accounts being followed by a large number of users. The Wikipedia articles of the former are discovered, linking to a higher level of categories and hierarchies, which become an implicit expression of the users’ interests.

The following studies exploit the social semantics in OSNs to propose query expansion techniques for providing an enriched coverage of information needs:
The study in Reference [56] describes a query expansion framework that takes into account the users’ preferences, which are derived by analyzing microblog posts and hashtags related to the targeted users.

Another query expansion approach is proposed in Reference [57]. It takes into consideration the similarity problem between tags and users’ profiles with respect to the query terms. Its aim is to assist users by refining and formulating their queries and by providing them with information relevant to their interests.

The research effort of Reference [55] we have presented in Section 4.1 attempts to associate tweets with a given event by utilizing their structured information. The application of query expansion techniques and the relationships derived from the semantified data result in those associations.

The story-tracking framework of Reference [72] is modeled as a pattern-mining and real-time retrieval problem. The most popular news stories, assigned with hashtags, are detected by mining frequent hashtag pattern sets. Using query expansion on the original hashtags, new story articles are retrieved. The pattern set structure enables hierarchical and multiple-linkage representation of the articles.

The authors in Reference [58] attempt to identify several hashtags relevant to a given query that can be used to expand it, thus leading to more accurate content retrieval. The proposed method is applied on microblog messages, and by leveraging some well-known statistical techniques, probabilistic language models are created over the hashtags found in that corpus.

Another study in the area of query expansion that considers the users’ input queries is introduced in References [115] and [105]. Specifically, the authors proposed an algorithmic approach that can dynamically create a set of suggested queries, according to viral content in Twitter, when considering the initial input query. In contrast to the work in Reference [58], apart from hashtags, the expansion set may also contain other accounts and URLs, resulting in a more enriched coverage of information needs. Moreover, the framework is not entirely based on statistical techniques and probabilistic models but on the capture-recapture methodology [46], thus making the set dynamic by providing a survival probability on the entities. The capture-recapture experiments are first introduced and applied to wildlife biological studies, where the species under study are captured, marked (type, timestamp, etc.) and then released into their environment. If an already marked subject is captured again on a subsequent trapping instance, then it is characterized as recaptured. According to the number of recaptured individuals of each species under study, critical metrics are estimated (i.e., population size, birth, death, and survival rate).

In Section 4.1 several studies [29, 30, 32, 62] utilize semantic technologies and related protocols to provide expanded query suggestions or to represent user preferences and similarities. Specifically, the authors of Reference [29] proposed a framework for enriching Twitter messages with semantic relationships that are extracted by analyzing microblog posts. These relationships are exploited towards the provision of query suggestions to the users and are identified among persons, events, and products. The authors introduce a methodology for identifying similar users and organizations according to geospatial and topic entities. The study in Reference [59] uses Semantic Web technologies for the creation of user profiles by analyzing Twitter posts. To capture the users’ interests, the URLs of news articles found in tweets are used. Lexical analysis is applied on their content to discover the relationships between the entities in news articles (representing the interests) that are then semantically related to those tweets. The framework in Reference [32] exploits ontologies and Linked Data for extracting information about scientific events from Twitter. Finally, an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users is proposed in Reference [62]. The entities found at the posts are mapped to classes of DBpedia ontology and are used for the labeling of clusters. Moreover, the topical
similarities among individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on these graphs, a quasi-clique community detection algorithm is applied for the discovery of topic clusters without predefining their target number.

In Section 3.3 and Section 3.4.2, we presented studies that attempt to adequately describe user characteristics, in order for similarities among them to be discovered. In Reference [38], a matrix factorization framework with social regularization is proposed for improving the accuracy of recommender systems. Each social link is weighted based on the similarity among the users, allowing the exploitation of friends differently according to the rating similarity.

The friend recommendation problem in Flickr is studied in Reference [63] from the viewpoint of network correlation. The authors assume that each user has many different social roles in OSNs. During each role, different social sub-networks are formed that are aligned in order for the correlations among them to be found through a weighted tag feature selection. When recommendations are made, the similarities of the tag features among the new and existing users are calculated. The more similar the tags are, the closer the users should be.

The author in Reference [68] proposes a semantic follower recommender system in Twitter that exploits the users’ tweets to build interest profiles. An interest graph is created using specific semantic knowledge graphs containing a variety of topics, which are then mapped to the users according to their semantic relevance to the topics. Using graph theory algorithms, user interest similarity is calculated, which is used during the recommendation process.

A framework for discovering similar accounts in Twitter based only on the “List” feature is proposed in Reference [64]. This functionality allows the users to create their own lists by adding any account they wish. The authors claim that this feature is considered a form of crowdsourcing. The hypothesis of the methodology is that when two accounts are present in the same list, they should be similar or related to each other. Therefore, the proposed measurement relies on the number of lists that a specified account and a potentially similar one are listed together.

Another study on the discovery and suggestion of similar Twitter users is described in Reference [113]. It is based exclusively on the users’ broadcasted messages in terms of Twitter entities (that is, mentions, replies, hashtags, and URLs) these messages contain. Consequently, a higher number of common entities in the messages of different users indicates greater similarity among them.

4.2.2 Topic and Event-oriented Matching. As we have already mentioned, social semantics patterns can be used to identify users’ interests or topics of discussion such as real-life events.

The studies described in References [23, 4, 5], and [95] are specialized in discovering the most influential authors in Twitter on a specific topic. Specifically, in Reference [23], the authors suggest various metrics based on tweets (seed tweet, retweets, and replies), mentions, as well as friendship relationships. In Reference [4], which was also described in Section 4.2.1, these metrics are discovered by considering properties and measures on user-post graphs; while in Reference [5], presented in Section 3.4.1, influencers are regarded as those generating tweets of high quality. A different kind of social influence, a persuasive one, is proposed in Reference [95]. The proposed measurement depends on topical information, the users’ authority, and the characteristics of relationships among individuals. Section 3.2.1 presents more detailed information on the study.

The authors in Reference [92] propose a content-centered model of flow analysis for investigating the Influence Maximization problem on topic-specific influencers. As also described in Section 3.2.1, this analysis is not based on the users’ relationships, but on the content of the transmitted messages. Influencers are discovered by exploiting information flow patterns, their position, as well as the number of flow paths they participate in.
A framework for determining the relevance of Twitter messages for a given topic is introduced in Reference [54]. Two kinds of features are identified: those relevant to the user query and therefore created once; and those not related to this query but are inherent posts and are therefore created as soon as they are altered.

In Section 4.1, we presented an approach that associates tweets with a given event by utilizing their structured information [55]. The application of query expansion techniques and the relationships are derived from the semantified data result in those associations.

Another topic-oriented framework for Twitter is presented in Reference [51]. Its aim is to discover the users’ topics of interest by examining the content and the entities found, which may be mentions or plain text (in OSNs, the mentions are words prefixed with “@”). The Wikipedia knowledge base is leveraged to disambiguate those entities and the topics of interest to be defined (e.g., the term “apple” may refer to the fruit or to the multinational technology company).

The work in Reference [67] was described in Section 4.2.1 and presents an LDA [59] topic profile modeling approach for the discovery of users’ interests. Therefore, a POS tagger extracts the nouns from their tweets, which are then provided to a search engine to retrieve the top related web pages. Using LDA on the content of these web pages discovers final topics.

Topic profiling using the Wikipedia knowledge base is also studied in Reference [52]. Specifically, the topics are discovered based on the posted content, namely, hashtags, of Twitter accounts and their friendship relationships. The celebrities (accounts of popular people) who are followed are the primary indicators of interest that have been derived from their Wikipedia classification. The indicators along with the posted hashtags infer the accounts’ areas of interest.

Similarly, the work in Reference [53] focused on identifying the interests of Twitter users by exploiting the topic categories and sub-categories they have generated. The methodology is based on the hierarchical relationships of Wikipedia content for inferring user interests. According to the authors, those structures can be incorporated into existing systems to provide more personalized content by utilizing broader and higher-level concepts (e.g., the concept “basketball” is more generic that the term “NBA”).

The authors in Reference [62] proposed an ontology-assisted topic modeling technique for determining the topical similarities among Twitter users. The entities found in the posts are mapped to classes of DBpedia ontology and are used to label clusters. Moreover, the topical similarities between individuals on different topics are calculated using ranking techniques, which define the structure of the resulting graphs. Based on the graphs, a quasi-clique community detection algorithm is applied for discovering topic clusters without predefining their target number. The methodology was also presented in Section 4.1.

Another topic-oriented framework, also presented in Section 3.3, which uses the DBpedia knowledge base, is proposed in Reference [71]. Specifically, the context of Twitter is mapped to DBpedia entities and graph-based centrality theory is applied for assigning weights to the entities of the examined messages.

Topic profiling is also exploited in recommendation systems and examples of such studies have been proposed in References [65, 68], and [103], which were also presented in Section 3.4.2. The first two (References [65] and [68]) describe methodologies for the creation of semantic followee recommender systems for Twitter. These studies are based on the classification of the content of tweets and the users that generated them and on semantic knowledge graphs containing a variety of topics being mapped to users, respectively. The recommender described in Reference [103] discovers personal interests by applying a distributed learning supervised algorithm and by taking into consideration explicit social features such as the users’ topic-level influence, topic information, and social relations.
A framework for discovering topic-specific experts in Twitter by employing two distinct metrics is presented in Reference [70]. The first metric measures the users’ global authority on a given topic, while the other metric provides the similarity between the users’ generated tweets and that topic. By leveraging the topical influence and similarity, the users who have the highest-ranking scores are regarded as experts in that domain. This study is described in detail in Section 3.4.1.

In Section 3.2.1, we described a multi-topic influence propagation model based on user relationships, posts, and social actions [75]. The influence score consists of direct and indirect influence, related to different topics. The distribution of users’ topics of interest depends on the content of the disseminated messages. The proposed topic-dependent algorithm is applied and a multi-topical network is created to identify multi-topic influential users.

Another framework exploiting both user interests and social influence is described in Reference [94], where its authors investigate the rebroadcasting behavior of users in OSNs. This study is presented in Section 4.2.1. In the same section, Reference [1] presented a framework using supervised learning for discovering the communities the users belong to and identifying the most influential ones.

### 4.3 Community Detection

Community detection is not only useful for the analysis of OSNs, but also for gaining insights into the structure and characteristics of complex networks. The aim is to group their nodes into potentially overlapping sets sharing common attributes and characteristics. The following studies propose different approaches to detect communities in OSNs:

In contrast to the traditional techniques, the approach in Reference [47] uses not only network topology but also the attributes of the nodes for developing a methodology towards the detection of overlapping communities.

Another algorithm for detecting communities is presented in Reference [49]. It is based on the content of the edges derived from the users’ pair-wise interactions. According to the authors, this algorithm delivers better perception of the communities, because it depicts more effectively the nature of social interactions.

As described in Section 4.1, the methodology in Reference [48] aims at discovering the latent communities in large-scale networks and generates weighted semantic ones. The latter are created using the content derived from users’ comments.

The approach in Reference [50] was presented in Section 3.3 and detects communities using node similarity techniques over a virtual network that is created on top of the original one. New virtual edges are added by considering node similarity under the Jaccard similarity metric. Similarly, Reference [62] proposes a topic modeling technique among Twitter users using the DBpedia ontology for community detection.

Community detection can also be used for identifying influential users in OSNs. Specifically, the work in Reference [76] proposes such a methodology by applying maximum flow algorithms on a weighted representation of the network by considering structural features such as shortest path, betweenness, closeness, and degree centralities. More information can be found in Section 3.2.1.

Another framework, also presented in Section 3.2.1, for discovering top persuasive users in OSNs is described in Reference [99]. The framework is based on machine learning techniques and depends on three factors: influence, entity similarity, and structural equivalence. Each of them is assigned a probability (denoting the likelihood of persuasion) that has been derived by using a machine learning algorithm. Finally, another study that uses communities to identify influential users is presented in Reference [27]. The authors analyzed the social activity and the interconnections of the users inside the communities they belong to. The communities are utilized to develop a methodology for monitoring how influential content spreads from influencers to their followers.
### Table 3. Classification of Referenced Works

| Category                  | NS                | Activities | Context | KB | Semantic Web | Entities |
|---------------------------|-------------------|------------|---------|----|--------------|----------|
|                           |                   | (R)P       | I       | P  | T            | RDF      | O | H | URL |
| Social Modeling           | 48, 62, 65, 69, 71, 112, 114 | 29, 30, 31, 32, 48, 55, 60, 62, 65, 68, 69, 71, 74, 112, 114 | 30, 32, 48, 65, 55, 68, 62, 69, 71, 74, 93, 74 | 29, 30, 31, 32, 55, 62, 65, 68, 69, 71, 74, 93, 114 | 29, 30, 31, 32, 55, 60, 62, 68, 69, 71, 74, 112, 114 | 31, 32, 30, 60, 31, 74, 112, 114 |
|                           |                   | 1, 4, 5, 6, 13, 16, 26, 38, 62, 63, 64, 69, 84, 99 | 1, 4, 5, 6, 20, 26, 29, 30, 32, 38, 55, 56, 58, 62, 67, 68, 69, 72, 84, 91, 94, 105, 113, 115 | 1, 4, 5, 6, 20, 26, 30, 56, 57, 67, 68, 69, 84, 94 | 1, 4, 5, 6, 20, 26, 30, 56, 57, 67, 68, 69, 84, 94 | 30, 32, 32, 55, 113, 30, 32, 62, 69, 62, 68, 94 |
| Topic and Event-oriented  | 1, 4, 5, 23, 51, 52, 53, 54, 55, 62, 65, 64, 69, 70, 71, 75, 79, 92, 95, 103 | 1, 4, 5, 23, 51, 52, 53, 54, 62, 65, 64, 69, 70, 71, 75, 79, 92, 94, 103 | 45, 23, 51, 52, 53, 54, 55, 62, 65, 64, 69, 70, 71, 75, 92, 103 | 51, 52, 53, 54, 55, 62, 65, 68, 71 | 55, 62, 23, 52, 54 | 55, 62, 23, 54, 67 |
| Community Detection       | 27, 47, 48, 49, 50, 62, 76, 99 | 27, 48, 62, 48, 99 | × | 62 | 62 | × | 62 | × | × |

NS refers to network structure, (R)P to (re)-posts, I to interactions, P to profiling and personalization, T to topics, O to ontologies, and H to hashtags.

#### 4.4 Comparison of Related Works

Similarly to Section 3.5, we provide here comparative insights (Table 3) from the above reviewed articles that refer to online social semantics. For each reviewed article, the first column denotes its category according to our classification scheme (see Figure 1). In the rest of the columns, we place the reference number of the articles or mark of "No" (×) in cases where the studies employ or propose metrics and characteristics based on:

(a) Network Structure (NS): includes social follow-up relationships or other types of network linking (e.g., based on mention, reply actions).

(b) Behavioral/Conversational activities: posts (P), re-posts (RP), favorites/likes (FL), mentions (M), replies (R), Contextual analysis: works for building user profiles or providing personalized information (P) and those applied on specific topics (T).

(c) Use of Knowledge Bases (KB): works that use publicly available, open, or crowdsourcing-based resources (e.g., Wikipedia, DBpedia, and so on).

(d) Use of semantics: modeling unstructured data using RDF protocol, use of (existing or new) ontologies (O).

(e) OSN Entities: hashtags (H) and web URLs distributed in social content.
5 ASSESSING INFLUENTIAL CONTENT WITH SEMANTICS

In Sections 3 and 4, we studied the impact of influence and the role of semantics in OSN analysis. Their combination can be used to assess information dynamics as well as for the qualitative assessment of viral user-generated content. In this section, we provide insights into different quantitative indicators that are related to the user-generated content itself (e.g., by traditional textual techniques, sentiment analysis, crowdsourcing enrichment), the user’s behavior and topical interests (e.g., reposting and sharing content, topical similarity, and community metrics), as well as the user’s structural position represented by various centrality measures.

According to the authors in Reference [120], metrics regarding retweets are the best quantitative indicators that show a preference for reading a tweet over another. From the readers’ perspective, a tweet being retweeted several times is more attractive than a tweet with a lot of mentions. The authors of this study concluded that the relationship among users and authors, as well as the authors’ visual and lexical information, namely, their username, display name, and avatar, are the best qualitative indicators that have the strongest effect on the retweeting and reading processes. Retweets as quality indicators are also considered for measuring the appreciation of other users on the generated posts [1]. This attribute—along with metrics related to the users’ activity (represented by the number of hashtags and mentions found at their messages), the characteristics of the communities they belong to, and their structural position (represented by the in-degree centrality)—is highly used to calculate users’ influence.

In Section 4, we described Reference [94] wherein the rebroadcasting behavior of users is studied by examining three aspects, namely, “role of content,” “content-user fit,” and “social influence.” The “content-user fit” measures the relation between the content of the message with respect to the user interests. Influence measures the susceptibility of users for identifying those being affected by the reposting behavior of others. To discover the users’ interests, the LDA methodology is applied on each message. The study concludes that the rebroadcasting of messages does not depend only on its content but also on its relevance with a user.

The methodology described in Reference [8] attempts to identify Twitter users’ asserting positive or negative influence. It depends on three factors, namely, on the users’ structural position, the sentiment, and the textual quality of their posts. The first factor derives as a combination of various centrality measures, while the second one is obtained by applying a sentiment analysis technique on the user-generated content. Finally, the quality of the posted messages can greatly affect users’ influence. Quality is directly dependent on the content and how well-written and comprehensive these messages are.

Authors of Reference [26] performed sentiment analysis to classify social messages into three categories. Specifically, the “MISNIS” framework is presented whose goal is to discover topic-oriented influential Twitter users by applying the PageRank algorithm on a graph representing users’ mentions found in Portuguese tweets. As far as the detection of the topics is concerned, instead of performing naive string matching based on the characters of a hashtag, a fuzzy word similarity algorithm is applied by using all the contents of a message. Consequently, more relevant tweets for a topic are retrieved even if they do not contain the exact hashtags or other user indicated keywords.

The authors in Reference [5] also attempt to identify topic-oriented influential users by regarding them as those generating tweets of high quality. Toward this end, quality is evaluated according to a set of parameters such as the topical similarity, the reposting patterns, and the influence degree on that particular topic of the users who re-shared those messages.

In Reference [84], the authors proposed an approach to identify topic-sensitive opinion leaders in online communities. Their expertise and interests are derived from their tags found in the
description of the products. A computational approach measures the leaders’ influence and ranks them according to not only the link structure of the network, but also their expertise and interests. The influence depends on the topical similarity between reviewers on a specific topic.

The work in Reference [66] describes a service for proposing news articles to Twitter users. It is based not only on terms (e.g., plain words or hashtags), which are mined from both the users’ and their friends’ timeline, but also incorporates additional factors that affect the interest of a user on a tweet, the number of retweets, the influence of its creator, and the quality of the social posts. Indicators of the latter are the numbers of reposts, likes, URLs, hashtags, and its length.

Personalization issues for recommendation purposes are also examined in Reference [103]. The authors incorporate influential features (e.g., users’ topic-level influence, topic information) and their relations among OSN users (e.g., retweeting behavior) for improving recommendation results in thematic categories. In addition, social influence and its propagation can also be used as quality indicators in recommendation systems. In the work described in Reference [14], influence is considered as an attribute that propagates among users in OSNs. The authors’ proposed framework calculates the influence that social relationships have on users’ rating behaviors and incorporates it into recommendation proposals. Other works that employ semantics for identifying user preferences and interests, as well as similarity between users and/or topics, are described in References [32, 62, 68], and [69].

A study that evaluates the quality of articles on Wikipedia by investigating their usage on Twitter is presented in Reference [121]. This is achieved by analyzing three aspects, namely, the language used in tweets in their referenced Wikipedia articles, the Twitter-related content features of such articles (e.g., URLs, hashtags, names of entities), and the correlation between the number of tweets/retweets and edits in their related articles. The authors discovered that the language of the tweets and the referenced Wikipedia articles are not always the same, mainly because of the low quality or the absence of equivalent entries in the user’s native language. Moreover, it was found that the impact of a tweet/retweet about a certain topic is not related to crowdsourcing-based metrics (e.g., edits, discussions) on the same Wikipedia topic.

The authors of Reference [89] describe a framework that exploits influence for evaluating and enhancing communication issues between governmental agencies and citizens (OSNs users). The authors evaluate the quality of the agencies’ responses with respect to the citizens’ requests, analyze their sentiments, and suggest influential users for agencies to obtain a new audience.

The authors of Reference [106] analyze and compare a variety of measurements in OSNs that affect user influence. These were grouped based on various criteria, namely, neighborhood (i.e., number of influencers, personal network exposure), structural diversity (i.e., active community metrics), locality, temporal measures (i.e., retweet time delay), cascade measures (i.e., size, path length), and metadata (i.e., presence of links, mentions, hashtags). Moreover, based on several learning algorithms, the authors propose methods to calculate the users’ retweeting probability.

Influential users are discovered in Reference [13] by applying an eigenvector centrality–based algorithm as an extension of PageRank, which considers both the users’ topical similarities, also identified by applying the LDA methodology, and follow-up relationships. It is suggested that “homophily” (that is, the tendency of people to create social relations with other individuals who have almost identical interests) is responsible for the majority of “Follower-Following” relationships.

Another interesting methodology for measuring user influence based on the quality of the content is proposed in Reference [19], which was further described in Section 3. Initially, users who disseminate qualitative content are considered those with high Follower-to-Followee ratio. Then, by incorporating additional factors deriving from user interactions, such as retweets, replies, user mentions, and keyword similarity, user influence is properly adjusted. According to the classification methodology, in the case of spam detection, the users’ influence is reduced.
In Reference [95] the authors analyzed the so-called persuasion-driven social influence based on topic. Specifically, several influence measurements, in terms of influence propagation for quantifying user-to-user influence probability, incorporate users’ social persuasiveness. Based on the proposed metrics, the framework exploits the topical information, the user’s authority, and the characteristics of relationships (such as direct or indirect connections among users) among individuals.

In Section 3, we also presented the “Influence Metric” measurement [111] and its improved version [112] in the calculation of the importance and influence of Twitter accounts. Therefore, a social function is presented incorporating the activity (or passivity) of a user and the number of followers and followees. In the improved version, a new qualitative factor is incorporated into the measurement based on the established $h$-index measurement. Its aim is to reflect other users’ actions and preferences over the content of the created posts, thus enhancing the influence of quality users.

An important aspect in modeling the content quality in OSNs is the credibility of users, because it can also be directly dependent on the degree of trustworthiness of the information. One suggested solution toward this direction is crowdsourcing neutralized evaluation based on the accuracy, clarity, and timeliness of information since its creation. When user-generated content is created, topic and/or domain labeling enhances its credibility. In addition to the provision of personalized social content, OSNs should also allow their users to define their own custom filtering rules, based on their interests, toward the selection and evaluation of labeled content. The adoption of such an approach should ultimately enable the social community to assess the credibility of those labels through collective intelligence and crowdsourcing processes. A different solution involves the provision of personalized information as well as topic and/or domain labeling according to the users’ interest profiles. The authors of Reference [122] proposed several solutions that when applied with a variety of data mining approaches, such as classic artificial intelligence and real-time ones, can result in an enhanced experience of social networking capable of providing the means for exchanging information of high credibility and quality.

6 CONCLUSIONS

In this article, we have reviewed two major aspects of OSNs, namely, the online social influence and the role of semantics in OSNs (Section 3 and Section 4, respectively), while in Section 5, we discussed how we may combine both aspects toward the qualitative assessment and modeling of user-generated content. To perform a more detailed analysis and to adequately cover all perspectives of the aforementioned aspects, we analyzed the reviewed works according to a proposed hierarchical classification scheme (Figure 1).

All of the related studies regarding influence measurements in OSNs conclude that the number of user followers/followees does not always produce high influence in his/her community, despite affecting it to a certain extent. The most important factors that affect users’ influence can be categorized as:

- **User-oriented**: interaction with other users and similar activities (e.g., creating new messages), relationship details (number of followers, following users, friends, and so on), as well as structural network characteristics and attributes (e.g., position, shortest paths, closeness, eccentricity, centrality, and degree).
- **Content-oriented**: viral content (e.g., hashtags, mentions), thematic categorization.
- **Quality-oriented**: where quality is measured by the user’s social knowledge and the degree of engagement with other users (e.g., the number of retweets/shares, favorites/likes, replies, and so on).
Usually, the users are highly influential mainly on specific topics and less on others. Therefore, we found that two types of influence exist: a topic-specific one, and a global one spreading through the whole network. Several recent studies propose that social influence should be incorporated into recommendation systems to leverage past behavior and latent relationships among users, as well as to improve their performance. In parallel, social semantics are exploited in the analysis of users’ behavior, interests, and preferences to help recommenders suggest informative content, similar users, and other personalized information.

The literature that we have reviewed in this work has confirmed that influence and information flow are two interdependent concepts of OSNs, since they affect one another positively or negatively. Studies on dissemination of information have shown that the largest cascades are mainly generated by influential users who have a significant number of followers. Usually, a large number of those—not so highly influential—followers initiate short diffusion chains that quickly merge into a large single structure. The dynamics of that information flow can be quantified by considering the following social diversities:

- User activity or passivity.
- User influence and susceptibility.
- User relationships in terms of interaction (e.g., mentions, replies) and friendship features.
- Reposting characteristics (e.g., volume, speed, time interval, number of hops).
- Homophily and entity similarity.
- Network attributes, structure, and user topology.
- Content and structure of messages (e.g., topics, presence of URLs or hashtags, formality of language).

In addition, the study of social information spread (diffusion and propagation) is strongly connected with community detection. Communities in OSNs promote certain topics and can be treated as the logical grouping of social actors that share common interests, ideas, or beliefs. There are two possible sources of information that can be used towards their detection: the network structure and the features and attributes of the node-users.

Finally, OSN users often create messages that are characterized by the highly unstructured and informal language with many typographical errors, lack of structure, limited length, and high contextualization. Consequently, microblogging retrieval systems should deal with data and structure sparseness. To overcome those limitations and to contextualize the semantic meaning of microblog content, modern approaches focus on exploiting the existence of social semantics and user-generated content by identifying entities in them. These entities are used as keywords to indicate the topics of the messages, description of real-life events, as well as to reveal behavioral patterns and building interest profiles, thus enabling the interrelation of semantically related terms and the social proximity or similarity between profiles and interests. Often, those entities are linked to knowledge bases (e.g., Google Knowledge Graph, DBpedia) or are represented as concepts extracted from ontologies using Semantic Web vocabularies to transform unstructured data into Linked Data.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments and suggestions that helped us to improve the content, presentation, and organization of this article.

REFERENCES

[1] S. Räbiger and M. Spiliopoulou. 2015. A framework for validating the merit of properties that predict the influence of a twitter user. Exp. Syst. App. 42, 5 (2015), 2824–2834. Retrieved from http://dx.doi.org/10.1016/j.eswa.2014.11.006
[2] I. Anger and C. Kittl. 2011. Measuring influence on Twitter. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW’11), Stefanie Lindstaedt and Michael Granitzer (Eds.), ACM, New York, NY, 1–4. Retrieved from http://dx.doi.org/10.1145/202428.2024326

[3] D. King, D. Ramirez-Cano, F. Greaves, I. Vlaev, S. Beales, and A. Darzi. 2013. Twitter and the health reforms in the English national health service. Health Pol. 110, 2, (2013), 291–297. Retrieved from http://dx.doi.org/10.1016/j.healthpol.2013.02.005

[4] B. Sun and V. TY Ng. 2012. Identifying influential users by their postings in social networks. In Proceedings of the 3rd International Workshop on Modeling Social Media (MISM’12). ACM, New York, NY, 2012, 1–8. Retrieved from http://dx.doi.org/10.1145/2310057.2310059

[5] S. Kong and L. Feng. 2011. A tweet-centric approach for topic-specific author ranking in micro-blog. In Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA’11), J. Tang, I. King, L. Chen, J. Wang (Eds.), Springer, Berlin, 138–151. Retrieved from http://dx.doi.org/10.1007/978-3-642-25853-4_11

[6] L. A. Overbey, C. Paribello, and T. Jackson. Identifying influential Twitter users in the 2011 Egyptian revolution. In Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP’13), Ariel M. Greenberg, William G. Kennedy, and Nathan D. Bos (Eds.). Springer-Verlag, Berlin, 377–385. Retrieved from http://dx.doi.org/10.1007/978-3-642-37210-0_41

[7] V. Lampos, N. Aletras, D. Preotiuc-Pietro, and T. Cohn. 2014. Predicting and characterising user impact on twitter. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL’14). pp. 405–413.

[8] C. A. S. Bigonha, T. N. C. Cardoso, M. M. Moro, M. A. Gonçalves, and V. A. F. Almeida. 2012. Sentiment-based influence detection on Twitter. J. Braz. Comput. Soc. 18, 3 (2012), 169–183. Retrieved from http://dx.doi.org/10.1007/s13173-011-0051-5

[9] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone’s an influencer: Quantifying influence on Twitter. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM’11). ACM, New York, NY, 65–74. Retrieved from https://doi.org/10.1145/1935826.1935845

[10] E. S. Sun, I. Rosenn, C. A. Marlow, and T. M. Lento. 2009. Gesundheit! modeling contagion through Facebook news feed. In Proceedings of the International AAAI Conference on Weblogs and Social Media.

[11] H. Kwak, C. Lee, H. Park, and S. B. Moon. 2010. What is Twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web (WWW’10), ACM, New York, NY, pp. 591–600. Retrieved from http://dx.doi.org/10.1145/1772690.1772751

[12] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM’10).

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM’10), ACM, New York, NY, 261–270. Retrieved from http://dx.doi.org/10.1145/1718487.1718520

[14] T. Yuan, J. Cheng, X. Zhang, Q. Liu, and H. Lu. 2015. How friends affect user behaviors? An exploration of social relation analysis for recommendation. Knowl.-based Syst. 88 (2015), 70–84. Retrieved from http://dx.doi.org/10.1016/j.knosys.2015.08.005

[15] F. Riquelme and P. G. Cantergiani. 2016. Measuring user influence on Twitter: A survey. Inf. Proc. Manag. 52, 5 (2016), 949–975. Retrieved from https://doi.org/10.1016/j.ipm.2016.04.003

[16] N. Liu, L. Li, G. Xu, and Z. Yang. 2014. Identifying domain-dependent influential microblog users: A post-feature based approach. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI’14). 3122–3123.

[17] P.-Y. Huang, H.-Y. Liu, C.-T. Lin, and P.-J. Cheng. 2013. A diversity-dependent measure for discovering influencers based approach. In Proceedings of the 13th Mexican International Conference on Artificial Intelligence (MICAI’14), Human-inspired Computing and Its Applications, A. F. Gelbukh, F. C. Espinoza, and S. N. Galicia-Haro (Eds.). Lecture Notes in Computer Science, Vol. 8536. Springer, 245–256.

[18] L. B. Jabeur, L. Tamine, and M. Boughanem. 2012. Active microbloggers: Identifying influencers, leaders, and discussers in microblogging networks. In Proceedings of the 19th International Symposium on String Processing and
Information Retrieval (SPIRE’12), L. Calderón-Benavides, C. N. González-Caro, E. Chávez, and N. Ziviani (Eds.). Lecture Notes in Computer Science, Vol. 7608. Springer, 111–117.

[22] D. M. Romero, S. Asur, W. Galuba, and B. A. Huberman. Influence and passivity in social media. In Proceedings of the 20th International Conference Companion on World Wide Web (WWW’11), ACM, New York, NY, 113–114. Retrieved from http://dx.doi.org/10.1145/1963192.1963250

[23] A. Pal and S. Counts. Identifying topical authorities in microblogs. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM’11), ACM, New York, NY, 45–54. Retrieved from https://doi.org/10.1145/1935826.1935843

[24] E. Kafeza, A. Kanavos, C. Makris, and P. Vikatos. Predicting information diffusion patterns in Twitter. In Proceedings of the Conference on Artificial Intelligence Applications and Innovations (AIAI’14), IFIP Advances in Information and Communication Technology, L. Iliadis, I. Maglogiannis, H. Papadopoulos (Eds.), 436, Springer, Berlin. Retrieved from https://doi.org/10.1007/978-3-662-44654-6_8

[25] T. Berners-Lee. 2012. 5 Star Open Data. Retrieved from http://5stardata.info/.

[26] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista. 2017. MISNIS: An intelligent platform for Twitter topic mining. Exp. Syst. Appl. 89 (2017), 374–388. Retrieved from http://dx.doi.org/10.1016/j.eswa.2017.08.001

[27] N. Barbieri, F. Bonchi, and G. Manco. 2013. Cascade-based community detection. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM’13), ACM, New York, NY, 33–42. Retrieved from http://dx.doi.org/10.1145/2433396.2433403

[28] J. Mueller and G. Stumme. 2017. Predicting rising follower counts on Twitter using profile information. In Proceedings of the 9th International ACM Web Science Conference (WebSci’17). ACM, New York, NY, 121–130. Retrieved from https://doi.org/10.1145/3091478.3091490

[29] I. Celik, F. Abel, and G.-J. Houben. 2011. Learning semantic relationships between entities in Twitter. In Proceedings of the Conference on Web Engineering (ICWE’11). Lecture Notes in Computer Science, S. Auer, O. Diaz, G.A. Papadopoulos (Eds.) 6757, 167–181, Springer.

[30] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. 2011. Semantic enrichment of Twitter posts for user profile construction on the social web. In Proceedings of the 8th Extended Semantic Web Conference on the Semantic Web: Research and Applications (ESWC’11), Vol. Part II, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis (Eds.). Springer-Verlag, Berlin, 375–389.

[31] J. Shinavier. 2010. Real-time #SemanticWeb in <= 140chars.” In Proceedings of the International Conference on World Wide Web (WWW’10).

[32] L. De Vocht, S. Softic, M. Ebner, and H. Mühlburger. 2011. Semantically driven social data aggregation interfaces for research 2.0. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW’11), S. Lindstaedt, M. Granitzer (Eds.), ACM, New York, NY, 43–52. Retrieved from http://dx.doi.org/10.1007/978-3-642-20161-5_44

[33] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. 2017. Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. Phys. A: Stat. Mech. Applic. (2017), 278–288. Retrieved from http://dx.doi.org/10.1016/j.physa.2016.11.002

[34] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana. 2017. Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. Phys. A: Stat. Mech. Applic. (2017), 278–288. Retrieved from http://dx.doi.org/10.1016/j.physa.2016.11.002

[35] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. 2011. Recommender systems with social regularization. In Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 287–296.

[36] M. Bouguessa, and L. B. Romdhane. 2015. Identifying authorities in online communities. ACM Trans. Intell. Syst. Technol. 6, 3 (2015), Article 30, 23. Retrieved from http://dx.doi.org/10.1145/2700481

[37] J. Yang, and S. Counts. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. In Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM’10).
[41] K. Lerman and R. Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM’10).

[42] J. Yang and J. Leskovec. 2010. Modeling information diffusion in implicit networks. In Proceedings of the IEEE International Conference on Data Mining (ICDM’10). IEEE Computer Society, Washington, DC, 599–608. Retrieved from http://dx.doi.org/10.1109/ICDM.2010.22

[43] M. Kimura, and K. Saito. 2006. Tractable models for information diffusion in social networks. In Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD’06), Knowledge Discovery in Databases. Lecture Notes in Computer Science. J. Färnkranz, T. Scheffer, M. Spiliopoulou (Eds.), 4213. Springer, 259–271. Retrieved from http://dx.doi.org/10.1007/11871637_27

[44] E. Bakshy, I. Rosenn, C. Marlow, and L. A. Adamic. The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web (WWW’12). ACM, New York, NY, 519–528. Retrieved from http://dx.doi.org/10.1145/2187836.2187907

[45] D. M. Romero, B. Meeder, and J. M. Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In Proceedings of the 20th International Conference on World Wide Web (WWW’11). ACM, New York, NY, 695–704. Retrieved from http://dx.doi.org/10.1145/1963405.1963503

[46] K. H. Pollock, J. D. Nichols, C. Brownie, and J. E. Hines. 1990. Statistical inference for capture-recapture experiments. Wild. Monogr. 109 (1990), 3–97. Retrieved from https://doi.org/10.2307/2532321

[47] J. Yang, J. J. McAuley, and J. Leskovec. 2013. Community detection in networks with node attributes. In Proceedings of the IEEE 13th International Conference on Data Mining (ICDM’13). Retrieved from http://dx.doi.org/10.1109/ICDM.2013.167

[48] Z. Y. Xia and B. Zhan. 2012. Community detection based on a semantic network. Knowl.-based Syst. 26 (2012), 30–39. Retrieved from https://doi.org/10.1016/j.knosys.2011.06.014

[49] G. J. Qi, C. C. Aggarwal, and T. S. Huang. 2012. Community detection with edge content in social media networks. In Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE’12). IEEE Computer Society, Washington, DC, 534–545. Retrieved from http://dx.doi.org/10.1109/ICDE.2012.77

[50] K. AlFalahi, Y. Atif, and S. Harous. 2013. Community detection in social networks through similarity virtual networks. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM’13). 1116–1123. Retrieved from https://doi.org/10.1145/2492517.2500261

[51] M. Michelson and S. A. Maclskassy. 2010. Discovering users’ topics of interest on Twitter: A first look. In Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data (AND’10). ACM, New York, NY, 73–80. Retrieved from http://dx.doi.org/10.1145/1871840.1871852

[52] K. H. Lim and A. Datta. 2013. Interest classification of Twitter users using Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration (WikiSym’13). ACM, New York, NY, Article 22. Retrieved from http://dx.doi.org/10.1145/2491055.2491078

[53] P. Kapanipathi, P. Jain, C. Venkatramani, and A. P. Sheth. 2014. User interests identification on Twitter using a hierarchical knowledge base. In Proceedings of the Extended Semantic Web Conference (ESWC’14), The Semantic Web: Trends and Challenges. Lecture Notes in Computer Science. V. Presutti, C. d’Amato, F. Gandon, M. d’Aquin, S. Staab, A. Tordai (Eds.), 8465. Springer, 8465–8478. Retrieved from https://doi.org/10.1007/978-3-319-07443-6_8

[54] K. Tao, F. Abel, C. Hau, and G.-J. Houben. 2012. Twinder: A search engine for Twitter streams. In Proceedings of the 12th International Conference on Web Engineering (ICWE’12). 153–168.

[55] H. S. Packer, S. Samangoeji, J. S. Hare, N. Gibbins, and P. H. Lewis. 2012. Event detection using Twitter and structured semantic query expansion. In Proceedings of the 1st International Workshop on Multimodal Crowd Sensing (Crowd-Sens’12). ACM, New York, NY, 7–14. Retrieved from http://dx.doi.org/10.1145/2390034.2390039

[56] D. Zhou, S. Lawless, and V. Wade. 2012. Improving search via personalized query expansion using social media. Inf. Retr. 15, 3-4 (2012), 218–242. Retrieved from https://doi.org/10.1007/s10791-012-9191-2

[57] M. Reda B., H. Hacid, M. Bouzeghoub, and J. Dai-gremont. 2011. Personalized social query expansion using social bookmarking systems. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’11). 1113–1114. Retrieved from http://dx.doi.org/10.1145/2009916.2010075

[58] M. Efron. Hashtag retrieval in a microblogging environment. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’10). ACM, New York, NY, 787–788. Retrieved from http://dx.doi.org/10.1145/1835449.1835616

[59] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (2003), 993–1022.

[60] M. Hepp. 2010. HyperTwitter: Collaborative knowledge engineering via Twitter messages. In Knowledge Engineering and Management by the Masses. Lecture Notes in Computer Science, P. Cimiano, H. S. Pinto (Eds.), 6317. Springer, 451–461. Retrieved from http://dx.doi.org/10.1007/978-3-642-16438-5_35
7:36 G. Razis et al.

[61] D. Gayo-Avello. 2013. Nepotistic relationships in Twitter and their impact on rank prestige algorithms. *Inf. Proc. Manag.*, 49, 6 (2013), 1250–1280. Retrieved from https://doi.org/10.1016/j.ipm.2013.06.003

[62] K. Slab bek oorn, T. Noro, and T. Tokuda. 2016. Ontology-assisted discovery of hierarchical topic clusters on the social web. *J. Web Eng.*, 15 (2016), 361–396.

[63] S. Huang, J. Zhang, L. Wang, and X.-S. Hua. 2016. Social friend recommendation based on multiple network correlation. *IEEE Trans. Multim.*, 18, 2 (2016), 287–299. http://dx.doi.org/10.1109/TMM.2015.2510333

[64] N. Kamungu lkassam and T. Leelanupab. 2016. Power of crowdsourcing in twitter to find similar/related users. In *Proceedings of the 15th International Joint Conference on Computer Science and Software Engineering (JCSSE’16)*. 1–6. Retrieved from http://dx.doi.org/10.1109/JCSSE.2016.7748852

[65] B. Deb, I. Mukherjee, S. N. Srirama, and E. Vainnikko. 2016. A semantic followee recommender in Twitter using topicmodel and Kalman filter. In *Proceedings of the 12th IEEE International Conference on Control and Automation (ICCA’16)*. 649–656. Retrieved from http://dx.doi.org/10.1109/ICCCA.2016.7505352

[66] A. Nigam, S. Aguinaga, and N. V. Chawla. 2016. Connecting the dots to infer followers’ topical interest on Twitter. In *Proceedings of the 13th International Conference on Advances in Computing, Communications and Informatics (ICACCI’16)*. 1–6. Retrieved from https://doi.org/10.1109/BESC.2016.7804498

[67] D. P. Karidi. 2016. From user graph to topics graph. In *Proceedings of the IEEE 32nd International Conference on Data Engineering Workshops (ICDEW’16)*. 121–123. Retrieved from http://dx.doi.org/10.1109/ICDEW.2016.7495629

[68] C. Besel, J. Schlöterer, and M. Granitzer. 2016. On the quality of semantic interest profiles for online social network consumers. *SIGAPP Appl. Comput. Rev.*, 16, 3 (2016), 5–14. Retrieved from https://doi.org/10.1145/3015297.3015298

[69] W. Wei, G. Cong, C. Miao, F. Zhu, and G. Li. 2016. Learning to find topic experts in Twitter via different relations. *IEEE Trans. Knowl. Data Eng.*, 28, 7 (2016), 1764–1778. Retrieved from http://dx.doi.org/10.1109/TKE.2016.2539166

[70] F. Kalloubi, E. H. Nfaoui, and O. Elbeqqali. 2016. Microblog semantic context retrieval system based on linked open data and graph-based theory. *Exp. Syst. Appl.*, 53, C (2016), 138–148. http://dx.doi.org/10.1016/j.eswa.2016.01.020

[71] G. Poghosyan and G. Ifrim. 2016. Real-time news story detection and tracking with hashtags. In *Proceedings of the 2nd Workshop on Computing News Storylines*. 20–29.

[72] H. Chen, X. Cui, and H. Jin. 2016. Top-k followee recommendation over microblogging systems by exploiting diverse information sources. *Fut. Gen. Comput. Syst.*, 55, C (2016), 534–543. Retrieved from https://doi.org/10.1016/j.future.2014.05.002

[73] Y. Wang, J. Liu, Y. Huang, and X. Feng. 2016. Using hashtag graph-based topic model to connect semantically related words without co-occurrence in microblogs. *IEEE Trans. Knowl. Data Eng.*, 28 (2016), 1919–1933. Retrieved from http://dx.doi.org/10.1109/TKDE.2016.2531661

[74] X. Tang, Q. Miao, S. Yu, and Y. Quan. 2016. A data-based approach to discovering multi-topic influential leaders. *PLOS One* 11 7 (2016), Retrieved from http://dx.doi.org/10.1371/journal.pone.0158855

[75] V. Jaitly, P. Chowriappa, and S. Dua. A framework to identify influencers in signed social networks. In *Proceedings of the 17th Pacific Asia Conference on Information Systems (PACIS’13)*, J. Lee, J. Mao, and J. Y. L. Thong (Eds.). Retrieved from http://aisel.aisnet.org/pacis2013/20

[76] J. Wei, G. Mengdi, W. Xiaoxi, and W. Xianda. 2016. A new evaluation algorithm for the influence of user in social network. *China Commun.*, 13, 2 (2016), 200–206. Retrieved from http://dx.doi.org/10.1016/j.cc.2016.07.002

[77] K. Almgren and J. Lee, J. Mao, and J. Y. L. Thong (Eds.). Retrieved from http://aisel.aisnet.org/pacis2013/20

[78] K. Almgren and J. Lee. 2016. Applying an influence measurement framework to large social network. *J. Netw. Technol.*, 7 (2016), 6–15.

[79] W.-X. Lu, C. Zhou, and J. Wu. 2016. Big social network influence maximization via recursively estimating influence spread. *Knowl.-based Syst.*, 113 (2016), 143–154. Retrieved from https://doi.org/10.1016/j.knosys.2016.09.020

[80] N. Ohsaka, T. Akiba, Y. Yoshida, and K. Kawarabayashi. 2016. Dynamic influence analysis in evolving networks. *Proc. VLDB End.*, 9, 12 (2016), 1077–1088. http://dx.doi.org/10.14778/2994509.2994525

[81] W. Chai, W. Xu, M. Zou, and X. Wen. 2013. ACQR: A novel framework to identify and predict influential users in micro-blogging. In *Proceedings of the 17th Pacific Asia Conference on Information Systems (PACIS’13)*, J. Lee, J. Mao, and J. Y. L. Thong (Eds.). Retrieved from http://aisel.aisnet.org/pacis2013/20

[82] Q. Zhang, J. Wu, H. Yangx, W. Luy, G. Long, and C. Zhang. 2016. Global and local influence-based social recommendation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM’16)*. 1917–1920. Retrieved from http://dx.doi.org/10.1145/2983323.2983873

[83] M. A. Al-garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, and V. Chang. 2016. Identifying the influential spreaders in multilayer interactions of online social networks. *J. Intell. Fuzz Syst.*, 51, 5 (2016), 2721–2735. Retrieved from http://dx.doi.org/10.3233/JIFS-169112

ACM Computing Surveys, Vol. 53, No. 1, Article 7. Publication date: February 2020.
[84] Q. Miao, Y. Meng, and J. Sun. 2016. Identifying the most influential topic-sensitive opinion leaders in online review communities. In Proceedings of the IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA’16), 330–335. Retrieved from http://doi.org/10.1109/ICCCBDA.2016.7529579

[85] Y. Singer. 2016. Influence maximization through adaptive seeding. ACM Sigecom Exch. 15, 1 (2016), 32–59. Retrieved from http://dx.doi.org/10.1145/2994501.2994503

[86] C. Francalanci and A. Hussain. 2017. Influence-based Twitter browsing with navigtweet. Inf. Syst. 64 (2017) 119–131. Retrieved from https://doi.org/10.1016/j.is.2016.07.012

[87] A. Nargundkar and Y. S. Rao. 2016. Influencerank a machine learning approach to measure influence of Twitter users. In Proceedings of the International Conference on Recent Trends in Information Technology (ICRTIT’16). 1–6. Retrieved from http://dx.doi.org/10.1014/2016.7569535

[88] G. Song, Y. Li, X. Chen, X. He, and J. Tang. 2017. Influential node tracking on dynamic social network: An interchange greedy approach. IEEE Trans. Knowl. Data Eng. 29, 2 (2017), 359–372. Retrieved from http://dx.doi.org/10.1109/TDE. 2016.2620141

[89] A. Dennett, S. Nepal, C. Paris, and B. Robinson. 2016. TweetRipple: Understanding your Twitter audience and the impact of your tweets. In Proceedings of the 2nd IEEE International Conference on Collaboration and Internet Computing (CIC’16). 256–265. Retrieved from http://dx.doi.org/10.1109/CIC.2016.043

[90] Y. Yang, N. V. Chawla, R. N. Lichtenwalter, and Y. Dong. 2016. Influence activation model: A new perspective in social influence analysis and social network evolution. Retrieved from https://arxiv.org/abs/1605.08410.

[91] N. Hassan, M. E. El-Sharkawi, and N. El-Tazi. 2016. Measuring user’s susceptibility to influence in Twitter. In Proceedings of the Social Data Analytics and Management Workshop, co-located with VLDB’16.

[92] K. Subbian, C. Aggarwal, and J. Srivastava. 2016. Mining influencers using information flows in social streams. ACM Trans. Knowl. Disc. Data 10, 3 (2016). Retrieved from http://dx.doi.org/10.1145/2815625

[93] G. Razis, I. Anagnostopoulos, and P. Saloun. 2016. Thematic labeling of Twitter accounts using DBpedia properties. In Proceedings of the 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA’16). 106–111. Retrieved from http://dx.doi.org/10.1109/SMA.2016.7753393

[94] Y. Zhang, W. W. Moe, and D. A. Schweidel. 2017. Modeling the role of message content and influencers in social media rebroadcasting. Int. J. Res. Mark. 34, 1 (2017), 100–119. Retrieved from http://dx.doi.org/10.1016/j.ijresmar.2016.07.003

[95] X. Yi, X. Shen, W. Lu, T. S. Chan, and F. L. Chung. 2016. Persuasion driven influence analysis in online social networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN’16). 4451–4456. Retrieved from http://dx.doi.org/10.1109/IJCNN.2016.7727782

[96] H.-J. Hung, D.-N. Yang, and W.-C. Lee. 2016. Routing and scheduling of social influence diffusion in online social networks. In Proceedings of the IEEE 36th International Conference on Distributed Computing Systems (ICDCS’16). 437–446. Retrieved from http://dx.doi.org/10.1109/ICDCS.2016.45

[97] S. Peng, A. Yang, L. Cao, S. Yu, and D. Xie. 2017. Social influence modeling using information theory in mobile social networks. Inf. Sci. 379 (2017), 146–159. http://dx.doi.org/10.1016/j.ins.2016.08.023

[98] M. R. Bouadjenek, H. Hacid, and M. Bouzaglou. 2016. Social networks and information retrieval, how are they converging? A survey, a taxonomy, and an analysis of social information retrieval approaches and platforms. Inf. Syst. 56 (2016), 1–18. Retrieved from http://dx.doi.org/10.1016/j.is.2015.07.008

[99] X. Fang and P. J. Hu. 2016. Top persuader prediction for social networks. (3 May 2016), 2620141.

[100] Y. Yang, Z. Wang, J. Pei, and E. Chen. 2017. Tracking influential nodes in dynamic networks. IEEE Trans. Knowl. Data Eng. 29, 11 (2017), 2615–2628. Retrieved from http://doi.org/10.1109/TDE.2017.2734667

[101] J. Veijalainen, A. Semenov, and M. Reinikainen. 2015. User influence and follower metrics in a large Twitter dataset. In Proceedings of the 11th International Conference on Web Information Systems and Technologies (WEBIST’15). 487–497. Retrieved from http://dx.doi.org/10.5220/0005410004870497

[102] C. C. Chen, S.-Y. Shih, and M. Lee. 2016. Who should you follow? combining learning to rank with social influence for informative friend recommendation. Dec. Supp. Syst. 90 (2016), 33–45. Retrieved from http://dx.doi.org/10.1016/ j.dss.2016.06.017

[103] D. Li, Z. Luo, Y. Ding, J. Tang, G.-Z. Sun, X. Dai, J. Du, J. Zhang, and S. Kong. 2017. User-level microblogging recommendation incorporating social influence. J. Assoc. Inf. Sci. Technol. 68, 3 (2017), 553–568. Retrieved from https://doi.org/10.1002/asi.23681

[104] K. Zhuang, H. Shen, and H. Zhang. 2017. User spread influence measurement in microblog. Multim. Tools Appl. 76, 3 (2017), 3169–3185. Retrieved from http://dx.doi.org/10.1007/s11042-016-3818-z

[105] I. Anagnostopoulos, G. Razis, P. Mylonas, and C. N. Anagnostopoulos. 2013. Query expansion with a little help from Twitter. Engineering Applications of Neural Networks. Communications in Computer and Information Science 384, 20–29. Retrieved from http://dx.doi.org/10.1007/978-3-642-41016-1_3
