Identifying Top-*k* Nodes in Social Networks: A Survey

RANRAN BIAN, University of Auckland and Pingar
YUN SING KOH and GILLIAN DOBBIE, University of Auckland
ANNA DIVOLI, Pingar

Top-*k* nodes are the important actors for a subjectively determined topic in a social network. To some extent, a topic is taken as a ranking criteria for identifying top-*k* nodes. Within a viral marketing network, subjectively selected topics can include the following: Who can promote a new product to the largest number of people, and who are the highest spending customers? Based on these questions, there has been a growing interest in top-*k* nodes research to effectively identify key players. In this article, we review and classify existing literature on top-*k* nodes identification into two major categories: top-*k* influential nodes and top-*k* significant nodes. We survey both theoretical and applied work in the field and describe promising research directions based on our review. This research area has proven to be beneficial for data analysis on online social networks as well as practical applications on real-life networks.

CCS Concepts: • **Information systems** → **Data mining**; *Social networking sites*; • **General and reference** → *Surveys and overviews*;

Additional Key Words and Phrases: Top-*k* nodes identification, social network graphs

ACM Reference format:
Ranran Bian, Yun Sing Koh, Gillian Dobbie, and Anna Divoli. 2019. Identifying Top-*k* Nodes in Social Networks: A Survey. *ACM Comput. Surv.* 52, 1, Article 22 (February 2019), 33 pages.
https://doi.org/10.1145/3301286

1 INTRODUCTION

In this fast-paced digital age, people around the world are now more connected with others than ever before. Of the many communication and collaboration channels, social networks have become very popular among different communities. The general public frequently use connection social networks, such as Twitter and Facebook, to express opinions on trending topics. Marketers construct informational networks to observe consumer buying behaviors. Research collaborations are recorded in academic networks such as Digital Bibliography & Library Project (DBLP), which is a bibliographic reference on major computer science publications.

A social network can be modeled as a graph that consists of nodes and edges. Each node represents an actor, while each edge shows a connecting relationship between two actors. Intuitively, the representation of nodes and edges varies accordingly to interested actors and relationships in

This work is supported by Callaghan Innovation, under an R&D Student Fellowship Grant, contract number: PTERN1502.
Authors' addresses: R. Bian, Y. S. Koh, and G. Dobbie, School of Computer Science, University of Auckland; emails: rbia002@aucklanduni.ac.nz, ykoh@cs.auckland.ac.nz, g.dobbie@auckland.ac.nz; R. Bian and A. Divoli, Pingar; email: anna.divoli@pingar.com.

0360-0300/2019/02-ART22 $15.00
https://doi.org/10.1145/3301286

the network. For example, in Twitter, each user can be represented as a node, while an edge is formed when there is a "following" relationship between users X and Y. Furthermore, each tweet (a posting message made on Twitter) can also be a node. An edge exists between tweet A and B, if B is a retweet of A's message (i.e a tweet message forwarded by someone else). In viral marketing networks, each consumer can be a node, while an edge is formed if a consumer successfully persuades another person to purchase a new product. In DBLP, individual nodes represent different academics, while edges indicate that the connected researchers have published a paper together.

An important factor for an organization's success is the ability to advertise their products to potential customers. To do this efficiently, the organization may want to know who are the high-priority customers that should be targeted. Top-k nodes identification in social networks is an extension of this question. Identifying key nodes can be useful for increasing product adoption rates in advertising or searching for domain experts. Research in this field has received a lot of interest due to three benefits: (1) It brings order to search results so that the contributing nodes can be ranked by their significance, authority, and/or influence; (2) it can be utilized to increase the efficiency of marketing and advertisement campaigns; and (3) its ability to improve the utility of gathered information.

Nowadays, with the massive amounts of data available, it is not always practical to analyze everything in a dataset due to resource constraints. Instead, sometimes, it is more effective to explore the most significant or influential actors (the top-k nodes) in a network. In this survey, we review and classify the current work on identifying top-k nodes in social networks. As seen from existing literature, this research area has applications to different domains, such as advertisements in viral marketing (Kempe et al. 2003), information circulation (Gruhl et al. 2004), and domain experts search (Zhu et al. 2015; Subbian et al. 2016b).

Since the first algorithmic formulation of the top-k nodes identification problem in 2001 (Domingos and Richardson 2001), there have been a large number of publications on various kinds of algorithms and applications in this area. With over a decade of research, it is time to perform an overview of this field and examine what more can be done in this research area. First, we provide general descriptions of concepts that are used extensively in the rest of this article. Please note that more detailed descriptions of these concepts are provided later in Sections 2.6, 2.7, and 2.8.

Topic. A topic is the area of user interest and can be used as a ranking criteria for identifying top-k nodes.
Top-k nodes (T). Top-k nodes are the k important actors for a subjectively determined topic in a social network, where k is a user-specified integer.
Top-k Influential nodes (I). Top-k influential nodes are the k actors that are capable of generating maximum influence and widest information spread to their connected nodes in a social network, where k is a user-specified integer. Top-k influential nodes will sometimes be shortened to influential nodes in this article.
Top-k Significant nodes (S). Given a particular topic, top-k significant nodes are the k actors whose intrinsic attributes are most relevant to the given topic in a social network, where k is a user-specified integer. Top-k significant nodes will sometimes be shortened to significant nodes in this article.

1.1 Related Surveys

In comparison with other related surveys (Lappas et al. 2011; Sun and Tang 2011; Guille et al. 2013; Probst et al. 2013; Riquelme and González-Cantergiani 2016), our work has three distinct differences:

(1) We extend the focus from either influential or significant nodes to a broader concept—the top-k nodes. As a result, our survey includes not only research on influence maximization but also studies on identifying the significant nodes in social networks.
(2) As opposed to some existing surveys, e.g., Riquelme and González-Cantergiani (2016), which focuses on only one type of social network (Twitter in this case), our work reviews related work in a wide range of social networks, such as Amazon, DBLP, and Wiki.
(3) More comprehensive and recent literature are reviewed and an extensive coverage of the subject is provided, i.e., applied work in different application domains, and top-k influential nodes identification in dynamic social networks, which were not addressed in previous surveys.

1.2 Contributions

Our main contributions are summarized as follows. First, we define an extended concept, top-k nodes, to provide more comprehensive coverage of the field. Second, both theoretical and applied work of identifying the top-k nodes are reviewed and classified in a novel way. Finally, some promising research directions are discussed based on our survey.

1.3 Survey Organization

In this article, we conduct a high-level overview of the top-k nodes identification algorithms, methodologies, and applications. With a rich body of literature in this area, we organize our discussions into the following four topics: (1) top-k influential nodes identification, (2) top-k significant nodes identification, (3) applications of identifying top-k nodes in various networks, and (4) research directions. The remainder of the article is also organized in the corresponding four sections (Sections 3 to 6). In addition, we provide some preliminary concepts in Section 2 and conclude the survey study in Section 7.

2 PRELIMINARIES

We introduce the social network–related preliminary concepts, followed by traditional node centrality measures and then different influence diffusion models for identifying top-k influential nodes. Finally, formal definitions of influence maximization, influential nodes, and significant nodes are provided. In addition, relationships among top-k nodes, influential nodes, and significant nodes are discussed.

2.1 Social Network

In our survey, social networks contain two concepts. First, it is a network of social interactions and personal relationships. This kind of social network existed long before the likes of Facebook and Twitter and has mainly been used to describe entity relationships. Early research on this type of network was conducted by social scientists. It has become an important research area in computer science in recent years. Second, social networks are profile sites or virtual communities where users can share interests and ideas or discover friends through posting comments, messages, and images.

2.2 Static Network versus Dynamic Network

As dynamic networks evolve, new nodes and edges will be introduced. An example of this is shown in telecommunication networks, where transient links are added between two participant nodes based on texts or calls between them. These dynamic networks with transient interactions can be represented as graph streams; however, due to high computational complexity and disk storage requirements, these graph streams typically require real-time methods. In contrast, static networks

have fixed topology, structure, and information, which can be treated as a snapshot of the dynamic network at a distinct time t. Therefore, offline computational methodologies can be applied on these static networks (Aggarwal and Subbian 2014).

2.3 Social Network Graph

A social network can be modeled as a graph $G = \{V, E\}$, where V is a set of nodes $\{v_1, v_2, \ldots, v_i\}$ and E is a set of edges $\{e_1, e_2, \ldots, e_j\}$. In the network graph, V represents actors and E represents social interactions and relationships between the actors. For example, when an edge e_j exists between nodes v_l and v_p, the two corresponding actors are considered connected/related to each other. Neighbour nodes of v_i is represented as $N = \{n_1, n_2, \ldots, n_m\}$, which refers to all nodes that are directly connected to v_i. Social networks will be described occasionally as *network graphs* in this article.

2.4 Node Centrality Measures

In existing literature, a number of node-based centrality measures have been proposed and defined for evaluating the importance of nodes in selected network graphs. As this survey focuses on identifying top-k nodes, we introduce two of the measures that are relevant to this survey.

2.4.1 Degree Centrality. Historically first and conceptually the simplest node measure is degree centrality, where a "degree of a node is the number of edges the given node has" (Wasserman and Faust 1994). Degree centrality c_{v_i} of node v_i is defined to be the degree of the node (Equation (1)):

$$c_{v_i} = deg(v_i). \tag{1}$$

In the academic collaboration network (Figure 1), the degree centrality of Diana node is 4. From the perspective of traditional social network analysis, a node with a larger degree centrality is normally considered as more important.

2.4.2 Closeness Centrality. "Closeness centrality is defined as the average length of shortest paths between node v_i and all reachable nodes of v_i in the network graph" (Wasserman and Faust 1994). The closeness centrality is represented in Equation (2):

$$Closeness(v_i) = \frac{\sum_{v_j \in R \setminus v_i} dist(v_i, v_j)}{|R| - 1}. \tag{2}$$

Given that R is the group of nodes that can be traversed from node v_i and that R contains the node v_j, the function $dist$ returns the length of shortest path between node v_i and node v_j (Tseng and Chen 2012). The shortest path indicates the minimum number of edges connecting two nodes.

In the academic collaboration network (Figure 1), the shortest path between Diana and Lucy is 1 rather than 2 (Diana-Celina-Lucy). Following Equation (2), we can calculate that the closeness centrality of Diana node is 1. From the perspective of traditional social network analysis, a node with a smaller closeness centrality is normally considered more important.

2.5 Influence Diffusion Models

Since influential nodes are an essential component of the top-k nodes, we introduce two influence diffusion models that are widely used for exploring influence of nodes. When modeling the "spread of an innovation through a social network graph G, each node v_i can have either an *active* (an adopter of the innovation) or *inactive* (not yet an adopter) status" (Kempe et al. 2003). Two ground settings for influence diffusion models discussed in this article are

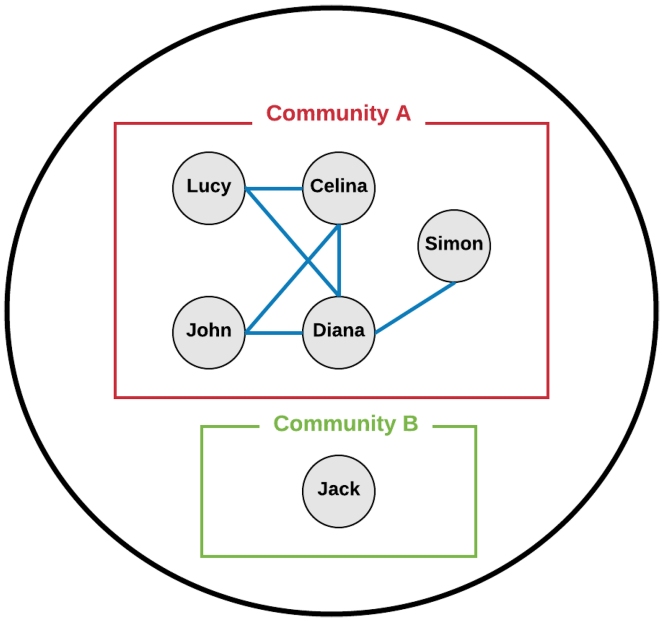

Fig. 1. Academic collaboration network.

(1) v_i has monotonically increased tendency to become active as more of its neighbour nodes n_m (previously defined in Section 2.3) become active.
(2) v_i can switch in only one direction: from being inactive to being active.

The influence diffusion model progresses as follows for an initially inactive node v_i: With the unfolding of this process, an increasing number of v_i's neighbour nodes n_m become active. At certain points, this might influence v_i to become active and v_i's decision might subsequently trigger a status change of the remaining inactive n_m. The process runs until no further triggering can happen (Kempe et al. 2003).

The Linear Threshold (LT) model (Granovetter 1987) and Independent Cascade (IC) model (Goldenberg et al. 2001) are two of the most basic and widely studied influence diffusion models. The LT is receiver-centric where the core idea is to find out whether a node can be activated given a fraction of active neighbour nodes, whereas the IC model is sender-centric and focuses on individual node's activation attempts on its inactive neighbour nodes. Therefore, the two models reflect different views on the influence diffusion process.

2.5.1 Linear Threshold Model. In the model, a node v_i is influenced by each neighbouring node n_m according to a weight w_{v_i,n_m}. For any inactive node, v_i, we select a *threshold* θ_{v_i} between the interval [0,1]. The value θ_{v_i} represents the weighted fraction of v_i's neighbour nodes n_m that must be active for v_i to become active, and this represents the tendencies of nodes being influenced by neighbouring active nodes, becoming activators themselves (Equation (3)),

$$\sum_{n_m \; active \; neighbour \; of \; v_i} w_{v_i,n_m} \geq \theta_{v_i}. \tag{3}$$

Given a random set of active nodes, A, and random thresholds, θ_{v_i}, for each inactive node, the diffusion process will begin to iterate in deterministic and discrete steps. For each iteration, t, all nodes that were active in steps $t-1$ or lower will remain active and if the total neighbouring

weight, n_m, for a target node v_i exceeds its threshold θ_{v_i}, v_i will be activated, and this process will continue until no more activations are possible. (Kempe et al. 2003; Granovetter 1987).

2.5.2 Independent Cascade Model. Also starting with an initial set of active nodes, A, the independent cascade model performs diffusion based on the following randomized rule. For a given iteration, t, if a target node, v_i, was activated in step $t-1$, then it will have the chance to activate each inactive neighbouring node, n_m, based on a history-independent probability p_{v_i,n_m}. If v_i succeeds in activating n_m, then n_m will become and remain active in steps $t+1$ onwards, repeating the same process for its neighbouring nodes and v_i will no longer be able to influence the network. (Kempe et al. 2003; Goldenberg et al. 2001).

2.6 Influence Maximization Definition

Based on the influence diffusion models described in the previous section, the influence maximization problem is defined as follows.

Definition 2.1 (Influence Maximization). "Influence maximization is an optimization problem, which requires selection of a good initial set of active nodes A. The influence of A is measured by the number of active nodes at the end of an influence diffusion process. The influence maximization problem aims at finding a k-node set of maximum influence" (Kempe et al. 2003). The meaning of active nodes varies with interested topics. Possible meanings can include but are not limited to adoption of a new research area, purchase of a recommended product, or receiving a new piece of information.

2.7 Influential Nodes Versus Significant Nodes

Given the influence maximization definition in Section 2.6 and the definition of Topic in Section 1, we formally define top-k influential nodes and top-k significant nodes.

Definition 2.2 (Top-k Influential Nodes). Let $G = \{V, E\}$ denote a social network graph, where $V = \{v_1, v_2, \ldots, v_i\}$ is a set of nodes and $E = \{e_1, e_2, \ldots, e_j\}$ is a set of edges in G. We define top-k influential nodes as a user-specified number of v_i that leads to maximum influence spreading to their connected nodes in G.

Definition 2.3 (Top-k Significant Nodes). Let $G = \{V, E\}$ denote a social network graph, where $V = \{v_1, v_2, \ldots, v_i\}$ is a set of nodes and $E = \{e_1, e_2, \ldots, e_j\}$ is a set of edges in G. We define top-k significant nodes as a user-specified number of v_i, whose intrinsic attributes are most relevant to a specified topic in G.

2.8 Relationships among Top-k Nodes, Top-k Influential Nodes, and Top-k Significant Nodes

A given social network can contain a wide array of topics, such as "who are the influential users?" and "who are the subject experts?" From the perspective of viewing the network under the context of a single topic as opposed to the set of all possible topics, we illustrate the relationships among top-k nodes (T), top-k influential nodes (I), and top-k significant nodes (S). When we consider only a single topic for a given social network, if the topic is more relevant to influence maximization of k number of nodes to their connected nodes, then the set of top-k nodes will be the same for both top-k influential and top-k significant nodes, we define this as the Top-k Influential Nodes Scenario. However, if we view the topic based on intrinsic attributes (such as authority or representativeness) of k number of nodes rather than their influence on neighbouring nodes, then the top-k nodes will be the same as the set of top-k significant nodes, which we define as the Top-k Significant Nodes

Scenario. *From the perspective of all possible topics in a social network, top-k influential nodes is a subset of top-k significant nodes, denoted as* $I \subseteq S$.

In a viral marketing network, each existing or potential customer is a node (actor), while an edge is formed between two customers if they have social connections. Topics of interest such as "Who can promote a new product to the largest number of people?" are examples of influence maximization on groups of customers and can be considered as examples of Top-k Influential Nodes Scenario, whereas topics that are focused on non-influential properties of the nodes such as "Who are the highest spending customers?" are examples of the Top-k Significant Nodes Scenario.

To further illustrate differences between the Top-k Influential Nodes Scenario and the Top-k Significant Nodes Scenario, we employ an academic collaboration network graph (Figure 1). In the graph, each node represents an individual academic. An edge between two nodes exists if there is a research collaboration between the two corresponding academics. Red and green rectangles indicate two separate research communities A and B in the network. The table in Figure 1 shows information derived from the academic collaboration network: The number of collaborators each academic has and the research community that a particular academic belongs to. Based on Figure 1, we provide examples of the Top-2 Influential Nodes Scenario and the Top-2 Significant Nodes Scenario below:

> *Top-2 Influential Nodes Scenario.* Topic of interest: Who are the two most influential academics in the network? By having the largest numbers of collaborators, Diana and Celina will have a higher chance of promoting new research ideas to co-workers than others. In this case, Diana and Celina are the top-2 nodes, the top-2 influential nodes, and the top-2 significant nodes.
> *Top-2 Significant Nodes Scenario.* Topic of interest: Who are the two academics that represent all existing research communities in the network? In this case, Jack and Diana are the top-2 nodes as well as the top-2 significant nodes. Despite being an isolated researcher, Jack will still be selected as one of the top (significant) nodes due to his unique representativeness of the research community B. While the most-connected node, Diana, in research community A is the top (significant) node in her community.

In the academic collaboration network (Figure 1), when the topic of interest is more relevant to maximum influence (promoting new research ideas to collaborative academics), Diana and Celina are the top-2 influential and significant nodes. However, when the topic of interest is shifted to find the total number of communities in the network, the concept of influential nodes is no longer applicable, and Jack and Diana become the top-2 significant nodes in the network.

3 TOP-K INFLUENTIAL NODES IDENTIFICATION

As we progress further into the digital age, there has been a steady increase in the importance and influence of social media and networks. The ability for users to share their thoughts, statuses, and activities in these social mediums has established a new level of connectivity between different groups and niches of people, such that a single user can influence millions of their followers. Hence, it would be of interest for companies to select these influential individuals as their initial user group to produce the greatest level of coverage when advertising their products.

We start our survey with approaches for identifying influential nodes. Different kinds of approaches for identifying influential nodes are reviewed in Sections 3.1 to 3.4. In addition, Section 3.5 discusses a relatively new research area: Identifying Top-k influential nodes in dynamic social networks. Table 1 provides an overview of research work covered in this section.

Table 1. List of Some Approaches for Top-k Influential Nodes Identification

Category of Approach	Related Research
Greedy	(Kempe et al. 2003) (Kimura et al. 2007) (Leskovec et al. 2007) (Chen et al. 2009)
Centrality Measure	(Ilyas and Radha 2011) (Chen et al. 2012) (Kim and Yoneki 2012) (Wei et al. 2013)
Topic	(Tang et al. 2009) (Liu et al. 2010) (Barbieri et al. 2013) (Aslay et al. 2014)
Network Content (or Topology)	(Wang et al. 2010) (Subbian et al. 2014) (Zhou et al. 2014) (Chen and He 2015)
Dynamic Network	(Subbian et al. 2013) (Zhuang et al. 2013) (Subbian et al. 2016a) (Subbian et al. 2016b)

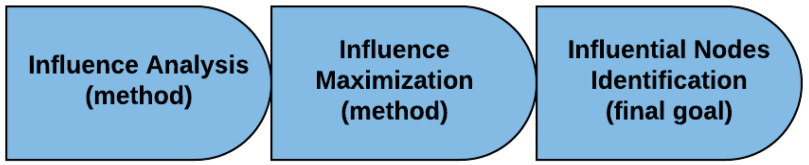

Fig. 2. Core idea of greedy algorithms for identifying influential nodes.

3.1 Greedy-Based Approaches and Improvements

One important category of approaches for identifying influential nodes is based on greedy algorithms. "The core idea of the algorithm is to calculate the influence of each individual, and take turns to choose the node maximizing the marginal influence value until k number of influential nodes are selected" (Wang et al. 2010). Figure 2 illustrates this core idea in a procedural way: The process of influential nodes identification begins with analysing the nodes' influence and then it iteratively maximizes the influence of an initial set of nodes. The final goal is to obtain a group of initial nodes with maximum influence, which are the influential nodes. To describe the process of information propagation, the majority of work reviewed in this section use influence diffusion models (Kempe et al. 2003; Leskovec et al. 2007; Chen et al. 2010), such as the Linear Threshold (Granovetter 1987) or the Independent Cascade model (Goldenberg et al. 2001) (previously described in Section 2.5).

In 2001, Domingos and Richardson (2001) were the first to explore information and influence propagation as a computational problem. With their proposed probabilistic solution, they designed viral marketing strategies and analyzed diffusion processes using a data mining approach. Two years later, Kempe, Kleinberg, and Tardos (2003) categorized the problem of finding the most influential individuals as an optimization problem. Additionally, Kempe et al. provided the first provable approximation guarantees for the influence maximization problem, where the influence propagation process is modeled to reflect the effects of "word of mouth" for the "promotion of new products" (Kempe et al. 2003). When identifying top-k influential nodes, we are interested in finding the most influential "mouths" that can generate the largest possible influence cascade. The problem statement contains two sub-problems: "The most influential mouths" describes the problem of finding the top-k influential nodes, whereas "generate largest possible influence cascade" corresponds to the influence analysis and maximization problem. These two sub-problems are both incorporated into Section 3.

Kempe et al. define influence maximization as "a problem of identifying a small set of seed nodes in a social network that maximizes the spread of influence under certain influence diffusion model" (Kempe et al. 2003). Although there are various models for influence propagation in network graphs, the authors (Kempe et al. 2003) choose the Linear Threshold (Granovetter 1987) and Independent Cascade (Goldenberg et al. 2001) models (previously described in Section 2.5) in their research. On the basis of submodular functions, the proposed analysis framework demonstrated that the natural hill-climbing greedy algorithm that achieves "a solution that is provably within 63% of optimal" (Kempe et al. 2003). In conjunction to the provable guarantees, experiments were also conducted to show that their approximation algorithm significantly out-performs node-selection heuristics based on degree and closeness centrality (previously described in Section 2.4). One of their major findings observed that focusing only on clustered centrality-based nodes may not generate maximum influence. On the contrary, targeting nodes with most possible additional marginal gain results in superior performance over the two centrality-based heuristics.

Following their research on influence maximization through a social network (2003), Kempe et al. (2005) define a natural and general model for the influence propagation process, termed the decreasing cascade model. The model begins with a set of "active" nodes that spreads influence in a cascading way depending on a probabilistic rule. Their problem statement focuses on choosing a target set of individuals for initial activation so that the cascade process is able to result in a largest possible active set. Kempe et al. provide provable approximation guarantees for selecting a target set of size k using a simple greedy algorithm in the proposed decreasing cascade model. A research direction described is to investigate which are the most general influence diffusion models and what provable performance guarantees can be achieved for those models. Kimura and Saito (2006) proposed two natural special models (SPM, SP1M) of the IC model (Goldenberg et al. 2001) such that they can effectively compute the number of influenced nodes given an initial set of influential nodes. Similarly to Kempe et al. (2005), the authors also provided provable performance guarantees for the simple greedy algorithm in the proposed models. Their experiments with large-scale social networks demonstrated that when using the proposed two models for identifying influential nodes: (1) They can provide close approximation to the IC model if the propagation probabilities between links are small, and (2) they can be scalable and much faster than the IC model.

Leskovec et al. (2007) developed a "lazy-forward" optimization method for choosing a group of nodes to detect out-break, i.e., the spread of virus, as early as possible. Their proposed algorithm, Cost-Effective Lazy Forward (CELF), greatly reduces the number of calculations on the node influence propagation, which gains up to 700 times more efficiency over the simple greedy algorithm. Chen et al. (2009) attempted to further improve the efficiency of identifying influential nodes from the following two directions: (1) Develop new algorithms (NewGreedy and MixedGreedy) on top of the simple greedy algorithm. However, this resulted in unremarkable efficiency improvements, (2) designed a new heuristic algorithm: DegreeDiscount, which is more efficient and scalable than the greedy strategy. However, the DegreeDiscount heuristic is a derivation of the *Uniform Independent Cascade model* where propagation probabilities of all edges are identical.

DegreeDiscount's limitation is later addressed in a study by Chen et al. (2010) with a new heuristic that accommodates the general IC model (Goldenberg et al. 2001). The authors underline two critical weaknesses with existing influential node discovery techniques: (1) Existing algorithms have poor scalability, thus will perform poorly with large-sized graphs, i.e., Kempe et al. (2003) and Leskovec et al. (2007); (2) these algorithms have either low scalability or have un-influential initial nodes and therefore have low influence spread.

To resolve these two issues, Chen et al. adopted a simple tuneable parameter for users to control "the balance between efficiency (in terms of running time) and effectiveness (in terms of influence spread) of the algorithm" (Chen et al. 2010). Their solution results in a more efficient greedy

Table 2. Comparisons of Some Greedy-based Approaches for Identifying Influential Nodes

Approach	Heuristic approach	Performance guarantees provided	Scalable
Simple greedy (Kempe et al. 2003)	No	Yes	No
Improved greedy (Kimura et al. 2007)	Yes	No	Yes
CELF (Leskovec et al. 2007)	No	Yes	No
NewGreedy (Chen et al. 2009)	No	Yes	No
MixedGreedy (Chen et al. 2009)	No	Yes	No
DegreeDiscount (Chen et al. 2009)	Yes	No	Yes

algorithm when selecting nodes in each iteration. When comparing the performance of their algorithm with existing techniques such as simple greedy (Kempe et al. 2003), their results were shown to be (1) more scalable with linear growth in running time beyond million-sized graphs, while others are exponential, and (2) faster execution time and spread of influence for both real-world and synthetic datasets.

Chen et al. (2010) closed an open question left by Kempe et al. (2003) by proving that influence computation in the LT model (Granovetter 1987) is NP-hard. In addition, the authors showed that "computing influence in directed acyclic graphs (DAGs) can be done in linear time" (Chen et al. 2010). Based on the fast computation in DAGs, Chen, Yuan, and Zhang proposed a "scalable influence maximization algorithm tailored for the LT model" (Chen et al. 2010).

Narayanam and Narahari (2011) developed a novel way of discovering influential nodes in social networks with the Shapley value concept (Shapley 1950), which is commonly used in cooperative game theory. The ShaPley value-based Influential Nodes (SPIN) maps the information diffusion process to "the formation of coalitions in a cooperative game" (Narayanam and Narahari. 2011). For computing the network values of each node, Shapley values are used to determine the marginal contributions each node makes to the influence propagation process. The experiments with both synthetic and real-world datasets showed that SPIN works comparatively well as the simple greedy strategy (Kempe et al. 2003) for maximizing influence yet by consuming much less running time.

Liu et al. (2014) provided a bounded linear approach for identifying influential nodes in large-scale social networks. A quantitative metric, termed Group-PageRank, which can be computed in near constant time, is also proposed to address the scalability issue of the influence maximization problem. The authors developed two "lazy-forward" greedy algorithms based on the bounded linear approach and the Group-PageRank, respectively. The evaluation results showed that the two greedy algorithms can effectively and efficiently identify influential nodes with both of them being scalable for large-scale social networks.

We compare some greedy-based approaches for identifying influential nodes in Table 2. The simple greedy algorithm (Kempe et al. 2003) provides the first-ever provable performance guarantees for the influence maximization problem. However, it is computationally expensive and not applicable for large-scale social networks, and later approaches address the efficiency and scalability limitations. The improved greedy algorithm proposed by Kimura et al. (2007) achieves a large reduction in computational cost by removing edges that do not contribute to information diffusion and does the propagation on a subgraph. The CELF algorithm optimizes the simple greedy algorithm using the "submodularity property of the influence maximization objective" (Leskovec et al. 2007) to reduce the number of calculations on the node influence propagation. Chen et al. (2009) proposed three different algorithms: NewGreedy, MixedGreedy, and DegreeDiscount. The key idea behind NewGreedy is to eliminate the edges that will not contribute to influence propagation from the original graph to get a smaller graph and performs the influence diffusion on

the smaller graph. The first iteration of MixedGreedy employs the NewGreedy algorithm, while the remaining iterations use the CELF algorithm. DegreeDiscount assumes that influence spread of a node is increased with the increase in degree centrality of the node. The authors suggested that DegreeDiscount should be adopted when efficiency is essential, while MixedGreedy can be used for identifying influential nodes when maximum influence spread is a priority (Wang et al. 2010). An interesting fact shown in Table 2 is that scalable greedy-based approaches lack provable performance guarantees. This interesting fact is also described as "algorithms applies various heuristics without provable approximation guarantee" (Song et al. 2017).

3.2 Centrality Measure-Based Approaches

In terms of influence maximization, some research in Section 3.1 experimentally demonstrated that their greedy-based approaches out-perform traditional centrality measures, such as the degree and closeness centrality (Wasserman and Faust 1994) (previously described in Section 2.4). However, there are existing approaches for identifying influential nodes using centrality measures. In this section, we review new centrality measures that have been shown to be more effective than traditional measures for maximizing influence.

Ilyas and Radha define centrality as *"a measure to assess the criticality of a node's position"* (Ilyas and Radha 2011). The ability to find influential nodes is shown in many existing methodologies, such as Eigenvalue Centrality (EVC) (Bonacich 1987). However, these techniques often target a single set of influential nodes and cluster them within one neighbourhood of nodes. This does not reflect the properties of social network graphs, where there could be multiple influential neighbourhoods. The approach proposed by Ilyas and Radha (2011) uses Principal Component Centrality (PCC) to form social hubs, groups of nodes in a network, whose centrality scores are higher than their neighbours. While EVC forms a single cluster of nodes with the highest centrality scores, PCC considers additional factors, such as the weighting of eigenvectors, when computing centrality of nodes. The authors applied both EVC and PCC to real-world datasets (i.e., Facebook and Orkut) and found a significant increase in the number of influential neighbourhoods discovered.

Chen et al. (2012) underlined the issues with using traditional centrality measures for influential node identification. Degree centrality methods are simple but irrelevant, while closeness centrality methodologies are inapplicable for large-scale networks due to the computational complexity and running time. The authors then proposed a semi-local centrality measure to balance between both centrality measures. Their results on four real-world networks showed much faster computational efficiency while providing more effective results than degree centrality methodologies. However, this methodology was not compared with greedy-based approaches for mining influential nodes.

Kim and Yoneki (2012) studied the problem of maximizing influence diffusion through social networks. The authors underlined the weaknesses of existing techniques that use arbitrary node propagation and proposed the Influential Neighbors Selection (INS) scenario to select the most effective group of neighbouring nodes to propagate. Four selection strategies were proposed, and the results showed that *highest degree selection* had favorable results for short-term diffusions but *random selection* performed better in long-term scenarios. *Highest weight and volume selections* showed similar results to *highest degree selection*, but the additional communication costs meant they were less favorable.

Based on the Dempster-Shafer evidence theory (Dempster 1967; Shafer 1976), Wei et al. (2013) proposed a new evidential centrality measure for identifying influential nodes in weighted networks. Their approach considers not only degree and weight of nodes but also status of the nodes in a weighted network. By experimentally comparing with other measures such as the degree and closeness centrality, the proposed measure showed comparative performance for identifying influential nodes. Again, the evidential centrality was not compared with any greedy-based

Table 3. Comparisons of Various Centrality Measure-based Approaches

Approach	Identify influential node neighbourhoods	Identify influential nodes
PCC (Ilyas and Radha 2011)	Yes	Yes
Semi-local centrality measure (Chen et al. 2012)	No	Yes
Four selection strategies (Kim and Yoneki 2012)	Yes	Yes
Evidential centrality measure (Wei et al. 2013)	No	Yes

approaches for top-k influential nodes identification. This common trait of centrality measure-based approaches attracts our attention to the issue. Future methodologies in this category need to compare performances with greedy-based approaches to be more convincing on effectiveness and efficiency of their proposed measures.

The techniques presented in this section showed many similarities. Both Ilyas and Radha (2011) and Wei et al. (2013) evaluate centrality measures while using Susceptible-Infected related models to evaluate the performances of their proposed models. The evaluation of traditional centrality measures (degree centrality, betweeness centrality, and closeness centrality) was also a common theme in the work by Chen et al. (2012) and Wei et al. (2013), and both studies present heuristics that outperform these centrality measures in terms of influence diffusion while maintaining lower computational complexity. Kim and Yoneki (2012) adopt the IC model to select the most influential neighbours of a node. Experimental results on two evaluation metrics were presented. The first metric was Pearson correlation coefficient between closeness centrality and the proposed four selection strategies. For the second metric, those selection methods were compared against each other by computing the ratio of the number of activated nodes to the total number of nodes in the network. The similarities and differences of these approaches are summarized in Table 3.

3.3 Topic-Based Approaches

In a viral marketing network, top-k influential nodes are those that, when convinced to adopt a product, shall influence others in the network and lead to a largest possible number of new adoptions. Although real-world product purchasers have different degrees of interest on various topics, e.g., some customers are only interested in the latest smart-phones, while others tend to pay more attention to new computer models; however, both the greedy and centrality measure-based approaches are topic blind. In this aspect, the two kinds of approaches treat all possible topics in a network as if they were the same and focus only on general inherent attributes of nodes (such as marginal influence gain or centrality measure value) regardless of any particular topic or context. This problem is addressed in Tang et al. (2009), Liu et al. (2010), Barbieri et al. (2013), and Aslay et al. (2014), and we review these topic-based approaches in this section.

The Topical Affinity Propagation heuristic proposed by Tang, Sun, Wang, and Yang (2009) describes the topic-aware influences in large-scale social networks. This tool allows users to perform meaningful and valuable influence analysis on real-world datasets by (1) finding the influential nodes on a given topic and (2) identifying the social influences of the neighbouring nodes for a particular node. The authors accomplish this by integrating learning algorithms into their Topical Factor Graph (TFG) model. The features of the TFG model allow users to find both the local information of nodes (such as topic-level influences and next most probable propagation) and global information (connectivity between any two nodes). In conjunction with these components, the authors adopted distributed learning techniques to train the TFG model due to the size and complexity of large networks. The programming model, Map-Reduce, partitions large social networks into subgraphs so multiple machines can process and collect values associated with nodes

in both internal and external subgraphs. When adopting the TFG model with real-world datasets, their results showed that the distributed learning approach has good scalability performance, and topic-level influences can improve the performance of influential nodes finding.

Liu et al. (2010) focused on investigating how to mine topic-level influence in heterogeneous networks. The authors solve the problem in two steps: (1) They proposed a model that combines the "heterogeneous link" (Liu et al. 2010) information with the text associated with nodes in the network to determine topic-aware direct influence, and (2) from the direct influence that was validated, a topic-aware algorithm was proposed to determine indirect influence between nodes. Their experiments with Twitter, Digg, and citation networks showed that the proposed approach can unveil useful influence patterns in heterogeneous networks.

Barbieri et al. (2013) extended the traditional IC model (Goldenberg et al. 2001) to be a Topic-aware Independent Cascade (TIC) model. Their experiment results showed that the TIC model can describe real-world influence–driven propagations more accurately than the state-of-the-art topic-blind models.

As a first step toward enabling social-influence online analytics in support of viral marketing decision making, Aslay et al. (2014) proposed an efficient index for a general type of topic-aware viral marketing queries. Given the computational challenges related to the enormous number of potential queries and some other aspects, the authors employed a tree-based index, INFLEX, to obtain a solution for a limited number of possible queries. Their experiment results showed that with the index, the targeted queries can be answered in milliseconds instead of several days compared to existing offline computation methodologies.

Zhou et al. (2014) explored topic or preference-based top-*k* influential nodes identification in social networks. The proposed mining algorithm, GAUP, finds influential nodes on a given topic in two stages; First, by computing user preferences and projecting them into a "reduced latent space and a VSM-based model" (Zhou et al. 2014). In the second stage, GAUP utilizes the greedy-based approaches, e.g., CELF (Leskovec et al. 2007) to find influential nodes for a particular topic. When evaluated with an academic network, GAUP was shown to maximize influence spread for a given topic.

Although identifying topical influential nodes in social networks is an interesting research direction, there has not been significant research performed in this area. From three different aspects, Table 4 compares four topic-based approaches reviewed in this section.

3.4 Network Content or Topology-Based Approaches

With the increasing quantity of content in social networks, it is possible to conduct content-centric mining of influencers. In this section, we review network information- (or network structure) based approaches for identifying top-*k* influential nodes, which provides a different perspective from the topic-based approaches.

Community-based Greedy Algorithm (CGA) (Wang et al. 2010) takes the community property of mobile social network (MSN) into consideration. Two major components of the CGA are (i) an algorithm for community detection and (ii) a dynamic programming model for choosing communities to identify influential nodes. Wang et al. extended the Independent Cascade model (Goldenberg et al. 2001) to account for the edge weight of MSNs and provide provable performance guarantees for CGA. In terms of efficiency and approximation error rate, CGA is proven to outperform some state-of-the-art greedy-based approaches for identifying top-*k* influential nodes. Similarly to Wang et al. (2010) and Bozorgi et al. (2016) also utilized the network community structure in their research. A community-based algorithm, INCIM, was proposed for identifying influential nodes under the Linear Threshold model (Granovetter 1987). To summarize, the network community structure is used to find the influential communities, while a node's influence is determined by

Table 4. Inputs, Outputs, and Characteristics of Various Topic-based Approaches

Work	Inputs	Outputs	Characteristics
TFG (Tang et al. 2009)	• a social network • a prior topic distribution for each node (inferred input)	topic-wise user-to-user influence strength	• works for large-scale networks • ability to find both the local and global information of nodes
Topic-level influence mining in heterogeneous networks (Liu et al. 2010)	an heterogeneous social network with nodes that are users and documents	topic distribution and user-to-user influence	a probabilistic model for the joint inference of the topic distribution and topic-wise user-to-user influence
TIC (Barbieri et al. 2013)	• a log of past propagations • model parameters	more accurate descriptions of influence driven propagations	proposed TIC model by extending IC model
INFLEX (Aslay et al. 2014)	a directed social graph where the edges are associated with a topic-dependent user-to-user social influence strength	a set of users that should be targeted in a viral marketing campaign for a given topic	• answer queries online within few milliseconds • based on the TIC model

a combination of its local and global influences. Their evaluation results demonstrated that INCIM outperforms other approaches, such as the simple greedy algorithm, in the quality of the identified influential nodes while maintaining a reasonably low running time and memory consumption for large graphs.

PageRank (PR) (Brin and Page 1998) and Topic-sensitive PageRank (TSPR) (Haveliwala 2002) were originally proposed for the purpose of ranking webpages. However, in recent decades, a number of research on identifying influential nodes adopt, extend, and integrate PR or TSPR into their own methodologies (Chen and He 2015; Weng et al. 2010). Some existing research utilize PR and (or) TSPR in their experimental comparison analysis (Romero et al. 2011; Weng et al. 2010). Therefore, we briefly introduce the core ideas of these two algorithms: PageRank applies academic citation literature and orders webpages by calculating the number of citations and backlinks. Detailed descriptions of the PageRank algorithm can be found in Page et al. (1999). PageRank has been often adopted in research for identifying influential nodes due to numerous reasons: First, in a graph composed of tens of millions of webpages, each webpage can be represented as an individual node; Second, Brin and Page (1998) state that "*a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank.*" As, correspondingly, for top-k influential nodes, a node can have a high influence if there are many nodes that directly connect to it or if there are some nodes that directly connect to it and have a high influence. In 2002, Haveliwala (2002) proposed Topic-sensitive PageRank by computing multiple PageRank vectors, biased using various topics, to capture more accurately the notion of importance regarding each particular topic.

Romero et al. (2011) acknowledged the importance of user passivity when finding influential nodes in social media. The proposed general model for influence analysis uses the concept of passivity in a social network and develops an algorithm for quantifying the influence of all the nodes in the network. Their method utilizes the network structural properties as well as the diffusion behaviors among users. The influence of a user is determined by both the size of the influenced audience and their passivity. It is claimed that the model outperforms other measures of influence, i.e., PageRank (Brin and Page 1998) and degree centrality. The authors stated that high popularity

does not necessarily imply high influence. A similar opinion is also expressed in the greedy-based approaches such as Kempe et al. (2003).

Subbian et al. (2014) questions the lack of actual social value of network collaborations in influence diffusion models and defined individual social capital as social value generated through collaborations with peers of high influence. Through their proposed algorithm, SoCap, they were able to find influencers based on the hypothesis that people with high social capital indicates high influence within a network. Due to this, SoCap differentiates itself from other measures of influence, e.g., degree centrality and PageRank, as it finds high social value nodes through multiple collaborations. Unlike the greedy-based approaches, SoCap does not use any underlying influence diffusion models. A value-allocation model is also developed to compute the social capital and allocate the fair share of this capital to each individual involved in the collaboration.

A Context-Aware and Trust-Oriented influencer-finding method (Zhu et al. 2015), named CT-Influence, incorporates social contexts, such as the preferences of participants with the social relationship and trust between participants. The experiments with two real-world datasets showed that CT-Influence greatly outperforms SoCap (Subbian et al. 2014) in terms of effectiveness and efficiency for identifying influential nodes.

Chen and He (2015) take hostile relations in Online Social Networks (OSNs) into consideration when integrating the PageRank algorithm (Brin and Page 1998) on signed OSNs. The authors used the integrated PageRank to discover influential nodes in OSNs with both friend and hostile relations, which correspond, respectively, to positive and negative edges on signed networks. Their experiment results for selecting top-k influential nodes on real-world datasets indicated that the proposed method performs better than some algorithms, such as the original PageRank (Brin and Page 1998).

In this section, we reviewed network topology or content-based approaches for identifying top-k influential nodes. The network topology–oriented methods focus on exploiting the community structure property of social networks. The content-based techniques tend to extract various aspects of information from a given social network, such as friend and hostile relations, user passivity, social value of collaborations, the preferences of participants, the content or topic of tweets, and so on. For the presented techniques, we notice that PageRank was a popular comparison technique and a Twitter dataset was frequently analyzed. We provide comparisons of techniques discussed in this section in Table 5.

3.5 Identifying Top-k Influential Nodes in Dynamic Social Networks

In the previous four sections, we have addressed literature on top-k influential nodes identification in static social networks. This section discusses the existing work on identifying influential nodes in dynamic graphs, which is a relatively new research area. The associated computational challenge is a major difference in the two network settings of static and dynamic networks. Static social network can be considered as a snapshot of dynamic network, which cannot fully represent some characteristics of real-world social networks, such as continuous network topology change, high-speed data transmission, large population of participants, and uncertain information diffusion processes. More computational resources are potentially demanded due to these distinct characteristics of dynamic social networks.

Aggarwal et al. (2012) are among the first who study influential nodes identification in dynamic social networks. A stochastic approach is designed by the authors to identify the information flow authorities with two types of methods: a globally optimized forward trace approach and a locally optimized backward approach. In addition, methods for determining the approximately optimal

Table 5. Comparisons of Various Network Content or Topology-based Approaches

Approach	Network topology-based approach	Network content-based approach	Key Characteristics
CGA (Wang et al. 2010)	Yes	No	• community detection • dynamically select communities to find influential nodes
INCIM (Bozorgi et al. 2016)	Yes	No	• influential community detection • the influence of each node is determined by its local and global influences
IPR (Chen and He 2015)	No	Yes	• signed social networks • friend and hostile relations are represented as positive and negative edges on signed networks respectively
TwitterRank (Weng et al. 2010)	No	Yes	the topical similarity between users and the link structure are both taken into account
IP (Romero et al. 2011)	No	Yes	determines the influence and passivity of users based on their information forwarding activity
SoCap (Subbian et al. 2014)	No	Yes	captures the individual social capital
CT-Influence (Zhu et al. 2015)	No	Yes	considers the social trust and relationships between participants and the preferences of participants

release points for a given pattern of information spread are also proposed. The performance of proposed methods was evaluated on both DBLP and ArnetMiner citation datasets.

Zhuang et al. (2013) underlined the influence maximization problem in dynamic networks as probing nodes in an unobserved network. The Maximum Gap Probing (MaxG) algorithm was proposed to provide an approximate optimal solution to probe a subset of nodes that can best unravel the actual influence diffusion process in the network. The authors claimed that MaxG is a general method and can be directly used to guide online marketing decisions in social networks. Experiment datasets used for evaluating the performance of MaxG were constructed from Twitter and the coauthor network of ArnetMiner. Unlike in MaxG, which focuses on probing a subset of nodes in a dynamic network, Han et al. (2017) adopt a divide-and-conquer strategy to capture the global evolution of a dynamic network by only probing the most active communities. Extensive experiments were conducted by the authors on Epinions, Slashdot, Twitter, and Inventor networks.

A content and network-based flow mining approach for dynamic influence analysis is proposed by Subbian et al. (2013). In their research, sequential patterns were dynamically mined from a combination of the keywords and the dynamic network in the social stream. These were then used to discover the most influential nodes in a dynamic and evolving network. Three years later, the authors develop an algorithm, InFlowMine (Subbian et al. 2016a), to determine flow patterns through content propagation on the dynamic network structure. The identified patterns were utilized for determining topic- or network-specific influential users or their combinations. Unfortunately, not all OSNs can provide sufficient context for discovering information flow patterns. Hence, this streaming method cannot be generalized for finding social influencers across OSNs. Three datasets were used in their evaluation experiments, which are Twitter, DBLP, and US patent datasets.

Vadoodparast and Taghiyareh (2015) focus on maximizing product adoption in dynamic networks by proposing a multiagent framework named MAFIM. The framework contains two kinds

of agents: modeling agents and solution provider agents. A dynamic network is viewed as consecutive static snapshots by these agents, and according to a selected budget assignment policy, each snapshot obtains its share from the budget defined by the sales manager. Their experiment results on real-world dataset, such as Slashdot, demonstrate that it is more effective to launch many short-lived campaigns rather than few long-lived ones.

Subbian et al. (2016b) proposed the first general keyword-based influence query and tracking model for streaming scenarios. With the constant evolution of topics over time, which potentially results in different identified influencers, the authors developed a method to maintain real-time influence scores of users in a social stream based on topic and time-sensitive information. The core idea of the method is to track information flow patterns in a treelike data structure across various paths of the network in a context and time-sensitive fashion. The data structure facilitates one pass computation of influencers for different contexts.

A Dynamic Independent Cascade (DIC) model and the concept of adaptive seeding strategy were proposed by Tong et al. (2017). Using a simple greedy algorithm, the authors provided a provable performance guarantee for their solution based on the DIC model. Additionally, an efficient heuristic algorithm with better scalability was also introduced. The superiority of the adaptive seeding strategies was demonstrated by empirically comparing with the state-of-the-art non-adaptive seeding approach (Kempe et al. 2003) and random strategy. Two kinds of real-world social networks used in their experiments were Hep and Wiki. Hep contains academic collaboration data of co-authorships in physics, while Wiki includes the Wikipedia voting data (Jure Leskovec 2010) from the inception of Wikipedia.

Song et al. (2017) designed an algorithm, called UBI, to solve the influential node tracking on dynamic social network problem. The core idea of their approach is to start from the influential seed set that was identified previously and implement node replacement to maximize the influence coverage. Three datasets evaluated in their experiments are mobile, HepPh, and HepTh networks. A very recent technique proposed by Wang et al. (2017) adopts the concept of sliding window to maintain "a set of k seeds with the largest influence value over the most recent social actions." The authors collected two real-world datasets, Reddit and Twitter, for their experiments.

Since many social networks, such as Twitter, are often available only in the form of social streams of user activity, there is a surge of recent research on social streams. However, our survey indicates that a very limited amount of work has been done on identifying top-k influential nodes in dynamic social networks. Among them, only a minority provide a provable performance guarantees (Tong et al. 2017; Wang et al. 2017; Han et al. 2017) or empirically compare their proposed methods with dynamic network–oriented approaches (Wang et al. 2017; Han et al. 2017). Three major categories of techniques presented in this section are (1) overall structure and information diffusion of the dynamic network focused approaches (Aggarwal et al. 2012; Tong et al. 2017), (2) network content or content flow patterns-based approaches (Subbian et al. 2013, 2016a, 2016b), and (3) identifying influential nodes in dynamic networks by modelling a subset of the network with the subset being a set of nodes or communities (Zhuang et al. 2013; Vadoodparast and Taghiyareh 2015; Song et al. 2017; Wang et al. 2017; Han et al. 2017). We conclude Section 3.5 by presenting two facts that exist among these works:

(1) *Model Change:* Traditional influence diffusion models (such as IC (Goldenberg et al. 2001) and LT (Granovetter 1987)) have a static network structure and edge propagation probabilities. Although they are widely used in the greedy-based approaches for identifying influential nodes, they are not adopted in dynamic network settings. Instead, more flexible models are proposed to accommodate dynamic network structure and edge propagation probabilities.

(2) *Content-Centric Method:* Network content, such as information flows and topic- or time-sensitive data, are well incorporated and utilized in approaches for identifying influential nodes in streaming social networks.

3.6 Section Summary

Since the first formalization of finding the influential actors problem by Kempe et al. (2003), there has been an increasing amount of research and development in the area. An interesting phenomenon observed is that the focus of this research area has been shifted from efficiency and scalability to dynamic networks in recent years. With various features provided by a wide variety of algorithms, applications will be required to select suitable algorithms to accommodate for its specific problem domain when mining top-k influential nodes.

Node and edge attributes have also been an underlying theme for top-k influential node identification. The research by Chen and He (2015) was based on friend and hostile attributes between nodes in the network, creating positive and negative edges on signed networks. Zu et al. (2015) incorporates social contexts into the network analysis, integrating humanistic factors into the properties of participating nodes.

4 TOP-K SIGNIFICANT NODES IDENTIFICATION

In this section, we focus on literature relevant to significant nodes identification. By reviewing recent work, we hope to shed some light on this research area.

4.1 Effector-Based Identification and Bridge Nodes

When we find nodes in a social network in a particular activation state, such as when a certain topic is popular, we can often observe the subset of root nodes that propagate the information and activate the neighbouring nodes in the network, known as *effectors*. Unlike influential nodes that focus on centrality measures, effector nodes are key connectors in a network due to their property as bridges between peripheral nodes and groups even when they might not have a high degree of neighbouring nodes themselves and, therefore, if removed, usually cause networks to be fractured and disjoint.

Given a social network graph G and an activation vector a, Lappas et al. (2010) defined k-effectors to be the set of nodes, once activated, that cause an activation pattern that is as similar as possible to the activation procedure observed at a. The authors proved that the k-effectors problem can be solved in polynomial time for social networks represented as a tree. This is accomplished through a dynamic programming approach by specifying the maximum k-effectors of sub-trees under one of the two following approaches: The root of the sub-tree is included in the set of effectors for the next recursion and then recurses on children with (k-1) budget. The second approach does not include the root and the children are recursed with k budget. With the method described, the authors were able to extract the most probable active tree that spans all the active nodes of the network. Using this, they could identify the optimal set of effectors in the social network tree, and, in return, they were able to extract interesting observations on the network and interactions between significant nodes.

The research by Li et al. (2017) attempts to identify sets of *star nodes* (which are also the significant nodes in this case) based on the scale of connectivity loss if the nodes were removed from the network. The authors account for typical immunization methodologies, such as acquaintance immunization, and performed analysis on two generic social networks. The authors discovered that certain immunization strategies, such as selected immunization, cause entire networks to be disjoint if the targeted star nodes were removed, while random and acquaintance immunization provides less destructive results with better running times. This methodology separates itself from

previous work, such as that by Lappas et al. (2010), as it views effectors from a different perspective. Immunization strategies are more generally adopted to prevent outbreaks and, in this case, to identify the loss of connectivity in the network if the top-*k* significant nodes are removed.

Borgatti (2006) underlines the weaknesses of top-*k* node identification using centrality-based approaches, where they do not account for key players' role in a network's cohesiveness. The author proposed a model that ranks significant nodes by both their centrality measures (KPP-Pos) and loss of graph cohesion if removed (KPP-Neg).

Borgatti's (2006) research incorporates concepts from Lappas et al. (2010) by valuing a significant node's centrality while also weighting the consequences if the node was breached. Similarly to Li et al. (2017), this research underlines the importance of the top-*k* effectors from a loss of network connectivity perspective. This concept has also been described by researchers such as Musial and Juszczyszyn (2009) as *bridge nodes*. In their research, Musial and Juszczyszyn describe bridge nodes as anchoring nodes that connect peripheral nodes with the rest of the network and, if lost, can cause diffusion loss between nodes in the network. The authors conducted an experiment on the Thurman office social network in an attempt to identify and extract bridge nodes and their properties. From their results, they have established that the bridge nodes were usually nodes with the highest social position and can be categorized based on their neighbouring connections with peripheral nodes or cliques.

4.2 Authoritative-Based Identification

With the significant growth of the Internet, the ability to find sources of information with high levels of quality and authority has become increasingly difficult. With many Online Question and Answer portals facing this problem, there has been an emergence of research aimed to distinguish the top-*k* set of significant nodes (or authoritative users in this context). Farahat et al. (2002) analyses documents scattered in the World Wide Web and determines the reliability and authoritativeness of these documents based on textual, non-topical cues. In their research, the authors discovered that when querying certain subjects, documents found by PageRank (Brin and Page 1998) were sometimes uninformative and even controversial. To estimate the authority of documents more accurately, the authors combined the analysis of textual content of documents with its linguistic features and found that this approach was often able to rank authoritative documents produced by professionals and subject-matter experts higher than PageRank.

Zhang et al. (2007) attempts to identify a set of users with high expertise on a Java Forum. In their research, they evaluated several network-based ranking algorithms such as HITS Authority (Kleinberg 1999) and separated users into five expertise ratings. From their results, they found several behavioral patterns among users of different expertise levels, such as newbies making few posts and experts answering other users' questions while asking very few themselves. These asker–helper interactions overlap with several properties of effector-based identification where information from a single source (which in this case, is knowledge of the answer to a specific question) is able to activate downstream lower-level users.

Jurczyk and Agichtein (2007b) present authoritative node identification in a more controlled environment. The authors crawled through almost half a million questions on Yahoo! Answers portal with three million corresponding answers in the attempt to estimate the authority of users while using the HITS algorithm (Kleinberg 1999) as a baseline. For each question, the authors extracted the number of answers, the sum of the answers' voted values, and the average number of stars of the best answer's author and drew comparisons between each of these categories based on the Pearson correlation coefficient. What they found was that authority was easier to identify in particular subject domains and using votes and stars produced results based on the popularity

and quality of the feedback, respectively. Unlike the research by Farahat et al. (2002), Jurczyk and Agichtein's dataset contains several properties that the authors leveraged.

Unlike scattered documents in the World Wide Web, the datasets from Zhang et al. (2007) and Jurczyk and Agichtein (2007b) separated questions into distinct subjects and domains, which reduces the possibility of irrelevant nodes or answers being found for a particular query. Second, with each question, there exists a set of user-defined properties (such as the Voting and Star system in Yahoo! Answers) that exposes a new layer to evaluate the quality and authority of a particular node, underlining the value of useful node attributes to users and researchers. All of the research presented attempts to identify top-k significant nodes from various angles, but the effectiveness and accuracy of the algorithms are heavily influenced by the dataset and its attributes.

4.3 Academic Datasets

In terms of identifying top-k significant nodes, there are some other studies investigating the significance of researchers in academic co-author networks. Nascimento et al. (2003) investigated the co-authorship graph obtained from all papers published at SIGMOD between 1975 and 2002. They utilized the evolution of minimum closeness centrality scores as the ranking criteria to evaluate the significance of authors. Moreover, Liu et al. (2005) built a weighted directional model to represent the co-authorship network, for which they defined AuthorRank as an indicator for the prestige of an individual author in the network. Under the assumption that program committee members can be regarded as prestigious actors in the field, their results are validated against conference program committee members in the same time period.

The evaluation results showed that the use of AuthorRank has clear advantages over traditional centrality measures, e.g., degree and closeness metrics. Zhou et al. (2007) acknowledge the remarkable success demonstrated by graph-theoretic approaches for ranking networked entities. However, they also pointed out that the majority of the methodologies can only be utilized on homogeneous networks, such as the citation network. To rank significant authors and documents, Zhou et al. proposed a framework for co-ranking entities of two different types in heterogeneous networks. Their method couples two random walks into a combined one to co-rank authors and their publications using information retrieved from several networks: the social network connecting the authors, the citation network connecting the publications, as well as the authorship network that ties the previous two together. Their results suggest that the rankings of authors and documents depend on each other. Zaïane et al. (2009) used an iterative random-walk algorithm to evaluate the relevance between significant authors for the purpose of discovering research communities. Their core idea is to use a random walk on the bipartite or tripartite model of DBLP data and generate a relevance score to measure the closeness between two entities. The relevance score is then used to rank entities based on importance of a given relationship.

4.4 Other Work

Among the numerous existing top-k significant node heuristics, the PageRank (Brin and Page 1998) and the HITS (Kleinberg 1999) algorithms are the most popular and considered as baselines for many later techniques. As described previously in Section 3.4, PageRank is a top-k heuristic that determines the quality and relevance of a webpage based on the search topic, which behaves like our definition of significant node identification. The HITS algorithm queries the World Wide Web in two steps: (1) a sampling stage that constructs a collection of webpages that are likely to be relevant authorities and (2) a weight propagation step that iteratively estimates a hub and authority score for each webpage and returns the highest scoring authorities of each topic. HITS has led to a number of future works such as work by Joel et al. (2001) that enhances the original

HITS algorithm with exponential inputs and Wu et al. (2006) that integrates collaborative tagging to support community detection, user, and document recommendations.

Tseng and Chen (2012) proposed a novel node ranking methodology for general real-world applications. This technique contains an unsupervised learning algorithm that requires no training data to produce a ranking list of top-k significant nodes. The authors implemented their methodology and experimented on a co-author network of from the DBLP computer science bibliography. The network was structured as an undirected graph where each node represents an author, and each edge between two nodes indicates publication collaboration between two authors. This process consists of two parameters and major phases. The user-defined parameters are the set of desired features and the number of significant nodes wanted. The first phase was an offline procedure where several features are extracted from each author in the network and sorted lists of author features are prepared; these lists are used as the input of the ranking algorithm in the second phase. The primary principle of the ranking strategy is to find top-k significant nodes with overall ranks higher than others. The methodology was able to generate different ranking lists when diverse sets of desired features are considered.

In addition to the ranking methodology, Tseng and Chen (2012) also claimed that the definition of significant nodes is application dependent as it varies with circumstances in different networks formed by diverse kinds of social connections. To accommodate different application characteristics, the proposed methodology requires a set of desired features as one user-specified parameter. While this technique provides a wide range of features to meet a variety of service demands, it has the tradeoff of requiring more feature-related data and lowering efficiency for processing the two separate phases. The preparation of sorted author lists demands large volumes of input data on various features to be processed offline, which might not always be realistic for real-world applications. Additionally, adequate data might not be readily available, and the application could require streaming data to be processed online rather than offline. The degree and closeness centrality measures are the two of four desired features in the experiment evaluation, where Tseng et al. avoid giving a specific weight on each desired feature for the purpose of providing a more objective assessment. It is highly likely that the resulting weighting-free ranking algorithm is not suitable for applications requiring subjective assessment.

4.5 Section Summary

From the categories described in Section 3, we found that processes for identifying top-k influential nodes were mostly greedy-based approaches, whereas for top-k significant nodes, we find a wider array of different methodologies adopted. Section 4.1 describes the characteristics of effectors, star nodes, and bridge nodes in social networks. While in Section 4.2, methodologies for identifying authoritative nodes are presented.

Additionally, we emphasized the importance of node attributes and network properties on how they can affect the methodology and process for identifying significant nodes and studying the network. The studies by Zhang et al. (2007) and Jurczyk and Agichtein (2007b) both contained datasets with very specific node and edge properties that influenced the direction of their research and heuristics.

In addition, there is a paper by Lappas et al. (2011) where more details on topics such as algorithms and systems for expert location in social networks are discussed.

5 APPLICATIONS OF IDENTIFYING TOP-K NODES IN VARIOUS NETWORKS

With the tremendous increase of usage in OSNs, research in the identification of top users of OSNs such as Facebook, Twitter, and LinkedIn has increased over the past decade. In this section, we review numerous applications of identifying top-k nodes, and some of these applications showed

Table 6. List of Key Applications of Top-k Nodes Identification

Domain	Related Research
Twitter	(Cha et al. 2010) (Weng et al. 2010) (Bakshy et al. 2011) (Drakopoulos et al. 2016)
Facebook	(Heidemann et al. 2010) (Kim and Han 2009)
Blogosphere	(Gruhl et al. 2004) (Java et al. 2006) (Java et al. 2007) (Huang et al. 2016)
Misinformation Control	(Budak et al. 2011) (Nguyen et al. 2012)
Community Question Answering	(Opsahl et al. 2010) (Zhang et al. 2007) (Guo et al. 2008) (Pal and Konstan 2010)
Networks with Complex Topologies	(Opsahl et al. 2010) (Wei et al. 2013) (Zhou et al. 2007)
Miscellaneous Applications	(Ghosh and Lerman 2010) (Aral and Walker 2012)

how the use of content can enhance the identification process. It is evident from the discussion of this section that top-k nodes identification is useful for a wide variety of inter-disciplinary domains such as Twitter or Blogosphere. Due to the limitation of length, we focus on only a small number of successful applications: (1) Twitter (Section 5.1), (2) Facebook (Section 5.2), (3) Blogosphere (Section 5.3), (4) Misinformation Control (Section 5.4), (5) Community Question Answering (Section 5.5), (6) Networks with complex topologies (Section 5.6), and (7) Miscellaneous Applications (Section 5.7). An overview of the key applications are summarized in Table 6.

5.1 Twitter

With the rising popularity and usage of online social networking mediums in the past decade, knowledge on how information is shared and distributed to users on these mediums has become increasingly valuable to many corporations that seek to advertise their products and services. Among these networking services, one of the most common is Twitter.

The social infrastructure of one of the most notable micro-blogging services, Twitter, is composed of twitterers, users who publish/provide tweets (with a limit of 140 characters) or content to all of their connected followers. The information distributed by a few core and influential twitterers (or tweets) could have a much greater impact and distribution of information than a large group of random ones. Hence, there has been an increasing interest in finding these core and influential twitterers (or tweets) from various perspectives, i.e., Weng et al. (2010) and Bakshy et al. (2011).

Weng et al. (2010) presented an approach of identifying influential twitters by employing an extension of the PageRank algorithm (Brin and Page 1998) (previously described in Section 3.4). The proposed algorithm, TwitterRank, considers "both the topical similarity between users and the link structure" (Weng et al. 2010). In contrast to many other existing research on Twitter datasets, the authors pointed out that the number of followers alone may not accurately represent the influence due to "reciprocity," where users follow their followers as an act of social courtesy as opposed to "homophily," where users follow due to mutual topic interests. To establish whether "homophily" exists in the Twitter community, the authors proposed two underlying questions:

(1) Are twitterers with "following" relationships more similar than those without according to the topics they are interested in?
(2) Are twitterers with "reciprocal following" relationships more similar than those without according to the topics they are interested in?

To answer these questions, Weng et al. analyzed large volumes of unlabeled tweets (content) to automatically distill topics and determine if the relationship between influential twitterers and followers is due to topic sensitive influence. Their experiment results showed that, first, "homophily" does exist in the context of Twitter, and the authors claim that they are the first to report this. Their observation justifies that there are some twitterers who in fact "follow" someone due to common topical interests instead of just playing a "number game." Second, their proposed approach outperforms the benchmark technique that is currently used by Twitter and other related algorithms, e.g., in-degree (i.e., the number of followers) and the original PageRank (Brin and Page 1998).

In addition to TwitterRank (Weng et al. 2010), three other Twitter-specific ranking algorithms are TunkRank (Tunkelang 2009), inDegreeRanking (Kwak et al. 2010), and IARank (Cappelletti and Sastry 2012). TunkRank adapts PageRank (Brin and Page 1998) and defines influence as the attention a user is able to give to the tweets he or she receives combined with the attention that his or her followers can give to him or her. Kwak et al. (2010) proposed to rank users based on the number of followers (in-degree) and found the produced ranking was similar to PageRank. Unfortunately, a high in-degree could be made up by simply creating fake usernames that follow a user whose ranking is to be increased, making it an easy loophole for spammers. The key characteristic of IARank is its ability to rank users on Twitter in near real time. The basis of IARank is information amplification potential of a user, which is evaluated by the capacity of the user to increase the audience of a tweet or another twitter that they would find interesting, by retweeting or mentioning.

TunkRank (Tunkelang 2009) converges to the final ranking in an iterative way. The convergence time is determined by the number of users considered in the ranking process. As opposed to IARank (Cappelletti and Sastry 2012), TunkRank is not capable of producing a ranking list in real time. TwitterRank (Weng et al. 2010) focuses on topical-oriented influential twitters mining by distilling available topics, and it is also not amenable to be calculated in real time.

Cha et al. (2010) compares three different measures of influence in Twitter (number of followers, number of retweets, and number of mentions) and finds that the most followed users do not necessarily score highest on the other two measures. Number of followers (in-degree) represents a user's popularity; however, it is not directly correlated to other important perspectives of influence such as engaging audience. *Retweets are driven by the content value of a tweet, while mentions are driven by the name value of the user. Such subtle differences lead to dissimilar groups of the top influential Twitter users.* Focusing on retweets and mentions, the authors investigate the dynamics of user influence across topics and time using a large amount of data collected from Twitter.

The research by Bakshy et al. (2011) studies the information propagation of influential sources when compared to "ordinary influential users." One of the key functionalities of Twitter is the ability for a user to "retweet" or repost the contents of another twitterer. However, due to the 140-character limit of tweets, a common strategy adopted by twitterers is to attach content to shortened URLs (e.g., bit.ly) to effectively condense content in an easily distinguishable form. Additionally, because of these unique tokens, it is easy for the authors to identify the depth or level of cascade from which a tweet is propagated from a single influential source. Therefore, the authors study the URLs that twitterers add to their tweets and the overall cost-effectiveness of marketing through a small group of key influential twitterers as opposed to a large group of average or under-performing ones. The focus of the Bakshy et al.'s (2011) study is explicitly on the overall influence of a post rather than the traditional qualitative measure of user influence, and, hence, their methodologies are based on the "influence score" of URL posts and the diffusion level of URL reposts from the original "seed" until termination of the cascade. Their results showed that while the majority of posted URLs do not spread (with an average cascade size of 1.14), some distinct ones are able to spread as far as nine generations of repost. Using these results, the authors were able

to compare the effectiveness of "seed" influencers with word-of-mouth campaigns and find that using a quantitative approach of influence of a post has similar results to the qualitative study of influential users. Both methodologies showed that since the majority of posts from a single source do not spread at all, targeting certain "seeds" with high rates of diffusion as opposed to multiple low-performing ones generates the most coverage.

Pal and Counts (2011) discovered topical authorities in Twitter. There are three steps in their approach: First, characterize Twitter users with social media-specific features, including both nodal and topical metrics. Second, cluster the users over feature space. Last, produce a list of the important authors for a specific topic utilizing a within-cluster ranking procedure.

In 2016, Drakopoulos et al. (2016) studied five Twitter-specific metrics for ranking Twitter influence. Based on concepts from system theory, a methodology for evaluating influence metrics is also proposed. In addition, the authors implemented the five metrics in Java over Neo4j, which is a graph database that provides production grade front or back-end social graph storage.

5.2 Facebook

Being the largest social network, Facebook has more than 400 million active users with an average of 130 friends (Facebook 2017). The Facebook dataset has been utilized in various top-k nodes identification research.

Heidemann et al. (2010) proposed an adapted PageRank algorithm to identify influential users in a social network according to activity links. The approach was evaluated on a Facebook dataset and found that more active users who are retained can be identified when drawing on users' prior communication activities or centrality measures, e.g., degree centrality. Kim and Han (2009) identify influential users in a network graph by first calculating degree centrality based on social links and then estimating an activity index. The proposed method was evaluated by observing the influence diffusion of a Facebook game. Their experiment results show that by targeting the identified influential users, game growth rates and number of new game adopters can be increased.

Although Facebook has been frequently used as experiment data for identifying top-k nodes, our literature research shows that so far only a few Facebook-oriented algorithms exist. This phenomenon differs significantly from the Twitter domain, as a number of top-k nodes identification approaches have been designed specifically for Twitter, such as TunkRank (Tunkelang 2009), TwitterRank (Weng et al. 2010), and IARank (Cappelletti and Sastry 2012).

5.3 Blogosphere

With an increasing amount of blog posters and readers, Blogospheres (a network of blogs) are an effective and inexpensive medium for companies to evaluate their advertising campaigns. Hence, it is interesting to observe the emerging research that has begun identifying top-k nodes in the weblog domain.

Gruhl et al. (2004) model the diffusion of topics among blogs, determined by the textual content of the weblog. The authors managed to characterize information propagation from two perspectives: topics and individuals. The proposed model makes it possible to "identify particular individuals who are highly effective at contributing to the spread of infectious topics" (Gruhl et al. 2004).

Java et al. (2006) presented an analysis of influence models, i.e., the Linear Threshold model (Granovetter 1987) and the Independent Cascade model (Goldenberg et al. 2001), on a large-scale and real-world blog graph. Their evaluation results suggest that PageRank (Brin and Page 1998) is an inexpensive approximation to the simple greedy algorithm (Kempe et al. 2003) for selecting an influential set of bloggers, which maximizes the spread of information on the blogosphere. To

Table 7. Comparisons of Various Techniques for Identifying Top-k Nodes in Blogosphere

Technique	Identify influential bloggers	Identify influential blogs	Utilize blog contents	Analyze influence diffusion models
(Gruhl et al. 2004)	Yes	No	Yes	No
(Java et al. 2006)	Yes	No	No	Yes
(Java et al. 2007)	No	Yes	Yes	No
(Huang et al. 2016)	Yes	No	Yes	No

recommend feeds and identify influential blogs automatically, the same authors (Java et al. 2007) found "blog feeds that matter" for a specific topic using folder names and subscriber counts.

Huang et al. (2016) presented a framework that contains a heuristic quantification model for ranking key microbloggers. The two major parts in the framework were (1) based on content of posts, a classifying approach with sliding-window to specify interested domains of microbloggers, and (2) a method for quantifying key microbloggers by taking both the influence and user activity into consideration.

Overall, in the blogosphere domain, it is observed that a number of research, e.g., Gruhl et al. (2004), Java et al. (2007), and Huang et al. (2016), utilize the contents of blogs to identify top-k nodes in the weblog networks. We further compare the techniques discussed in this section in Table 7.

5.4 Misinformation Control

Despite the benefits of interconnectivity provided by online social networks, there are existing threats, such as spread of misinformation, that can cause undesirable effects, such as widespread panic to the general public. Budak et al. (2011) described the misinformation control problem as "identifying a subset of individuals that need to be convinced to adopt the good campaign so as to minimize the number of people that adopt the bad campaign" (Budak et al. 2011). In addition, they formulated their description as an optimization problem, proved that it was NP-hard, and then provided performance guarantees for a greedy strategy.

From a different aspect of controlling misinformation, Nguyen et al. (2012) focused on how to limit rumor spread in OSNs by finding the smallest set of influential nodes whose decontamination with good information helps to control the viral propagation of misinformation. Their solution includes a greedy-based algorithm, *Greedy Viral Stopper* (GVS), and a community-based heuristic method. GVS greedily adds nodes with the best influence gain to the current solution and shows that the algorithm selects a small fraction of the total nodes from the optimal solution. The community-based method returns a selection of nodes to decontaminate in a timely manner. The authors verified their approaches on real-world OSNs such as Facebook.

There are two major differences in the two aforementioned research by Budak et al. (2011) and Nguyen et al. (2012): (1) The approach by Budak et al. was limited by a k-nodes budget where k is an constraint imposed by the user. This was not presented in Nguyen et al.'s approach, and (2) the heuristic proposed by Budak et al. assumes a high level of propagation, where the probability for good information spread is either 1 or zero. In contrast, Nguyen et al. accounts for arbitrary probabilities.

5.5 Community Question Answering

Top-k nodes identification has also been researched widely in the Community Question Answering (CQA) domain with a number of models proposed. Users can often find detailed information from subject-matter experts on Q&A portals such as "Stack Overflow" and "Yahoo! Answers."

However, the quality of information can range from detailed solutions to unconstructive criticism. Additionally, if the feedback for a topic is sparse, it can be difficult to differentiate the different quality of answers. Hence, the ability for users to identify the members that provide reliable and accurate answers is critical for the CQA domain.

Fisher et al. (2006) detects key authors ("answer people") in Usenet newsgroups. Their heuristic network analysis methodology uses nodal features to find "answer people" with high out-degree and low in-degree. Zhang et al. (2007) models CQA as an expertise graph and evaluates a number of ranking measures and algorithms for identifying users with high expertise in differently structured networks. Extended from Zhang's approach (Zhang et al. 2007), Guo et al. (2008) proposed topic-based models to identify appropriate users to answer a specific question. Jurczyk and Agichtein (2007a) presented an analysis of the link structure of a general-purpose Q&A community to identify authoritative users. Agichtein et al. (2008) uses data from a web-scale community question answering portal and extracts graph-based features, e.g., the degree distribution of users and their PageRank, and hub scores to determine individual user's relative importance. A model to identify authoritative actors based on the number of best answers provided by users is presented in Bouguessa et al. (2008), whereas Pal et al. (2010) discriminates experts on top of their preference in answering position.

5.6 Networks with Complex Topologies

The majority of research described to now performs analysis on unweighted networks with very basic properties and node-to-node relationships. However, these simple network structures will often obfuscate many important properties of a relationship such as intensity, duration, and human emotional factors and, therefore, are unable to characterize the complexity of real-world networks and relationships. Barrat et al. (2004) describe weighted networks as graphs that go beyond the topological point and underline the capacity and intensity of connections between complex entities.

Opsahl et al. (2010) presents an approach on evaluating node centrality in weighted networks. In this research, the authors proposed a generalized degree centrality measure that incorporates both the number of edges and their weights on Freeman's Electronic Information Exchange System (EIES) network (1978). As it is difficult to non-subjectively define a weight for each of these properties, the authors' analysis contains a proposed threshold value, a, that alters the weighting of either attribute. The non-deterministic and complex nature of weighted networks can also be seen in similar research by Wei et al. (2013). Also using Freemans EIES dataset, the authors performed centrality measures based on DempsterShafer's theory of evidence (Dempster 1967; Shafer 1976), using the degree and strength of the node. To evaluate the effects of different weightings of an edge, the authors also applied small thresholds to their centrality measures.

Another complex network topology that we are underlining is the homogeneity of social networks for top-k node identification. A homogeneous network has a singular type of object and relationship, which are represented as nodes and edges, respectively, whereas heterogeneous networks are those with multiple types of nodes and node-to-node relationships. The vast majority of the research in this survey has predominately been on homogeneous networks, such as in Kempe et al. (2003), Wang et al. (2010), and Weng et al. (2010), this is likely due to homogeneous network data being more readily available and easier to analyze. However, there is very little existing research that identifyies top-k nodes in heterogeneous networks. One example presented in this survey is the research by Zhou et al. (2007). The requirement to co-rank two different types of entities significantly increases the complexity of their heuristic, which involves coupling two random walks to rank authors with their respective publications in the heterogeneous network.

5.7 Miscellaneous Applications

Ghosh and Lerman (2010) are among the few researchers who study the problem of predicting influential users in online social networks. They classified influence diffusion process as non-conservative if it depends not only on the network structure but also on details of the dynamic processes occurring on it. The authors experimented with the social news aggregator, Digg, which allows users to post and vote for news stories. In their scenario, influence was defined as the dynamic number of in-network votes a user's post generates, which represents non-conservative information flow. A number of influence models were compared, and their experiment results showed that non-conservative models that capture the actual details of the dynamic process are better at predicting influential users on Digg. In addition, Ghoah and Lerman's experiments also found that the normalized α-centrality metric is one of the best predictors of influential users.

Aral and Walker (2012) conducted a randomized experiment that identifies influence and susceptibility to influence in the product adoption decisions of a representative sample of 1.3 million Facebook users. The experiment includes "a random manipulation of influence-mediating messages sent from a commercial Facebook application" (Aral and Walker 2012), which allows users to share information about various products. Their methodology avoids the biases inherent in traditional estimates of social contagion by randomly manipulating who receives the influence-mediating messages. Some interesting findings from their experiments are as follows: (1) younger users tend to be more susceptible to influence than older users; (2) influential users are less susceptible to influence than non-influential ones and they cluster in the network, while susceptible users do not; and (3) unlike some previous research, which claim that an individual's influence is determined only by his or her personal attributes, Aral and Walker's experiment results showed that the combination of influence, susceptibility, and the likelihood of spontaneous adoption contributes to an individual's importance to the diffusion of behaviors.

5.8 Section Summary

This section presents examples of real-world applications integrating top-k node detection techniques. Several techniques that analyze OSNs such as Facebook and Twitter have been reviewed, including those targeting social dynamics such as misinformation control. We have also reflected on work that targets weighted and heterogeneous networks and underlined the increased complexity of identifying top-k nodes in these kinds of networks when compared to traditional networks.

Emphasis on node and edge attributes can be seen throughout the research in this section. (1) Research on Twitter, including Tunkelang (2009) and Weng et al. (2010), all contain node attributes, such as the number of followers and retweets. (2) Community Question Answering datasets (Fisher et al. 2006; Agichtein et al. 2008) contain integrated user voting and scoring systems. (3) Networks with complex heterogeneous structures (Zhou et al. 2007) contain multiple types of nodes and edges, resulting in multi-layered analysis of different node and edge types.

6 RESEARCH DIRECTIONS

With the abundant amount of literature published in research into identifying top-k nodes, one may wonder whether we have solved most of the critical problems related to top-k nodes identification such that the solutions provided are refined enough for most of the social network analysis tasks. However, in our view, there are still several critical research problems that need to be solved before top-k nodes identification can become a sufficient tool for analyzing social networks.

Exploration of factors that create the top k-nodes. So far, very little literature has focused on the exploration of factors that establish the top k-nodes being the most significant and/or influential, and this can be an interesting research area. In a situation where a social network comes with

some known top-k nodes, the question is as follows: What are the factors that distinguish those top nodes from others? Understanding these factors is essential to improve interpretation and usability of top-k nodes mining.

Devotion of more systematic research effort on top-k significant nodes. In general, more progress has been made on the top-k influential nodes as compared to the top-k significant nodes. In-depth research can be conducted in the future to investigate the top-k significant nodes in various domains. With the proliferation of social networks, i.e., email communication networks, user interactive question answering networks, organization hierarchies, OSNs, and so on, more intelligent and practical solutions can be applied to the top-k significant nodes, e.g., maximizing the quality of identified top-k significant nodes by capturing and utilizing individuals' skill sets and their social interactions as well as user influence.

Identification of top-k nodes in dynamic social networks. In Section 3.5, we pointed out that identifying top-k nodes in a dynamic social network is a relatively new area. Since many social networks, such as Twitter, are only available in the form of social streams of user activities, there is an increasing value of research in identifying top-k nodes in dynamic social networks.

Identification of top-k nodes for multiple topics. Based on our review, there has been very little research into the problem of identifying top-k nodes for multiple topics. The vast majority of the existing literature in the field identifies top-k nodes only for a single topic. However, it is more practical and useful to perform top-k nodes mining for multiple topics, especially for companies and organizations with heterogeneous social networks.

Enhancement of searching variety for a single topic. When dealing with top-k nodes, we are interested in the subjectively determined topic. A wide range of searching capabilities can be utilized to identify top-k nodes for a given topic. For example, when searching for top researchers of a particular subject in an area, it is useful if the functionality of searching on subject name is integrated within the algorithm.

Development of more efficient, scalable and performance guaranteed algorithms for top-k influential nodes. As we have categorized and summarized in Section 3, abundant research has been dedicated to generic algorithm development for top-k influential nodes. A number of these research evaluated their approaches in large-scale social networks using metrics such as time consumption, memory usage, and true positive rate. Therefore, one possible future direction is to focus on efficiency, scalability, and performance guarantee improvement of such generic algorithms. This direction is also described in a literature survey (Probst et al. 2013) on finding influential users as "*efficiency and validity are crucial.*"

7 CONCLUSIONS

In this article, we present an overview of the current status and future directions of top-k nodes identification in social networks. We reviewed and classified the existing literature in this area into two major categories: top-k influential nodes and top-k significant nodes. In general, we feel that considerable progress has been made on the top-k influential nodes as compared to top-k significant nodes. The applications of top-k nodes identification in social networks are quite diverse and have been discussed in detail.

This survey aims to show that numerous research works have attempted to solve the top-k nodes identification problem by proposing various algorithms, methodologies, and frameworks. An interesting scope exists in future research targeting top-k significant nodes, since the work in this area is quite limited. Furthermore, the vast majority of the work on identifying influential nodes is designed for static networks, and, therefore, the research area of mining top-k nodes in dynamic networks is still relatively new.

ACKNOWLEDGMENTS

The authors sincerely thank the anonymous reviewers of the previous versions of this survey for their comments that helped to improve this manuscript, the New Zealand Callaghan Innovation and Pingar for funding the R&D Student Fellowship Grant, and Tobey Sheng-Yen Hung for providing useful suggestions and feedback on the article.

REFERENCES

Charu Aggarwal and Karthik Subbian. 2014. Evolutionary network analysis: A survey. *ACM Comput. Surv.* 47, 1, Article 10 (2014), 36 pages. DOI: http://dx.doi.org/10.1145/2601412

Charu C. Aggarwal, Shuyang Lin, and Philip S. Yu. 2012. On influential node discovery in dynamic social networks. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. 636–647. DOI: http://dx.doi.org/10.1137/1.9781611972825.55

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*. ACM, New York, NY, 183–194. DOI: http://dx.doi.org/10.1145/1341531.1341557

Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341. DOI: http://dx.doi.org/10.1126/science.1215842

Çigdem Aslay, Nicola Barbieri, Francesco Bonchi, and Ricardo A Baeza-Yates. 2014. Online topic-aware influence maximization queries. In *Proceedings of the 17th International Conference on Extending Database Technology (EDBT'14)*. 295–306.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. 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. DOI: http://dx.doi.org/10.1145/1935826.1935845

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Topic-aware social influence propagation models. *Knowl. Inf. Syst.* 37, 3 (2013), 555–584. DOI: http://dx.doi.org/10.1007/s10115-013-0646-6

Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 2004. Modeling the evolution of weighted networks. *Phys. Rev. E* 70, 6 (2004), 066149.

Phillip Bonacich. 1987. Power and centrality: A family of measures. *Am. J. Sociol.* 92, 5 (1987), 1170–1182.

Stephen P. Borgatti. 2006. Identifying sets of key players in a social network. *Comput. Math. Org. Theory* 12, 1 (2006), 21–34.

Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. 2008. Identifying authoritative actors in question-answering forums: The case of yahoo! answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, NY, 866–874. DOI: http://dx.doi.org/10.1145/1401890.1401994

Arastoo Bozorgi, Hassan Haghighi, Mohammad Sadegh Zahedi, and Mojtaba Rezvani. 2016. INCIM: A community-based algorithm for influence maximization problem under the linear threshold model. *Inf. Process. Manage.* 52, 6 (2016), 1188–1199. DOI: http://dx.doi.org/10.1016/j.ipm.2016.05.006

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 1-7 (1998), 107–117.

Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. ACM, New York, NY, 665–674. DOI: http://dx.doi.org/10.1145/1963405.1963499

R. Cappelletti and N. Sastry. 2012. IARank: Ranking users on twitter in near real-time, based on their information amplification potential. In *Proceedings of the 2012 International Conference on Social Informatics*. 70–77. DOI: http://dx.doi.org/10.1109/SocialInformatics.2012.82

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'10)*.

Duanbing Chen, Linyuan Lu, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. 2012. Identifying influential nodes in complex networks. *Physica A* 391, 4 (2012), 1777–1787. DOI: http://dx.doi.org/10.1016/j.physa.2011.09.017

Shubo Chen and Kejing He. 2015. Influence maximization on signed social networks with integrated pagerank. In *Proceedings of the 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity'15)*. 289–292. DOI: http://dx.doi.org/10.1109/SmartCity.2015.86

Wei Chen, Chi Wang, and Yajun Wang. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. ACM, New York, NY, 1029–1038. DOI: http://dx.doi.org/10.1145/1835804.1835934

Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*. ACM, New York, NY, 199–208. DOI: http://dx.doi.org/10.1145/1557019.1557047

W. Chen, Y. Yuan, and L. Zhang. 2010. Scalable influence maximization in social networks under the linear threshold model. In *Proceedings of the 2010 IEEE International Conference on Data Mining*. 88–97. DOI: http://dx.doi.org/10.1109/ICDM.2010.118

Arthur P. Dempster. 1967. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* 38, 2 (1967), 325–339. Retrieved from http://www.jstor.org/stable/2239146.

Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. ACM, New York, NY, 57–66. DOI: http://dx.doi.org/10.1145/502512.502525

Georgios Drakopoulos, Andreas Kanavos, and Athanasios Tsakalidis. 2016. Evaluating twitter influence ranking with system theory. In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST'16)*.

Facebook. 2017. Statistics. (2017). Retrieved from http://www.facebook.com/press/info.php?statistics.

Ayman Farahat, Geoff Nunberg, and Francine Chen. 2002. AuGEAS: Authoritativeness grading, estimation, and sorting. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02)*. ACM, New York, NY, 194–202. DOI: http://dx.doi.org/10.1145/584792.584827

D. Fisher, M. Smith, and H. T. Welser. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 3. 59b–59b. DOI: http://dx.doi.org/10.1109/HICSS.2006.536

Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. *Soc. Netw.* 1, 3 (1978), 215–239.

Rumi Ghosh and Kristina Lerman. 2010. Predicting influential users in online social networks. In *Proceedings of the Fourth SNA-KDD Workshop*. http://arxiv.org/abs/1005.4882.

Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Market. Lett.* 12, 3 (2001), 211–223. DOI: http://dx.doi.org/10.1023/A:1011122126881

Mark Granovetter. 1987. Threshold models of collective behavior. *Am. J. Sociol.* 83, 6 (1987), 1420–1443. http://www.jstor.org/stable/2778111.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web (WWW'04)*. ACM, New York, NY, 491–501. DOI: http://dx.doi.org/10.1145/988672.988739

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. 2013. Information diffusion in online social networks: A survey. *SIGMOD Rec.* 42, 2 (2013), 17–28. DOI: http://dx.doi.org/10.1145/2503792.2503797

Jinwen Guo, Shengliang Xu, Shenghua Bao, and Yong Yu. 2008. Tapping on the potential of q&a community by recommending answer providers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, New York, NY, 921–930. DOI: http://dx.doi.org/10.1145/1458082.1458204

Meng Han, Mingyuan Yan, Zhipeng Cai, Yingshu Li, Xingquan Cai, and Jiguo Yu. 2017. Influence maximization by probing partial communities in dynamic online social networks. *Trans. Emerg. Telecommun. Technol.* 28, 4 (2017), e3054–n/a. DOI: http://dx.doi.org/10.1002/ett.3054 e3054 ett.3054.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*. ACM, New York, NY, 517–526. DOI: http://dx.doi.org/10.1145/511446.511513

Julia Heidemann, Mathias Klier, and Florian Probst. 2010. Identifying key users in online social networks: A pagerank based approach. In *Proceedings of the International Conference on Information Systems*.

Lidong Huang, Wenxian Wang, Xiaozhou Chen, and Junhua Chen. 2016. *A Heuristic Method of Identifying Key Microbloggers*. Springer International Publishing, Cham, 417–426. DOI: http://dx.doi.org/10.1007/978-3-319-39601-9_37

Muhammad U. Ilyas and Hayder Radha. 2011. Identifying influential nodes in online social networks using principal component centrality. In *Proceedings of the 2011 IEEE International Conference on Communications (ICC'11)*. 1–5. DOI: http://dx.doi.org/10.1109/icc.2011.5963147

Akshay Java, Pranam Kolari, Tim Finin, Anupam Joshi, and Tim Oates. 2007. Feeds that matter: A study of bloglines subscriptions. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'07)*.

Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. 2006. Modeling the spread of influence on the blogosphere. In *Proceedings of the 15th International World Wide Web Conference*. 22–26.

Pawel Jurczyk and Eugene Agichtein. 2007a. Discovering authorities in question answer communities by using link analysis. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*. ACM, New York, NY, 919–922. DOI: http://dx.doi.org/10.1145/1321440.1321575

Pawel Jurczyk and Eugene Agichtein. 2007b. Hits on question answer portals: Exploration of link analysis for author ranking. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 845–846.

Jon Kleinberg Jure Leskovec, Daniel Huttenlocher. 2010. Wikipedia Vote Network. (2010). Retrieved from http://snap.stanford.edu/data/wiki-Vote.html.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*. ACM, New York, NY, 137–146. DOI: http://dx.doi.org/10.1145/956750.956769

David Kempe, Jon Kleinberg, and Éva Tardos. 2005. *Influential Nodes in a Diffusion Model for Social Networks*. Springer, Berlin, 1127–1138. DOI: http://dx.doi.org/10.1007/11523468_91

E. S. Kim and S. S. Han. 2009. An analytical way to find influencers on social networks and validate their effects in disseminating social games. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*. 41–46. DOI: http://dx.doi.org/10.1109/ASONAM.2009.59

Hyoungshick Kim and Eiko Yoneki. 2012. Influential neighbours selection for information diffusion in online social networks. In *Proceedings of the 2012 21st International Conference on Computer Communications and Networks (ICCCN'12)*. 1–7. DOI: http://dx.doi.org/10.1109/ICCCN.2012.6289230

Masahiro Kimura and Kazumi Saito. 2006. *Tractable Models for Information Diffusion in Social Networks*. Springer, Berlin, 259–271. DOI: http://dx.doi.org/10.1007/11871637_27

Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. 2007. Extracting influential nodes for information diffusion on a social network. In *AAAI*, Vol. 7. 1371–1376.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 5 (Sep. 1999), 604–632. DOI: http://dx.doi.org/10.1145/324133.324140

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue 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, 591–600. DOI: http://dx.doi.org/10.1145/1772690.1772751

Theodoros Lappas, Kun Liu, and Evimaria Terzi. 2011. *A Survey of Algorithms and Systems for Expert Location in Social Networks*. Springer US, Boston, MA, 215–241. DOI: http://dx.doi.org/10.1007/978-1-4419-8462-3_8

Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. 2010. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. ACM, New York, NY, 1059–1068. DOI: http://dx.doi.org/10.1145/1835804.1835937

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*. ACM, New York, NY, 420–429. DOI: http://dx.doi.org/10.1145/1281192.1281239

Hongbo Li, Jianpei Zhang, Jing Yang, Jinbo Bai, and Yuming Zhao. 2017. Efficient star nodes discovery algorithm in social networks based on acquaintance immunization. In *Proceedings of the 2017 2nd International Conference on Image, Vision and Computing (ICIVC'17)*. IEEE, 937–940.

Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, New York, NY, 199–208. DOI: http://dx.doi.org/10.1145/1871437.1871467

Qi Liu, Biao Xiang, Enhong Chen, Hui Xiong, Fangshuang Tang, and Jeffrey Xu Yu. 2014. Influence maximization over large-scale social networks: A bounded linear approach. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14)*. ACM, New York, NY, 171–180. DOI: http://dx.doi.org/10.1145/2661829.2662009

Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. 2005. Co-authorship networks in the digital library research community. *Inf. Process. Manage.* 41, 6 (2005), 1462–1480. DOI: http://dx.doi.org/10.1016/j.ipm.2005.03.012 Special Issue on Infometrics.

Joel C. Miller, Gregory Rae, Fred Schaefer, Lesley A. Ward, Thomas LoFaro, and Ayman Farahat. 2001. Modifications of kleinberg's HITS algorithm using matrix exponentiation and web log records. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 444–445.

Katarzyna Musiał and Krzysztof Juszczyszyn. 2009. Properties of bridge nodes in social networks. *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems* (2009), 357–364.

Ramasuri Narayanam and Yadati Narahari. 2011. A shapley value-based approach to discover influential nodes in social networks. *IEEE Trans. Autom. Sci. Eng.* 8, 1 (Jan. 2011), 130–147. DOI: http://dx.doi.org/10.1109/TASE.2010.2052042

Mario A. Nascimento, Jorg Sander, and Jeffrey Pound. 2003. Analysis of SIGMOD's co-authorship graph. *SIGMOD Rec.* 32, 3 (Sep. 2003), 8–10. DOI: http://dx.doi.org/10.1145/945721.945722

Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference (WebSci'12)*. ACM, New York, NY, 213–222. DOI: http://dx.doi.org/10.1145/2380718.2380746

Tore Opsahl, Filip Agneessens, and John Skvoretz. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Netw.* 32, 3 (2010), 245–251.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford InfoLab.

Aditya Pal and Scott Counts. 2011. 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. DOI: http://dx.doi.org/10.1145/1935826.1935843

Aditya Pal and Joseph A. Konstan. 2010. Expert identification in community question answering: Exploring question selection bias. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, New York, NY, 1505–1508. DOI: http://dx.doi.org/10.1145/1871437.1871658

Florian Probst, Laura Grosswiele, and Regina Pfleger. 2013. Who will lead and who will follow: Identifying influential users in online social networks. *Bus. Inf. Syst. Eng.* 5, 3 (2013), 179–193. DOI: http://dx.doi.org/10.1007/s12599-013-0263-7

Fabián Riquelme and Pablo González-Cantergiani. 2016. Measuring user influence on twitter: A survey. *Inf. Process. Manage.* 52, 5 (2016), 949–975. DOI: http://dx.doi.org/https://doi.org/10.1016/j.ipm.2016.04.003

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Part III*. Springer, Berlin, 18–33. DOI: http://dx.doi.org/10.1007/978-3-642-23808-6_2

Glenn Shafer. 1976. *A Mathematical Theory of Evidence*. Vol. 1. Princeton University Press.

L. S. Shapley. 1950. *A Value for N-person Games in Contributions to the Theory of Games*. Vol. 1. Princeton University Press.

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 (Feb. 2017), 359–372. DOI: http://dx.doi.org/10.1109/TKDE.2016.2620141

Karthik Subbian, Charu Aggarwal, and Jaideep Srivastava. 2013. Content-centric flow mining for influence analysis in social streams. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13)*. ACM, New York, NY, 841–846. DOI: http://dx.doi.org/10.1145/2505515.2505626

Karthik Subbian, Charu Aggarwal, and Jaideep Srivastava. 2016a. Mining influencers using information flows in social streams. *ACM Trans. Knowl. Discov. Data* 10, 3, Article 26 (2016), 26:1–26:28 pages. DOI: http://dx.doi.org/10.1145/2815625

Karthik Subbian, Charu C. Aggarwal, and Jaideep Srivastava. 2016b. Querying and tracking influencers in social streams. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM'16)*. ACM, New York, NY, 493–502. DOI: http://dx.doi.org/10.1145/2835776.2835788

Karthik Subbian, Dhruv Sharma, Zhen Wen, and Jaideep Srivastava. 2014. Finding influencers in networks using social capital. *Soc. Netw. Anal. Min.* 4, 1 (2014), 1–13. DOI: http://dx.doi.org/10.1007/s13278-014-0219-z

Jimeng Sun and Jie Tang. 2011. *A Survey of Models and Algorithms for Social Influence Analysis*. Springer US, Boston, MA, 177–214. DOI: http://dx.doi.org/10.1007/978-1-4419-8462-3_7

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*. ACM, New York, NY, 807–816. DOI: http://dx.doi.org/10.1145/1557019.1557108

Guangmo Tong, Weili Wu, Shaojie Tang, and Ding-Zhu Du. 2017. Adaptive influence maximization in dynamic social networks. *IEEE/ACM Trans. Netw.* 25, 1 (Feb. 2017), 112–125. DOI: http://dx.doi.org/10.1109/TNET.2016.2563397

Chi-Yao Tseng and Ming-Syan Chen. 2012. *Significant Node Identification in Social Networks*. Springer Berlin, 459–470. DOI: http://dx.doi.org/10.1007/978-3-642-28320-8_39

Daniel Tunkelang. 2009. A twitter analog to pagerank. *The Noisy Channel*.

M. Vadoodparast and F. Taghiyareh. 2015. A multi-agent solution to maximizing product adoption in dynamic social networks. In *Proceedings of the 2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP'15)*. 71–78. DOI: http://dx.doi.org/10.1109/AISP.2015.7123484

Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. 2010. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. ACM, New York, NY, 1039–1048. DOI: http://dx.doi.org/10.1145/1835804.1835935

Yanhao Wang, Qi Fan, Yuchen Li, and Kian-Lee Tan. 2017. Real-time influence maximization on dynamic social streams. *Proc. VLDB Endow.* 10, 7 (Mar. 2017), 805–816. DOI: http://dx.doi.org/10.14778/3067421.3067429

Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge University Press.

Daijun Wei, Xinyang Deng, Xiaoge Zhang, Yong Deng, and Sankaran Mahadevan. 2013. Identifying influential nodes in weighted networks based on evidence theory. *Physica A* 392, 10 (2013), 2564–2575. DOI: http://dx.doi.org/10.1016/j.physa.2013.01.054

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi 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. DOI: http://dx.doi.org/10.1145/1718487.1718520

Harris Wu, Mohammad Zubair, and Kurt Maly. 2006. Harvesting social knowledge from folksonomies. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*. ACM, 111–114.

Osmar R. Zaïane, Jiyang Chen, and Randy Goebel. 2009. *Mining Research Communities in Bibliographical Data*. Springer, Berlin, 59–76. DOI: http://dx.doi.org/10.1007/978-3-642-00528-2_4

Jun Zhang, Mark S. Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, NY, 221–230. DOI: http://dx.doi.org/10.1145/1242572.1242603

Ding Zhou, Sergey A. Orshanskiy, Hongyuan Zha, and C. Lee Giles. 2007. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM'07)*. 739–744. DOI: http://dx.doi.org/10.1109/ICDM.2007.57

Jingyu Zhou, Yunlong Zhang, and Jia Cheng. 2014. Preference-based mining of top-k influential nodes in social networks. *Fut. Gen. Comput. Syst.* 31 (2014), 40–47. DOI: http://dx.doi.org/10.1016/j.future.2012.06.011 Special Section: Advances in Computer Supported Collaboration: Systems and Technologies.

Feng Zhu, Guanfeng Liu, Yan Wang, An Liu, Zhixu Li, Pengpeng Zhao, and Lei Li. 2015. A context-aware trust-oriented influencers finding in online social networks. In *Proceedings of the 2015 IEEE International Conference on Web Services (ICWS'15)*. 456–463. DOI: http://dx.doi.org/10.1109/ICWS.2015.67

Honglei Zhuang, Yihan Sun, Jie Tang, Jialin Zhang, and Xiaoming Sun. 2013. Influence maximization in dynamic social networks. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*. 1313–1318. DOI: http://dx.doi.org/10.1109/ICDM.2013.145

Received September 2016; revised December 2017; accepted November 2018