Knowledge and Information Systems (2020) 62:3417–3455
https://doi.org/10.1007/s10115-020-01461-4

REGULAR PAPER

A survey on influence maximization in a social network

Suman Banerjee[1] · Mamata Jenamani[2] · Dilip Kumar Pratihar[3]

Received: 24 December 2018 / Revised: 1 March 2020 / Accepted: 7 March 2020 / Published online: 29 March 2020

Abstract
Given a *social network* with *diffusion probabilities* as edge weights and a positive integer k, which k nodes should be chosen for initial injection of information to maximize the influence in the network? This problem is popularly known as the *Social Influence Maximization Problem* (*SIM Problem*). This is an active area of research in *computational social network analysis* domain, since one and half decades or so. Due to its practical importance in various domains, such as *viral marketing*, *target advertisement* and *personalized recommendation*, the problem has been studied in different variants, and different solution methodologies have been proposed over the years. This paper presents a survey on the progress in and around *SIM Problem*. At last, it discusses current research trends and future research directions as well.

Keywords Social networks · Influence maximization · Approximation algorithm · Greedy strategy

1 Introduction

A social network is an interconnected structure among a group of agents formed for social interactions [85]. Nowadays, social networks play an important role in spreading information, opinion, idea, innovation, rumor, etc., at a large scale [14,90]. This spreading process has a huge practical importance in viral marketing [20,79]. Consider the case of promoting a brand by of a commercial house through online marketing, where the goal is to attract

Major part of this was done when the first author was a Ph.D. student at IIT Kharagpur. This work is financially supported by the project E-Business Center of Excellence (F.No.5-5/2014-TS.VII).

✉ Suman Banerjee
suman.b@iitgn.ac.in

Mamata Jenamani
mj@iem.iitkgp.ac.in

Dilip Kumar Pratihar
dkpra@mech.iitkgp.ac.in

[1] Department of Computer Science and Engineering, IIT Gandhinagar, Gujarat, India

[2] Department of Industrial and Systems Engineering, IIT Kharagpur, Kharagpur, India

[3] Department of Mechanical Engineering, IIT Kharagpur, Kharagpur, India

the users for purchasing a particular product. The best way to do this is to select a set of highly influential users and distribute them free samples. Many of them will like the item and influence their neighbors to try the product. These newly informed users will influence their neighbors. This cascading process will be continued, and ultimately a large fraction of the users will try for the product leading to a significant improvement in the earned revenue. Naturally, the number of free sample products will be limited due to economic reason. Hence, this process will be fruitful, if the free samples can be distributed among the highly influential users and the problem here bottoms down to select influential users from the network. This problem is known as the *Social Influence Maximization Problem*. Due to the potential applications of this problem in different domains, such as personalized recommendation [112], feed ranking [60], target advertisement [83], selecting influential twitters [4,131], selecting informative blogs [80], etc., recent years have witnessed a significant attention in the study of *influence propagation and maximization* in online social networks.

Social influence occurs due to the diffusion of information in the network. This phenomenon in a networked system is well-studied [34,65]. Specifically, there are two popularly adopted models to study the diffusion process, namely *Independent Cascade Model* (abbreviated as *IC Model*), which collects the independent behavior of the agents, and the other one is *Linear Threshold Model* (abbreviated as *LT Model*), which captures the collective behavior of the agents (detailed discussion is deferred till Sect. 2.5) [110]. In both the models, information is diffused in discrete time steps from some initially identified nodes and continued for several rounds. In SIM Problem, our goal is to maximize influence by selecting appropriate seed nodes.

To study the SIM Problem, a social network is abstracted as a *graph* with the users as the *vertex set* and *social ties* among the users as the edge set. It is also assumed that the *diffusion threshold* (a measurement of how hard to influence the user and given in a numerical scale; more the value, more hard to influence the user) is given as the *vertex weight* and influence probability between two users as *edge weight*. In this settings, the SIM Problem is stated as follows: For a given positive integer k, choose the set of k nodes, such that initial activation of them leads to the maximum number of influenced nodes.

There are two surveys by Li et al. [84] and Peng et al. [100] published recently. In Li et al.'s [84] survey, the main focus is on the solution methodologies of the problem. In the same line, Peng et al.'s survey [100] is focused on the solution methodologies and also, its real-life applications. However, both these surveys largely ignored the two other important issues. First one is the variants of the SIM Problem studied in the literature and also the second one is their hardness results under traditional as well as parameterized complexity theoretic framework. To bridge this gap, in this paper along with the solution methodologies, we include the above-mentioned two issues, which are not covered in the previous surveys.

1.1 Focus and goal of the survey

In this survey, we have mainly focused on three aspects of the SIM Problem, as mentioned below.

- Variants of this problem studied in the literature,
- Hardness results of this problem in both traditional and parameterized complexity framework,
- Different solution approaches proposed in the literature.

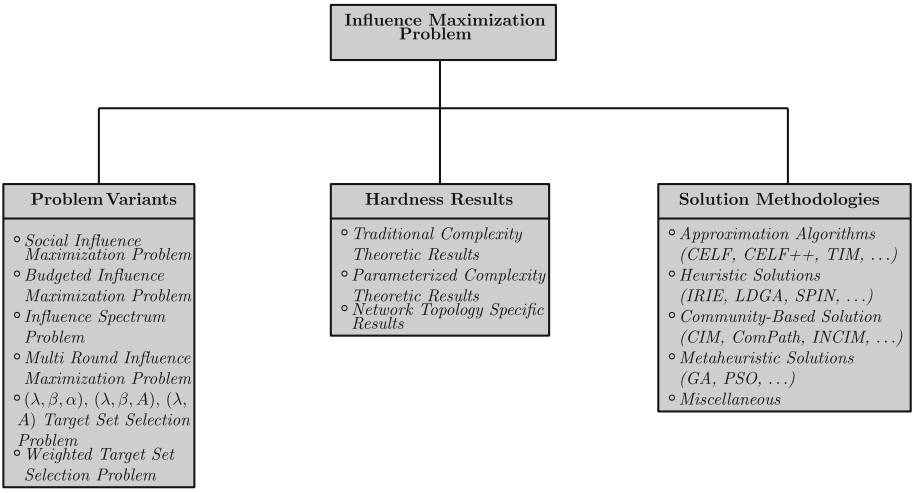

Fig. 1 Taxonomy of this survey

The overview of this survey is shown in Fig. 1. There are several other aspects of the problem, such as *SIM in the presence of adversaries*, *in a time-varying social network* and *in competitive scenario* which we have not considered in this survey.

The main goal of this survey is threefold:

- to provide comprehensive understanding about the SIM Problem and its different variants studied in the literature,
- to develop a taxonomy for classifying the existing solution methodologies and present them in a concise manner,
- to present an overview of the current research trend and future research directions regarding this problem.

We set the following two criteria for the studies to be included in this survey:

- Research work presented in the publication should produce theoretically or empirically better than some of the previously published results.
- The presented solution methodology should be generic, i.e., it should work for a network of any topology.

1.2 Organization of the survey

Rest of the paper is organized as follows: Sect. 2 describes some background material required to understand the subsequent sections of this paper. Section 3 formally introduces the SIM Problem and its variants studied in the literature. Section 4 describes the hardness results of this problem in both traditional and parameterized complexity theory framework. Section 5 describes some major research challenges in and around this problem. Section 6 describes the proposed taxonomy for classifying the existing solution methodologies in different categories and discuss them. Section 7 presents the summary of the survey and gives some future research directions. Finally, Sect. 8 presents concluding remarks regarding this survey.

Table 1 Symbols and notations

Symbols	Interpretation
$G(V, E, \theta, \mathcal{P})$	Directed, vertex and edge weighted social network
$V(G)$	Set of vertices of network G
$E(G)$	Set of edges of network G
U	Set of users of the network, i.e., $U = V(G)$
n	Number of users of the network, i.e., $n = \|V(G)\|$
m	Number of edges of the network, i.e., $m = \|E(G)\|$
θ	Vertex weight function of G, i.e., $\theta : V(G) \longrightarrow [0, 1]$
θ_i	Weight of vertex u_i, i.e., $\theta_i = \theta(u_i)$
$\mathcal{P}$	Edge weight function, i.e., $\mathcal{P} : E(G) \longrightarrow (0, 1]$
p_{ij}	Edge weight of the edge $(u_i u_j)$
$\mathcal{N}(u_i)$	Open neighborhood of vertex u_i
$\mathcal{N}[u_i]$	Closed neighborhood of vertex u_i
$[n]$	The set $\{1, 2, \ldots, n\}$
$\mathcal{N}^{in}(u_i)$	Incoming neighbors of vertex u_i
$\mathcal{N}^{out}(u_i)$	Outgoing neighbors of vertex u_i
$\deg^{in}(u_i)$	Indegree of vertex u_i
$\deg^{out}(u_i)$	Outdegree of vertex u_i
$\text{dist}(u, v)$	Number of edges in the shortest path between u and v
$\mathcal{S}$	Seed set for diffusion, i.e., $\mathcal{S} \subseteq V(G)$
k	Maximum allowable cardinality for the seed set, i.e., $\|\mathcal{S}\| \leq k$
r	Maximum allowable round for diffusion
$[m, n]$	The set $\{m, m+1, \ldots, n\}$ for $m \leq n$
$\mathbb{R}_0^+$	The set of positive real number including 0

2 Background

In this section, we have described relevant background topics up to required depth, such as *basic graph theory*, relation between SIM and existing graph theoretic problems, *approximation algorithm*, *parameterized complexity theory* and *information diffusion models* in social networks. The symbols and notations that have been used in the subsequent sections of this paper are given in Table 1.

2.1 Basic graph theory

Graphs are popularly used to represent most of the real-world networked systems including social networks [13,129]. Here, we have reported some preliminary concepts of *basic graph theory* from [36]. A graph is denoted by $G(V, E)$ where $V(G)$ and $E(G)$ are the *vertex set* and *edge set* of G, respectively. For any arbitrary vertex, $u_i \in V(G)$, its *open neighborhood* is defined as $\mathcal{N}(u_i) = \{u_j : (u_i u_j) \in E(G)\}$. *Closed neighborhood* of u_i will be $\mathcal{N}[u_i] = \{u_i\} \cup \mathcal{N}(u_i)$. *Degree* of a vertex is defined as the *cardinality* of its open neighborhood, i.e., $\deg(u_i) = |\mathcal{N}(u_i)|$. For any $S \subseteq V(G)$, its open neighborhood and close neighborhood will be $\mathcal{N}(S) = \{u_i \in V(G) \backslash S : \exists u_j \text{ and } (u_i u_j) \in E(G)\}$, and $\mathcal{N}[S] = S \cup \mathcal{N}(S)$, respectively.

Two vertices u_i and u_j are said to be *true twins*, if $\mathcal{N}[u_i] = \mathcal{N}[u_j]$ and *false twins*, if $\mathcal{N}(u_i) = \mathcal{N}(u_j)$. A graph is *weighted*, if real numbers are associated with its vertices or edges or both. In an undirected graph, an edge is an unordered pair of vertices, say $\{u, v\}$; however, in case of directed graph it is an ordered pair of vertices (u, v). The edges that join the same pair of vertices are known as parallel edges, and an edge whose both the end points are the same is known as *self-loop*. A graph is *simple*, if it is free from self-loop and parallel edges.

Information diffusion process in a *social network* is represented by a *simple*, *directed* and *vertex* and *edge weighted graph* $G(V, E, \theta, \mathcal{P})$. Here, $V(G) = \{u_1, u_2, \ldots, u_n\}$, the set of users of the network and $E(G) = \{e_1, e_2, \ldots, e_m\}$, the set of social ties among the users. θ and $\mathcal{P}$ are the *vertex* and *edge weight* functions, which assign numerical values in between 0 and 1 to each vertex and edge, respectively, as their weights, i.e., $\theta : V(G) \longrightarrow [0, 1]$ and $\mathcal{P} : E(G) \longrightarrow (0, 1]$. In *information diffusion*, vertex and edge weights are called node threshold and diffusion probability, respectively [56]. Larger value of θ_i becomes difficult to influence the user u_i. However, more the value of p_{ij}, it is more probable that u_i can influence u_j. For any user $u_i \in V(G)$, its *incoming neighbors* and *outgoing neighbors* $\mathcal{N}^{in}(u_i)$ and $\mathcal{N}^{out}(u_i)$ are defined as: $\mathcal{N}^{in}(u_i) = \{u_j | (u_j, u_i) \in E(G)\}$ and $\mathcal{N}^{out}(u_i) = \{u_j | (u_i, u_j) \in E(G)\}$, respectively. For any user $u_i \in V(G)$, its *indegree* and *outdegree* are defined as $\deg^{in}(u_i) = |\mathcal{N}^{in}(u_i)|$ and $\deg^{out}(u_i) = |\mathcal{N}^{out}(u_i)|$, respectively. A *path* in a directed graph is a sequence of vertices without repetition, such that between each consecutive vertex there will be an *edge*. Two users are connected in the graph G, if there exists a directed path between them. A directed graph is said to be connected, if there exists a path between each pair of users.

2.2 Relation between target set selection and other graph theoretic problems

The TSS Problem is a more generalized version of many standard graph theoretic problems discussed and mentioned in the literature, such as *dominating set with threshold* [58], *vector domination problem* [103], *k-tuple dominating set* [74] (in all these problems instead of multiple rounds, diffusion can run only for one round), *vertex cover* [17] (in this problem, vertex threshold is set equal to the number of neighbors of the node), *irreversible k-conversion problem* [41], *r-neighbor bootstrap percolation problem* [5] (where the threshold of each vertex is k or r, respectively) and *dynamic monopolies* [99]. (In this case, threshold is half of the neighbors of the user).

2.3 Approximation algorithm

Most of the *optimization problems* arising in real life are NP-hard [48]. Hence, we cannot expect to solve them by any deterministic algorithm in polynomial time. So, the goal is to get an approximate solution of the problem within affordable time. Approximation algorithms serve this purpose and also provide the worst-case guarantee on solution quality. For a maximization problem $\mathcal{P}$, let $\mathcal{A}$ be an algorithm, which provides its solution and $\mathcal{I}$ be the set of all possible input instances of $\mathcal{P}$. For an input instance I of $\mathcal{P}$, let $\mathcal{A}^*(I)$ be the optimal solution and $\mathcal{A}(I)$ be the solution generated by the algorithm $\mathcal{A}$. Now, $\mathcal{A}$ will be called an α-factor *absolute approximation algorithm*, if $\forall I \in \mathcal{I}$, $|\mathcal{A}^*(I) - \mathcal{A}(I)| \leq \alpha$ and α-factor *relative approximation algorithm*, if $\forall I \in \mathcal{I}, max\{\frac{\mathcal{A}^*(I)}{\mathcal{A}(I)}, \frac{\mathcal{A}(I)}{\mathcal{A}^*(I)}\} \leq \alpha$ $(\mathcal{A}(I), \mathcal{A}^*(I) \neq 0)$ [133]. Section 6.1 of this paper describes relative approximation algorithms for solving SIM Problem.

2.4 Parameterized complexity theory

Parameterized complexity theory is another way of dealing with NP-hard optimization problems [38]. It aims to classify computational problems based on the inherent difficulty with respect to multiple parameters related to the problem. There are several *complexity classes* in parameterized complexity theory. The class *Fixed- Parameter Tractable* (FPT) contains the problems, for which any problem with instances $(x, k) \in \mathcal{I}$, where x is the input, k is the parameter and $\mathcal{I}$ is the set of instances; its running time will be of $\mathcal{O}(f(k)|x|^{\mathcal{O}(1)})$, where $f(k)$ is the function depending on k only and $|x|$ denotes the length of the input. W hierarchy is the collection of complexity classes with the property $W[0] = \text{FPT}$ and $W[i] \subseteq W[j]$ for each $i \leq j$ [40]. Many normal computational problems occupy the lower levels of hierarchy, i.e., $W[1]$ and $W[2]$. In Sect. 4, we have described the hardness results of TSS Problem in parameterized complexity theoretic setting.

2.5 Information diffusion in a social network

Diffusion phenomenon in a networked system has got attention from different disciplines, such as *epidemiology* (how does disease spread in a human contact network?) [108], *social network analysis* (how does information propagate in a social network?) [138], *computer network* (how does computer virus propagate in an e-mail network?) [150], etc., *Information Diffusion* in an on-line social networks is a phenomenon by which word-of-mouth effect occurs electronically. Hence, the mechanism of information diffusion is very well-studied [72, 123]. To study the diffusion process, there are some models in the literature [59]. The nature of these models varies from *deterministic* to *probabilistic*. Here, we have described some well-studied *information diffusion models* from the literature.

– *Independent Cascade Model* (IC Model) [110] This is one of the well-studied probabilistic diffusion models used by Kempe et al. [67] in their seminal work of *social influence maximization*. In this model, a node can either be in active state (i.e., influenced) or in inactive state (i.e., not influenced). Initially (i.e., at $t = 0$), all the nodes except the seeds are inactive. Each active node (say, u_i) at time stamp t will get a chance to activate its currently *inactive* neighbor ($u_j \in \mathcal{N}^{out}(u_i)$ and u_j is inactive) with probability as their edge weight. If u_i succeeds, then u_j will become an active node in time stamp $t + 1$. A node can change its state from inactive to active but not from active to inactive. This cascading process continues until no more active node is there in a time stamp. Suppose, this diffusion process starts at $t = 0$ and continued till $t = \mathcal{T}$, and $\mathcal{A}_t$ denotes the set of active nodes till time stamp t, where $t \in [0, \mathcal{T}]$, then

$$\mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_t \subseteq \mathcal{A}_{t+1} \subseteq \cdots \subseteq \mathcal{A}_{\mathcal{T}} \subseteq V(\mathcal{G}).$$

Node u_i is said to be active at time stamp t, if $u_i \in \mathcal{A}_t \backslash \mathcal{A}_{t-1}$. $\mathcal{A}_0$ denotes the set of seed nodes, which is also represented by $\mathcal{S}$ in the literature. $\sigma(\mathcal{S})$ returns the set of influenced nodes, when the diffusion starts from the nodes in $\mathcal{S}$. Here, $\sigma(.)$ is the *social influence function* that assigns each subset of the nodes to its expected influence, i.e., $\sigma : 2^{V(G)} \longrightarrow \mathbb{R}_0^+$.

– *Linear Threshold Model* (*LT Model*) [110] This is another probabilistic diffusion model proposed by Kempe et al. [67]. In this model, for any node (say u_i), all its neighbors, who are activated just at the previous time stamp together make a try to activate that node. This activation process will be successful, if the sum of the incoming active neighbor's probability becomes either greater than or equal to the node's threshold, i.e., for all

$u_j \in \mathcal{N}^{in}(u_i)$, if $\sum_{u_j \in \mathcal{N}^{in}(u_i); u_j \in \mathcal{A}_t} p_{ji} \geq \theta_i$, then u_i will become active at time stamp $t+1$. This method will be continued until no more activation is possible. In this model, we can use the negative influence, which is not possible in IC Model. Later, several extensions of these two fundamental models have been proposed [139].
In both IC Model and LT Model, it is assumed that diffusion probability between two users is known. However, later there were several studies for computing diffusion probability [52,73,105–107].

- *Shortest Path Model* (*SP Model*) This is a special case of IC Model proposed by Kimura et al. [72]. In this model, an inactive node will get a chance to become active only through the shortest path from the initially active nodes, i.e., at $t = \min_{u \in \mathcal{A}_0, v \in V(G) \setminus \mathcal{A}_0} \{\text{dist}(u, v)\}$. A slightly different variation of SP Model proposed by the same author is *SP1 Model*, which tells that an inactive node will get a chance of activation at $t = \min_{u \in \mathcal{A}_0, v \in V(G) \setminus \mathcal{A}_0} \text{dist}(u, v)$ and $t = \min_{u \in \mathcal{A}_0, v \in V(G) \setminus \mathcal{A}_0} \text{dist}(u, v) + 1$.
- *Majority Threshold Model* (*MT Model*) This is the deterministic threshold model proposed by Valente [124]. In this model, the vertex threshold is defined as $\theta_i = \left\lceil \frac{\deg(u_i)}{2} \right\rceil$, which means that a node will become active, when at least half of its neighbors are already active in nature.
- *Constant Threshold Model* (*CT Model*) This is another deterministic diffusion model, where vertex threshold can be any value from 1 to its degree, i.e., $\theta_i \in [\deg(u_i)]$.
- *Unanimous Threshold Model* (*UT Model*) [17] This is the most influence resistant model of diffusion. In this model, for each node in the network, its threshold value is set to its degree, i.e., $\forall u_i \in V(G), \theta_i = \deg(u_i)$.

There are many other diffusion models, such as *weighted cascade model*, where edge weight will be the reciprocal of the degree of the node; *trivalency model*, where the edge weights are uniformly taken from the set: $\{0.1, 0.01, 0.001\}$, etc. Readers requiring a detailed and exhaustive treatment on information diffusion models may refer to [144].

3 SIM Problem and its variants

In the literature, SIM Problem has been studied since early two thousand. Initially, this problem was introduced by Domingos and Richardson in the context of viral marketing [37]. Due to its substantial practical importance across multiple domains, different variants of this problem have been introduced. In this section, we describe them one by one.

Basic SIM Problem [1] In the basic version of the *TSS Problem* along with a *directed social network* $G(V, E, \theta, \mathcal{P})$, we are given two integers: k and λ and asked to find out a subset of at most k nodes, such that after the diffusion process is over at least λ number of nodes are activated. Mathematically, this problem can be stated as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$, $\lambda \in [n]$ *and* $k \in \mathbb{Z}^+$.
Problem: *Basic TSS Problem [Find out an* $\mathcal{S} \subseteq V(G)$, *such that* $|\mathcal{S}| \leq k$, *and* $|\sigma(\mathcal{S})| \geq \lambda$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *and* $|\mathcal{S}| \leq k$.

Top k-node Problem/Social Influence Maximization Problem (SIM Problem) [89] This variant of the problem is most well-studied. For a given social network $G(V, E, \theta, \mathcal{P})$, this problem asks to choose a set $\mathcal{S}$ of k nodes (i.e., $\mathcal{S} \subseteq V(G)$ and $|\mathcal{S}| = k$), such that the maximum number of nodes of the network become influenced at the end of diffusion process, i.e., $\sigma(\mathcal{S})$ will be maximized. Most of the algorithms presented in Sect. 6 are solely developed for solving this problem. Mathematically, the *Problem of Top k-node Selection* will be like the following:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$ *and* $k \in \mathbb{Z}^+$.
Problem: *Top k-node Problem [Find out a* $\mathcal{S} \subseteq V(G)$ *where* $|\mathcal{S}| = k$ *such that and for any other* $\mathcal{S}' \subseteq V(G)$ *with* $|\mathcal{S}'| = k$, $\sigma(\mathcal{S}) \geq \sigma(\mathcal{S}')$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *and* $|\mathcal{S}| = k$.

Influence Spectrum Problem [95] In this problem, along with the social network $G(V, E, \theta, \mathcal{P})$, we are also given with two integers: k_{lower} and k_{upper} with $k_{upper} > k_{lower}$. Our goal is to choose a set $\mathcal{S}$ for each $k \in [k_{lower}, k_{upper}]$, such that social influence in the network ($\sigma(\mathcal{S})$) is maximum in each case. Intuitively, solving one instance of this problem is equivalent to solving $(k_{upper} - k_{lower} + 1)$ instances of SIM Problem. As viral marketing is basically done in different phases and in each phase, seed set of different cardinalities can be used, influence spectrum problem appears in a natural way. Mathematically, influence spectrum problem can be written as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$ *and* $k_{lower}, k_{upper} \in \mathbb{Z}^+$ *with* $k_{upper} > k_{lower}$.
Problem: *Influence Spectrum Problem [Find out a* $\mathcal{S} \subseteq V(G)$ *with* $|\mathcal{S}| = k$, $\forall k \in [k_{lower}, k_{upper}]$ *such that and for any other* $\mathcal{S}' \subseteq V(G)$ *with* $|\mathcal{S}'| = k$, $\sigma(\mathcal{S}) \geq \sigma(\mathcal{S}')$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *and* $|\mathcal{S}| = k$ *for each* $k \in [k_{lower}, k_{upper}]$.

Multi-Round Influence Maximization Problem [114] Most of the existing studies of influence maximization consider that the seed set selection is one shot task, i.e., the entire seed set has to be selected before the diffusion starts. However, in many real-world advertisement scenarios, it may be required that the viral marketing need to be conducted in multiple times. To model this scenario, 'Multi-Round Influence Maximization Problem' has been introduced by Sun et al. [114]. In this problem, along with the social network $G(V, E, \theta, \mathcal{P})$, we are also given with two integers: k and T. Here, T is the number of times the diffusion process needs to be conducted and k is the cardinality of the seed set. Here, the goal is to choose the seed nodes $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_T$, such that at the end of the entire diffusion process, the total number of influenced nodes becomes maximum.

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$ *and* $k, T \in \mathbb{Z}^+$.
Problem: *Multi Round Influence Maximization Problem [Find out* $\mathcal{S}_t^* \subseteq V(G)$ *with* $|\mathcal{S}_t^*| = k$, *for all* $t \in [T]$ *such that and for any other* $\mathcal{S}_t' \subseteq V(G)$ *with* $|\mathcal{S}_t'| = k$, $\rho(\mathcal{S}_1^* \cup \mathcal{S}_2^* \cup \cdots \cup \mathcal{S}_T^*) \geq \rho(\mathcal{S}_1' \cup \mathcal{S}_2' \cup \cdots \cup \mathcal{S}_T')$*].*
Output: *The Seed Sets for Diffusion* $\mathcal{S}_t^* \subseteq V(G)$ *and* $|\mathcal{S}_t^*| = k$ *for each* $t \in [T]$.

Here, $\rho(.)$ is the multi round influence function, whose input is the selected seed nodes and the output is the number of influenced nodes.

Target Set Selection Problem [17] In this variant, for a given social network $G(V, E, \theta, \mathcal{P})$, the goal is to select a seed set, whose initial activation leads to the complete influence in the network, i.e., all the nodes are influenced at the end of diffusion process. Formally, this problem can be posed as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$.
Problem: *Target Set Selection Problem [Find out a minimum cardinality seed set* $\mathcal{S} \subseteq V(G)$ *such that* $\sigma(\mathcal{S}) = n$ *].*
Output: *The minimum cardinality seed set* $\mathcal{S}$ *for diffusion.*

λ *Coverage Problem* [89]: This is another variant of SIM Problem, which considers the minimum number of influenced nodes required at the end of diffusion. For a given social network $G(V, E, \theta, \mathcal{P})$ and a constant $\lambda \in [n]$, this problem asks to find a subset $\mathcal{S}$ of its nodes with minimum cardinality, such that at least λ number of nodes will be influenced at the end of diffusion process. Mathematically, this problem can be described in the following way:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$ *and* $\lambda \in [n]$.
Problem: λ *Coverage Problem [Find out the latest cardinality subset* $\mathcal{S} \subseteq V(G)$ *such that* $\sigma(\mathcal{S}) \geq \lambda$ *].*
Output: *The minimum cardinality seed set* $\mathcal{S}$ *for diffusion.*

Weighted Target Set Selection Problem (WTSS Problem) [101] This is another (infect weighted) variant of SIM Problem. Along with a social network $G(V, E, \theta, \mathcal{P})$, we are given another *vertex weight function*, $\phi : V(G) \to \mathbb{N}_0$, signifying the cost associated with each vertex. This problem asks to find out a subset $\mathcal{S}$, which *minimizes* total *selection cost*, and also all the nodes will be influenced at the end of diffusion. Mathematically, this problem can be stated as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$, *vertex cost function* $\phi : V(G) \to \mathbb{N}_0$.
Problem: *Weighted TSS Problem [Find out the subset* $\mathcal{S} \subseteq V(G)$ *such that* $\phi(\mathcal{S})$ *is minimum and* $|\sigma(\mathcal{S})| = n$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *with minimum* $\phi(\mathcal{S})$ *value.*

r-round min-TSS Problem [16] It is a variant of SIM Problem, which considers the number of rounds required to complete the diffusion process. Along with a *directed graph* $G(V, E, \theta, \mathcal{P})$, we are given the maximum number of allowable rounds $r \in \mathbb{Z}^+$ and asks to find out a minimum cardinality seed set $\mathcal{S}$, which activates all the nodes of the network within r-round. Mathematically, this problem can be described as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$ *and* $r \in \mathbb{Z}^+$.
Problem: *r-round min-TSS Problem [Find out the most minimum cardinality subset* $\mathcal{S}$ *such that* $\cup_{i=1}^{r} \sigma_i(\mathcal{S}) = V(G)$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$.

Here, $\sigma_i(\mathcal{S})$ denotes the set of influenced nodes from the seed set $\mathcal{S}$ at the ith round of diffusion.

Budgeted Influence Maximization Problem (BIM Problem) [92] This is another variant of SIM Problem, which is recently gaining popularity. Along with a *directed graph* $G(V, E, \theta, \mathcal{P})$, we are given with a cost function $\mathcal{C} : V(G) \longrightarrow \mathbb{Z}^+$ and a fixed budget $\mathcal{B} \in \mathbb{Z}^+$. Cost function $\mathcal{C}$ assigns a nonuniform selection cost to every vertex of the network, which is the amount of incentive need to be paid, if that vertex is selected as a seed node. This problem asks for selecting a seed set within the budget, which maximizes the spread of influence in the network.

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$*, a cost function* $\mathcal{C} : V(G) \longrightarrow \mathbb{Z}^+$ *and affordable budget* $\mathcal{B} \in \mathbb{Z}^+$.
Problem: *Budgeted Influence Maximization Problem [Find out the seed set* $(\mathcal{S})$ *such that* $\sum_{u \in \mathcal{S}} \mathcal{C}(u) \leq \mathcal{B}$ *and for any other seed set* $\mathcal{S}'$ *with* $\sum_{v \in \mathcal{S}'} \mathcal{C}(v) \leq \mathcal{B}$, $|\sigma(\mathcal{S})| \geq |\sigma(\mathcal{S}')|$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *with* $\sum_{u \in \mathcal{S}} \mathcal{C}(u) \leq \mathcal{B}$.

(λ, β, α) *TSS Problem* [28]: This is another variant of TSS Problem, which considers the maximum cardinality of the seed set (β), maximum allowable diffusion rounds (λ), and number of influenced nodes at the end of diffusion process (α) all together. Along with the input graph $G(V, E, \theta, \mathcal{P})$, we are given with the parameters: λ, β and α. Mathematically, this problem can be stated as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$*, three parameters* $\lambda, \beta \in \mathbb{N}$ *and* $\alpha \in [n]$.
Problem: (λ, β, α) *TSS Problem [Find out the subset* $\mathcal{S} \subseteq V(G)$ *such that* $|\mathcal{S}| \leq \beta$, $|\cup_{i=1}^{\lambda} \sigma_i(\mathcal{S})| \geq \alpha$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *and* $|\mathcal{S}| \leq \beta$.

(λ, β, A) *TSS Problem* [28]: This is a slightly different from the (λ, β, α) TSS problem, in which instead of the required number of the nodes after the diffusion process, it explicitly maintains which nodes should be influenced. Along with the input social network

$G(V, E, \theta, \mathcal{P})$, we are also given with maximum allowable rounds (λ), maximum cardinality of the seed set (β), and set of nodes $A \subseteq V(G)$ need to be influenced at the end of diffusion process as input. This problem asks for selecting a seed set of maximum β elements, which will influence all the nodes in A within λ rounds of diffusion. Mathematically, the problem can be stated as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$, $A \subseteq V(G)$ *and two parameters* $\lambda, \beta \in \mathbb{N}$.
Problem: (λ, β, A) *TSS Problem [Find out the subset* $\mathcal{S} \subseteq V(G)$ *such that* $|\mathcal{S}| \leq \beta$, $A \subseteq \cup_{i=1}^{\lambda} \sigma_i(\mathcal{S})$*].*
Output: *The Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$ *and* $|\mathcal{S}| \leq \beta$.

(λ, A) *TSS Problem* [28]: This is slightly different from (λ, β, A) TSS Problem. Here, we are interested in finding the minimum cardinality seed set, such that within some fixed numbers of diffusion rounds (λ), a subset of the nodes (A) will be influenced. Mathematically, the problem can be stated as follows:

Instance: *A Directed Graph* $G(V, E, \theta, \mathcal{P})$, $A \subset V(G)$ *and* $\lambda \in \mathbb{N}$.
Problem: (λ, A) *TSS Problem [Find out the subset* $\mathcal{S}$ *such that* $A \subseteq \cup_{i=1}^{\lambda} \sigma_i(\mathcal{S})$ *and for any other* $\mathcal{S}'$ *with* $|\mathcal{S}'| < |\mathcal{S}|$ $A \nsubseteq \cup_{i=1}^{\lambda} \sigma_i(\mathcal{S}')$*].*
Output: *Minimum cardinality Seed Set for Diffusion* $\mathcal{S} \subseteq V(G)$.

We have described different variants of TSS Problem in social networks available in the literature. It is surprising to see that only Top-k-node Problem has been studied, in depth.

4 Hardness results of TSS problem

In this section, we have described the hardness results of SIM Problem under both traditional and parameterized complexity theoretic perspectives.

4.1 Traditional complexity theoretic results

Initially, the problem of social influence maximization was posed by Domingos and Richardson [37,104] in the context of viral marketing. However, Kempe et al. [67] were the first to investigate the computational issues of the problem. They were able to show that SIM Problem under IC and LT Model is a special case of *Set Cover Problem* and *Vertex Cover Problem*, respectively. Both the set cover and vertex cover problems are well-known *NP-hard* problems [48]. The conclusion is presented as Theorem 1.

Theorem 1 (Hardness and inapproximability result of SIM Problem) [67] *Social Influence Maximization Problem is NP-hard for both IC Model and LT Model and also NP-hard to approximate within a factor of* $\mathcal{O}(n^{(1-\epsilon)})$ *for all* $\epsilon > 0$.

Chen [17] studied the TSS Problem. His study was different from Kempe et al.'s [67] study in two ways. First, Kempe et al. [67] investigated the Top-k node problem, whereas Chen [17] studied the TSS Problem. Second, Kempe et al. [67] studied the diffusion process under

IC and LT Models, which are probabilistic in nature, whereas Chen [17] considered all the *deterministic diffusion models* like *majority threshold model*, *constant threshold model*, and *unanimous threshold model*. In general, for the TSS Problem, Chen [17] came up with a seminal result presented in Theorem 2.

Theorem 2 (Inapproximability result of TSS Problem) [17] *TSS Problem cannot be approximated with in the constant factor* $\mathcal{O}(2^{\log^{(1-\epsilon)} n})$ *unless* $NP \subseteq DTIME(n^{\text{polylog}(n)})$ *for any fixed constant* $\epsilon > 0$.

This theorem can be proved by a reduction from the *Minimum Representative Problem* given in [75]. Next, they have shown that in *majority threshold model* also, λ-*coverage problem* follows the similar result, as presented in Theorem 2. However, when $\forall u \in V(G)\ \theta(u) = 1$, the TSS Problem can be solved very intuitively as targeting one node in each component results in the activation of all the nodes of the network. Surprisingly, this problem becomes hard, when we allow the vertex threshold to be at most 2, i.e., $\forall u \in V(G)\ \theta(u) \leq 2$. They proved the following result in this regard.

Theorem 3 (Hardness result of TSS problem on bipartite graphs) [17] *The TSS Problem is NP-hard, when thresholds are at most 2, even for bounded bipartite graphs.*

This theorem can be proved by a reduction from a variant of 3-SAT Problem presented in [121]. Moreover, Chen [17] has shown that for *unanimous threshold model*, the *TSS Problem* is equivalent to *vertex cover problem*, which is a well-known NP-Complete Problem.

Theorem 4 (Equivalence between TSS problem and vertex cover problem) [17] *If all the vertex thresholds of the graph are unanimous (i.e.,* $\forall u \in V(G)$, $\theta(u) = \deg(u)$*), then the TSS Problem is identical to vertex cover problem.*

Chen [17] has also shown that if the underlying graph is tree, then the TSS Problem can be solved in polynomial time and they have also given the *ALG-tree* algorithm, which does this computation. Recently, Banerjee and Mathew [7] came up with an inapproximability result of the TSS Problem on bipartite graph by a reduction from the classical set cover problem. Formally, they proved Theorem 5.

Theorem 5 (Inapproximability result of TSS Problem on bipartite graphs) [7] *Unless P=NP, if the underlying influence graph is bipartite, the TSS Problem cannot have a polynomial time approximation algorithm with a performance guarantee better than a factor of* $\mathcal{O}(\log n_{\min})$*, where* $n_{\min}$ *is the smaller part in the bipartition.*

To the best of the authors' knowledge, there is no other literature, which focuses on the hardness analysis of the TSS Problem in traditional complexity theoretic perspective. We summarize these results in Table 2.

4.2 Parameterized complexity theoretic results

Now, we describe the hardness results based on the parameterized complexity theoretic perspective. For basic notions about *parameterized complexity*, readers may refer to [39]. Bazgan et al. [8] showed that SIM Problem under constant threshold model (CTM) does not have any parameterized approximation algorithm with respect to the parameter *seed set size*. Chopin et al. [26,27] studied the TSS Problem in parameterized settings with respect to the parameters related to network cohesiveness like *clique cover number* (number of cliques

Table 2 Hardness results of TSS Problem and its variants in traditional complexity theory perspective

Name of the problem	Diffusion/threshold model	Major findings
SIM Problem	IC Model	A special case of set cover problem and hence NP-hard [67]
	LT Model	A special case of vertex cover problem and hence NP-hard [67]
TSS Problem	MT Model	Not only NP-hard as well as cannot be approximated in the constant factor $\mathcal{O}(2^{\log^{(1-\epsilon)} n})$ unless $NP \subseteq DTIME(n^{\text{polylog}(n)})$ [17]
	CT Model with $\theta(u) = 1$, $\forall u \in V(G)$	Can be solved trivially by selecting a vertex from each component of the network [17]
	CT Model with $\theta(u) \leq 2$, $\forall u \in V(G)$	NP-Hard even for bounded bipartite graphs [17]
	UT Model	Identical to vertex cover problem and hence NP-hard [17]
	Any threshold model	Inapproximable better than a factor of $\mathcal{O}(\log n_{\min})$ for a bipartite graph [7]

required to cover all the vertices of the network [64]), *distance to clique* (number of vertices need to be deleted to obtain a clique), *cluster vertex deletion number* (number of vertices to delete in order to obtain a collection of disjoint cliques); parameters related to network density like *distance to cograph*, *distance to interval graph*; parameters related to sparsity of the network, namely *vertex cover number* (number of vertices to remove to obtain an edgeless graph), *feedback edge set number and feedback vertex set number* (number of edges or vertices to remove to obtain a forest), *pathwidth*, *bandwidth*. It is interesting to note that computing all the parameters except *feedback edge set number* is NP-hard problem. They studied the TSS Problem and came up with the following two important results related to the sparsity parameters of the network:

Theorem 6 (W1 hardness of TSS Problem) [27] *TSS Problem with majority threshold model is W[1] hard even with respect to the combined parameter feedback vertex set, distance to cograph, distance to interval graph and path width.*

Theorem 7 (FP tractability of TSS Problem) [27] *TSS Problem is fixed-parameter tractable with respect to the parameter bandwidth.*

For proving the above two theorems, authors have used reduction rules used in [98] and [97]. The results related to dense structure property of the network is given in Theorems 8 through 10.

Theorem 8 (W 1 hardness of TSS problem) *TSS Problem is W[1]-hard with parameter cluster vertex deletion number.*

Theorem 9 (NP-hardness and W2 hardness of TSS Problem) *TSS Problem is NP-hard and W2 hard with respect to the parameter target set size (k), even on graphs with clique cover number of two.*

Theorem 10 (FP tractability of TSS Problem) *TSS Problem is fixed-parameter tractable with respect to the parameter 'distance l to clique,' if the threshold function satisfies the following properties* $\theta(u) > g(l) \Rightarrow \theta(u) = f(\Gamma(u))$ *for all* $u \in V(G)$, $f : P(V(G)) \longrightarrow \mathbb{N}$ *and* $g : \mathbb{N} \longrightarrow \mathbb{N}$.

For a detailed proof of Theorems 8 through 10, readers may refer to [27]. All the results related to the parameterized complexity theory has been summarized in Table 3.

5 Major research challenges

Before entering into the critical review of the existing solution methodologies, in this section, we provide a brief discussion on major research challenges concerned with the SIM Problem. This will help the reader to understand which category of solution methodology can handle what challenge.

- **Trade of between accuracy and computational time** From the discussion in Sect. 4, it is now well understood that the SIM Problem is computationally hard from both traditional and parameterized complexity theoretic prospective, in general. Hence, for some given $k \in \mathbb{Z}^+$ obtaining the most influential k nodes within feasible time is not possible. In this scenario, the intuitive approach could be to use some heuristic method for selecting seed nodes. This will lead to less time for seed set generation. However, the number of influenced nodes generated by the seed nodes could be also arbitrarily less. In this situation, it is an important issue to design algorithms, which will run in affordable time and also, the gap between the optimal spread and the spread due to the seed set selected by an algorithm will be as small as possible.
- **Breaking the barrier of submodularity** In general, the social influence function $\sigma(.)$ is submodular (discussed in Sect. 6.1). However, in many practical situations, such as *opinion and topic specific influence maximization*, the social influence function may not be submodular [49,82]. This happens because one node can switch its state from positive opinion to negative opinion and the vice-versa. In this scenario, solving the SIM Problem may be more challenging due to the absence of submodularity property in the social influence function.
- **Practicality of the problem** In general, the SIM Problem takes many assumptions, such as every selected seed will perform up to expectation in the spreading process, influencing each node of the network is equally important, etc. This assumptions may be unrealistic in some situations. Assume the case of *target advertisement*, where instead of all the nodes, a set of target nodes are chosen and the aim is to maximize the influence within the target nodes [42,66]. In another way, due to the probabilistic nature of diffusion, a seed node may not perform up to expectation in the influence spreading process. Solving the SIM Problem and its variants will be more challenging, if we relax these assumptions.
- **Scalability** Real-life social networks have millions of nodes and billions of edges. So, for solving the SIM and related problems for real-life social networks, scalability should be an important issue for any solution methodology.
- **Theoretical challenges** For a computational problem, any of its solution methodology is concerned with two aspects. The first one is the *computational time*. This is measured as the execution time, when the methodology is implemented with real- life problem instances. The second one is the *computational complexity*. This is measured as the *asymptotic bound* of the methodology. Theoretical research on any computational problem always concerned with the second aspect of the problem. Hence, the theoretical challenge for the SIM Problem is to design algorithms with good asymptotic bounds.

Table 3 Hardness results of TSS Problem and its variants in parameterized complexity theory perspective

Name of the problem	Diffusion/threshold model	Parameter	Major findings
SIM Problem	Constant threshold model with $\theta(u) \in [\deg(u)]$	Seed set size	Does not have any parameterized approximation algorithm better than a factor of $\mathcal{O}(n^{1-\epsilon})$ [8]
TSS Problem	Majority threshold model	Feedback vertex set number, pathwidth, distance to cograph, distance to interval graph	The problem is $W[1]$-hard [27]
	General threshold model	Cluster vertex deletion number	The problem is $W[1]$-hard [27]
	Constant threshold model	Cluster vertex deletion number	The problem is fixed-parameter tractable [27]
	General threshold model	Seed set size	$W[2]$-hard even for graphs with clique cover number two [27]
	Majority threshold model, constant threshold model	Distance to clique	The problem is fixed-parameter tractable [27]

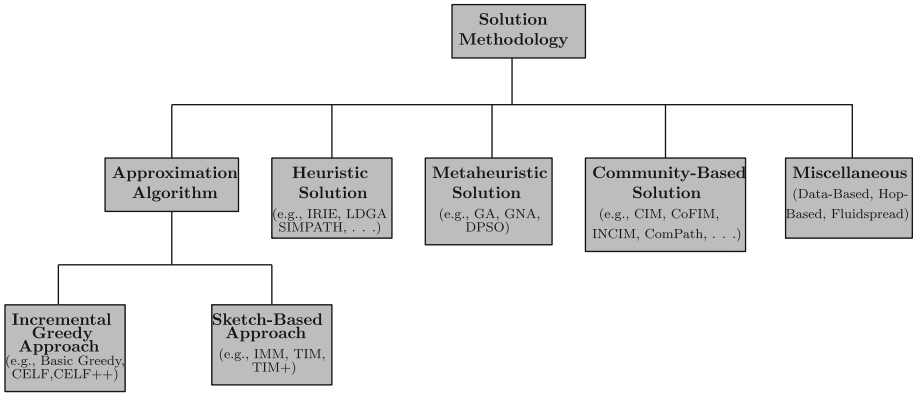

Fig. 2 Proposed taxonomy for classifying the solution methodologies

6 Solutions methodologies

Due to the inherent hardness of the SIM Problem, over the years, researchers have developed algorithms for finding seed set for obtaining near-optimal influence spread. In this section, the available solution methodologies in the literature have been described. At first, we describe our proposed taxonomy for classifying the solution methodologies. Figure 2 gives a diagrammatic representation of the proposed taxonomy and we describe them below.

- **Approximation algorithms with provable guarantee** Algorithms in this category give the worst-case bound for influence spread. However, most of them suffer from the scalability issues, which means, with the increase in the network size, running time grows heavily. Many of the algorithms of this category have near-optimal asymptotic bounds.
- **Heuristic solutions** Algorithms of this category do not give any worst-case bound on influence spread. However, most of them have more scalability and better running time compared to the algorithms of the previous category.
- **Metaheuristic solutions** Methodologies of this category are the metaheuristic optimization algorithms, and many of them are developed based on the evolutionary computation techniques. These algorithms also do not give any worst-case bound on influence spread.
- **Community-based solutions** Algorithms of this category use community detection of the underlying social network as an intermediate step to bring down the problem into the community level and improves scalability. Most of the algorithms of this category are heuristic and hence do not provide any worst-case bound on influence spread.
- **Miscellaneous** Algorithms of this category do not follow any particular property, and hence, we put them under this heading.

6.1 Approximation algorithms with provable guarantee

Kempe et al. [67–69] were the first to study the problem of social influence maximization as a *combinatorial optimization* problem and investigated its computational issues under two diffusion models, namely LT and IC models. In their studies, they assumed that the *social influence function*, $\sigma()$ is submodular and monotone. The function $\sigma : 2^{V(G)} \rightarrow \mathbb{R}_0^+$ will be submodular, if it follows the *diminishing return property*, which means $\forall\, \mathcal{S} \subseteq \mathcal{T} \subset V(G)$, $u_i \in V(G) \backslash \mathcal{T}$; $\sigma(\mathcal{S} \cup \{u_i\}) - \sigma(\mathcal{S}) \geq \sigma(\mathcal{T} \cup \{u_i\}) - \sigma(\mathcal{T})$ and σ will be monotone, if

for any $\mathcal{S} \subset V(G)$ and for all $u_i \in V(G)\backslash\mathcal{S}$, $\sigma(\mathcal{S} \cup u_i) \geq \sigma(\mathcal{S})$. They proposed a greedy strategy for selecting seed set presented in Algorithm 1.

Algorithm 1: Kempe et al.'s [67] Greedy Algorithm for *Seed Set Selection*. (**Basic Greedy**)

Data: Given Social Network $G(V, E, \theta, \mathcal{P})$ and some $k \in \mathbb{Z}^+$.
Result: Seed Set for diffusion $\mathcal{S} \subseteq V(G)$.

$\mathcal{S} \leftarrow \phi$;
for $i = 1$ *to* k **do**
  $u = \underset{u_i \in V(G)\backslash\mathcal{S}}{argmax} \ \sigma(\mathcal{S} \cup u_i) - \sigma(\mathcal{S})$;
  $\mathcal{S} \leftarrow \mathcal{S} \cup u$
return $\mathcal{S}$

Starting with the empty seed set ($\mathcal{S}$), Algorithm 1 iteratively selects node, which is currently not in $\mathcal{S}$, and inclusion of which to $\mathcal{S}$ causes the maximum marginal increment in $\sigma()$. Let us assume that $\mathcal{S}_i$ denotes the seed set at ith iteration of the *'for' loop* in Algorithm 1. In $(i+1)$th iteration, $\mathcal{S}_{i+1} = \mathcal{S}_i \cup \{u\}$, if $\sigma(\mathcal{S} \cup \{u\}) - \sigma(\mathcal{S})$ value becomes the maximum among all $u \in V(G)\backslash\mathcal{S}_i$. This iterative process will be continued until we reach the allowed cardinality of $\mathcal{S}$. Kempe et al. [67] showed that Algorithm 1 provides $(1 - \frac{1}{e} - \epsilon)$ with $\epsilon > 0$ for the approximation bound on influence spread, maintained in Theorem 11.

Theorem 11 *Algorithm 1 provides* $(1 - \frac{1}{e} - \epsilon)$ *with* $\epsilon > 0$ *factor approximation bound for the SIM Problem, i.e., if* $\mathcal{S}^*$ *is the* k *element optimal seed set, then* $\sigma(\mathcal{S}) \geq (1 - \frac{1}{e})\sigma(\mathcal{S}^*)$, *where* $e = \sum_{x=0}^{\infty} \frac{1}{x!}$.

Though Algorithm 1 gives good approximation bound on influence spread, it suffers from two major shortcomings. For example, for any given seed set $\mathcal{S}$, exact computation of the influence spread (i.e., $\sigma(\mathcal{S})$) is $\#P\text{-}Complete$. Hence, they approximate the influence spread by running a huge number of *Monte Carlo Simulations* (MCS), counting total number of influenced nodes in all simulation runs and taking average with the number of runs. However, recently Maehara et al. [88] developed the first procedure for exact computation of influence spread using *binary decision diagrams*. Second, the number of times influence function ($\sigma(.)$) needs to be evaluated is quite huge. For selecting a seed set of size k with $\mathcal{R}$ number of MCS runs in a social network having n nodes and m edges will require $\mathcal{O}(kmn\mathcal{R})$ number of influence function evaluations. Hence, application of this algorithm for a medium size networks (only consisting of 15000 nodes, though real-life networks are much larger) appears to be unrealistic [19], which means that the algorithm is not scalable enough.

In spite of having a few drawbacks, Kempe et al.'s [67] study is considered to be the foundational work on the SIM Problem. This study has triggered a vast amount of research in this direction. In most of the cases, the main focus was to reduce the scalability problem incurred by Basic Greedy algorithm in Kempe et al.'s work. Some of them landed with heuristics, in which the obtained solution could be far away from the optima. Still a few studies are there, in which scalability problem was reduced significantly without loosing approximation ratio. Here, we have listed the algorithms, which could provide approximation guarantee, whereas in Sect. 6.2, we have described all the heuristic methods.

– **CELF** For improving the scalability problem, Leskovec et al. [80] proposed a *Cost Effective Lazy Forward* (CELF) scheme by exploiting the submodularity property of the social influence function. The key idea in their study was: For any node, its marginal

gain in influence spread in the current iteration cannot be more than its marginal gain in the previous iterations. Using this idea, they were able to make a drastic reduction in the number of evaluations of the influence estimation function ($\sigma(.)$), which leads to significant improvement in running time though the asymptotic complexity remains the same as that of the *Basic Greedy Algorithm* (i.e., $\mathcal{O}(kmn\mathcal{R})$). The reported results in their paper show that CELF can speed up the computation process up to 700 times compared to Basic Greedy algorithm on benchmark data sets. This algorithm is also applicable to many other contexts, such as finding informative blogs in a *web blog network*, optimal placement of sensors in a water distribution network for detecting outbreaks, etc.

- **CELF++** Goyal et al. [54] proposed an optimized version of CELF by exploiting the submodularity property of social influence function and named it as CELF++. For each node u of the network, CELF++ maintains a table of the form $< u.mg1, u.prev\_best, u.mg2, u.f\_lag >$ where $u.mg1$ is the marginal gain in $\sigma(.)$ for the current $\mathcal{S}$; $u.prev\_best$ is the node with the maximum marginal gain among the users scanned till now in the current iteration; $u.mg2$ is the marginal gain in $\sigma(.)$ for u with respect to the $\mathcal{S} \cup \{prev\_best\}$ and $u.flag$ is the iteration number, when $u.mg1$ was last updated. The key idea in CELF++ is that, if $u.prev\_best$ is included in the seed set in the current iteration, then the marginal gain of u in $\sigma(.)$ with respect to $\mathcal{S} \cup \{prev\_best\}$ need not be recomputed in the next iteration. The reported results showed that CELF++ is 35–55% faster than CELF though the asymptotic complexity remains the same.
- **MIA and PMIA** Chen et al. [20] and Wang et al. [126] proposed *maximum influence arborescence* (MIA) and Prefix excluding MIA (PMIA) model of influence propagation. They computed the propagation probability from a seed node to a non-seed node by multiplying the influence probabilities of the edges present in the shortest path. *Maximum Influence Path* is the one having the maximum propagation probability, and they considered that influence spreads through local arborescence (a directed graph in which, for a vertex u called the root and any other vertex v, there is exactly one directed path from u to v) only. Hence, the model is called MIA. In PMIA (*Prefix excluding* MIA) model, for any seed s_i, its maximum influence path to other nodes should avoid all seeds that are before s_i. They proposed greedy algorithms for selecting seed set based on these two diffusion models. The reported results show that both MIA and PMIA can achieve a high level of scalability.
- **Static greedy** Cheng et al. [25] developed this algorithm for solving SIM Problem, which provides both guaranteed accuracy and high scalability. This algorithm works in two stages. In the first stage, R number of Monte Carlo snapshots are taken from the social network, where each edge (uv) is selected based on the associated diffusion probability p_{uv}. In the second stage, starting from the empty seed set, a node having the maximum average marginal gain in influence spread over all sampled snapshots will be selected as a seed node. This process will be continued until k nodes are selected. This algorithm has the running time of $\mathcal{O}(\mathcal{R}m + k\mathcal{R}m'n)$ and space requirement of $\mathcal{O}(\mathcal{R}m')$, where $\mathcal{R}$ and m' are the number of Monte Carlo samples and average number of active edges in the snapshots, respectively. The reported results show that the Static Greedy reduces the computational time by two orders of magnitude, while achieving the better influence spread compared to degree discount heuristic (DDH), maximum degree heuristic (MDH), prefix excluding maximum influence arborescence (PMIA) algorithms.
- **Borgs et al.'s method** Borgs et al. [9] proposed a completely different approach for solving SIM Problem under IC Model using *reverse reachable sampling technique*. Other than the MCS runs, this is a new approach for estimating the influence spread. Their algorithm is randomized and succeeds with the probability of $\frac{3}{5}$ and has the running

time of $\mathcal{O}((m+n)\epsilon^{-3}\log n)$, which improves the previously best known algorithm having the complexity of $\mathcal{O}(mnkPOLY(\epsilon^{-1}))$. The algorithm proposed by Borgs et al. is near-optimal, since the lower bound is $\Omega(m+n)$. This algorithm works in two phases. In the first phase, a hypergraph ($\mathcal{H}$) is generated stochastically from the input social network. The second phase is concerned with the seed set selection. This is done by repeatedly choosing the node with maximum degree in $\mathcal{H}$, deleting it along with its incidence edges from $\mathcal{H}$. The k-element set obtained in this way is the seed set for diffusion. This work is mostly theoretically enriched and lacking of practical experimentation.

- **Zohu et al.'s method** Zohu et al. [147] improved the approximation bound from $(1-\frac{1}{e})$ (which is approximately 0.63) to 0.857. They designed two approximation algorithms: The first algorithm works for the problem, where the cardinality of the seed set ($\mathcal{S}$) is not restricted and the second one works, when there is some restricted upper bound on the cardinality of seed set. They formulated the influence maximization problem as an optimization problem given below.

$$\max_{\mathcal{S}\subset V(G)} \sum_{u\in\mathcal{S}, v\in V(G)\setminus\mathcal{S}} p_{uv}, \tag{1}$$

where p_{uv} is the *influence probability* between the users: u and v. They converted this optimization problem into a *quadratic integer programming problem* and solved the problem using the concept of *semidefinite programming* [43].

- **SKIM** Cohen et al. [30] proposed a *Sketch-Based Influence Maximization* (SKIM) algorithm, which improves the Basic Greedy algorithm by ensuring in every iteration, with sufficiently high probability, or in expectation, the node we choose to add to the seed set has a marginal gain that is close to the maximum one. The running time of this algorithm is $\mathcal{O}(nl+\sum_{i=1}|E^i|+m\epsilon^{-2}\log^2 n)$, where l is the number of snapshots of G, E^i is the edge set of G^i. The reported results show that SKIM has high scalability over Basic Greedy, Two phase Influence Maximization (TIM), Influence Ranking and Influence Estimation (IRIE), etc., without compromising influence spread.
- **TIM** Tang et al. [118] developed a *Two-phase Influence Maximization* (TIM) algorithm, which has the expected running time of $\mathcal{O}((k+l)(n+m)\log n/\epsilon^2)$ with at least $(1-n^{-l})$ probability for some given k, ϵ and l. As its name suggests, this algorithm has two phases. In the first phase, TIM computes lower bound on the maximum expected influence spread among all k-sized sets and uses this lower bound to estimate a parameter ϕ. In the second phase, ϕ number of reverse reachability (RR) set samples have been picked up from the social network. Then, it derives a k-sized seed set that covers the maximum number of RR sets and returns as the final result. The reported results show that TIM is two times faster than CELF++ and Borgs et al.'s [9] Method, while achieving the same influence spread. To improve the running time of TIM, Tang et al. [118] proposed a heuristic which takes all the RR sets, generated in an intermediate step of the second phase of TIM as inputs. Then, it uses a greedy approach for the maximum coverage problem for selecting the seed set. This modified version of TIM is named as TIM$^+$. The reported results showed that TIM$^+$ is two times faster than TIM.
- **IMM** Tang et al. [119] proposed *Influence Maximization via Martingales* (IMM) (a kind of stochastic process, in which, for the given current and preceding values, the conditional expectation of the next value, will be the current value itself) which achieves a $\mathcal{O}((k+l)(n+m)\log n/\epsilon^2)$ expected running time and returns $(1-\frac{1}{e}-\epsilon)$ factor approximate solution with probability of $(1-n^{-l})$. IMM Algorithm also has two phases like TIM and TIM$^+$. The first phase is concerned with sampling RR sets from the given

social network, and the second phase is concerned with the seed set selection. In the first phase, unlike TIM and TIM$^+$, RR sets generated in the first phase are dependent because $(i+1)$-th RR set is generated based on whether first i of RR sets are satisfying stopping criteria or not. In IMM, the RR sets generated in the sampling phase are reused in node selection phase, which is not the case in TIM or TIM+. In this way, IMM can eliminate a lot of unnecessary computations, which leads to significant improvement in running time though asymptotic complexity remains the same as that of TIM. The reported results conclude that IMM outperforms TIM, TIM+, IRIE (described in Sect. 6.2) based on running time while achieving comparable influence spread.

- **Stop-and-stare** Nguyen et al. [94] developed the Stop-and-Stare Algorithm (SSA) and its dynamic version DSSA for *Topic-aware Viral Marketing* (TVM) problem. We have not discussed this problem, as it comes under topic aware influence maximization. However, this solution methodology can be used for solving SIM Problem with minor modification. They showed that the number of RR set samples used by their algorithms is asymptotically minimum. Hence, Stop-and-Stare is 1200 times faster than the state-of-the art IMM algorithm. We are not discussing the results as they are for the TVM problem and out of the scope of this survey.
- **BCT** Recently, Nguyen et al. [96] proposed *Billion-scale Cost-award Targeted* (BCT) algorithm for solving *cost-aware targeted viral marketing* (CTVM) introduced by them. We have not discussed this problem, as it comes under topic aware influence maximization. However, this solution methodology can be adopted for solving SIM Problem as well under both IC and LT Models and have the running time of $\mathcal{O}((k+l)(n+m)\log n/\epsilon^2)$ and $\mathcal{O}((k+l)n\log n/\epsilon^2)$, respectively. We are not discussing about the results, as they are for CTVM Problem and out of scope of this survey.
- **Nguyen et al.'s method** Nguyen et al. [92] studied the *Budgeted Influence Maximization Problem* described in Sect. 3. They have formulated the following optimization problem in the context of *Budgeted Influence Maximization*:

$$max \quad \sigma(\mathcal{S}) \tag{2}$$

$$\text{subject to,} \sum_{u\in\mathcal{S}} \mathcal{C}(\{u\}) \leq \mathcal{B} \tag{3}$$

Now, if $\forall u \in V(G)$, $\mathcal{C}(u) = 1$, then it will become the SIM Problem. To solve this problem, they proposed two algorithms. The first one is a modification of the Basic Greedy algorithm proposed by Kempe et al. [67] (Algorithm 1), and the second one was adopted from [70]. In the first algorithm, for all $u \in V(G)\backslash\mathcal{S}$, they computed the increment of influence in unit cost as follows:

$$\delta(u) = \frac{\sigma(\mathcal{S}\cup\{u\}) - \sigma(\mathcal{S})}{\mathcal{C}(\{u\})} \tag{4}$$

Now, the algorithm will choose u to include in the seed set ($\mathcal{S}$), if it maximizes the *objective function* as well as $\mathcal{C}(\mathcal{S}_i \cup \{u\}) \leq \mathcal{B}$. This iterative process will be continued until no more nodes can be added within the budget. However, this algorithm does not give any constant approximation ratio. This algorithm can be modified to get the constant approximation ratio, as given in Algorithm 2.

Algorithm 2: Nguyen et al.'s [92] greedy algorithm for BIM Problem.

Data: Given Social Network $G(V, E, \theta, \mathcal{P})$, cost function $\mathcal{C} : V(G) \longrightarrow \mathbb{Z}^+$ some $\mathcal{B} \in \mathbb{Z}^+$.

Result: Seed Set for diffusion $\mathcal{S} \subseteq V(G)$.

1 $S_1 =$ result of Naive Greedy;

2 $S_{max} = \underset{u \in V(G)}{argmax} \ \sigma(u)$;

3 $\mathcal{S} = argmax(\sigma(S_1), \sigma(S_{max}))$;

4 $return \ \ \mathcal{S}$

Theorem 12 *Algorithm 2 guarantees* $(1 - \frac{1}{\sqrt{e}})$ *approximate solution for BIM Problem.*

For the detailed proof of Algorithm 2, readers are referred to the appendix of [91].

Now, the presented algorithms have been summarized below. The main bottleneck in Kempe et al.'s [67] Basic Greedy algorithm is the evaluation of influence spread estimation function for a large number of MCS runs (say, 10,000). If we reduce the MCS runs directly, then the accuracy in computing influence spread may be compromised. So, the key scope for improvement is to reduce the number of evaluation of the influence estimation function in each MCS run. Both CELF and CELF++ exploit the submodularity property to achieve this goal and hence are found to be faster than Basic Greedy algorithm. On the other hand, Static Greedy algorithm uses all the randomly generated snapshots of the social network using MCS runs simultaneously. Hence, with the less number of MCS runs (say, 100) it is possible to have equivalent accuracy in spread. These four algorithms can be ordered in terms of maximum-to-minimum values of running time as follows: Basic Greedy $\succ$ CELF $\succ$ CELF++ $\succ$ Static Greedy.

Another scope of improvement in Kempe et al.'s [67] work was estimating the influence spread by applying some method other than the heavily time-consuming MCS runs. Borgs et al. [9] explored this scope by proposing a drastically different approach for spread estimation, namely reverse reachable sampling technique. The algorithms (such as TIM, TIM$^+$, IMM) which used this method were seemed to be much faster than CELF++ and also have competitive influence spread. Among TIM, TIM$^+$ and IMM , IMM was found to be the fastest one both theoretically (in terms of computational complexity) and empirically (in terms of computational time from experimentation) due to the reuse of the RR sets in the node selection phase. To the best of the authors knowledge, IMM is the fastest algorithm, which was solely proposed for solving SIM Problem. However, BCT algorithm developed by Nguyen et al. [96], which was originally proposed for solving CTVM problem, is the fastest solution methodology available in the literature that can be adopted for solving SIM Problem.

Now, from this discussion, it is important to note that the scalability problem incurred by the Basic Greedy algorithm had been reduced by the subsequent research. However, as the size of the social network data set has become gigantic, development of algorithms with high scalability remains the thrust area. The solution methodologies described till now have been summarized in Table 4. In the case of the algorithms, for which the complexity analysis has not been done by the author(s), the columns of the table are left blank. Pros and cons of the approximation algorithms for SIM and related problems have been summarized in Table 5.

Table 4 Approximation algorithms for SIM Problem and its variants

Name of the algorithm	Proposed by	Complexity	Applicable for	Model
Basic Greedy	Kempe et al. [67]	$\mathcal{O}(kmn\mathcal{R})$	SIM	IC and LT
CELF	Leskovec et al. [78]	$\mathcal{O}(kmn\mathcal{R})$	SIM	IC and LT
CELF++	Goyal et al. [54]	$\mathcal{O}(kmn\mathcal{R})$	SIM	IC and LT
MIA,PMIA	Chen et al. [20], Wang et al. [126]	$\mathcal{O}(nt_{i\theta} + kn_{o\theta}.n_{i\theta}(n_{i\theta} + \log n))$	SIM	MIA, PMIA
Static Greedy	Cheng et al. [25]	$\mathcal{O}(\mathcal{R}m + kn\mathcal{R}m)$	SIM	IC and LT
Brog et al.'s Method	Brogs et al. [9]	$\mathcal{O}(kl^2(m+n)\log^2 n/\epsilon^3)$	SIM	IC and LT
Zohu et al.'s Method	Zohu et al. [147]	–	SIM	IC and LT
SKIM	Cohen et al. [30]	$\mathcal{O}(nl + \sum_{i=1} \lvert E^i \rvert + m\epsilon^{-2}\log^2 n)$	SIM	IC and LT
TIM+, IMM	Tang et al. [118,119]	$\mathcal{O}((k+l)(n+m)\log n/\epsilon^2)$	SIM	IC and LT
Stop-and-Stare	Nguyen et al. [94]	–	TVM	IC and LT
Nguyen's Method	Nguyen et al. [92]	$\mathcal{O}(n^2(\log n + d) + kn(1+d))$	BIM	IC and LT
BCT	Nguyen et al. [96]	$\mathcal{O}((k+l)(n+m)\log n/\epsilon^2)$	SIM, BIM, CTVM	IC
BCT	Nguyen et al. [96]	$\mathcal{O}((k+l)n\log n/\epsilon^2)$	SIM, BIM, CTVM	LT

Table 5 Advantages and disadvantages of different algorithms

Name of the algorithm	Advantages	Disadvantages
Basic Greedy	Provides approximation guarantee	Huge computational time and poor scalability
	Easy to implement	Not usable for practical problems
CELF	Provides approximation guarantee	Computational time still high and do not have enough scalability
	Computational time is lesser than Basic Greedy	Not usable for large-scale social networks
CELF++	Provides approximation guarantee	Computational time is not good enough to work on real-world social networks
	Computational time is lesser than Basic Greedy and CELF	
MIA,PMIA	Provides approximation guarantee	Algorithm is parameter dependent, hence to obtain acceptable result parameter needs to be tuned properly
	First algorithm based on the MIA diffusion model	
Static Greedy	Provides approximation guarantee	Scalability is not enough to work with large-scale datasets
	Resolves the scalability-accuracy dilemma to some extent	Space requirement is more compared to Basic Greedy
Brog et al.'s Method	Provides approximation guarantee	Methodology is dependent on three parameters
	This method is theoretical break through in the influence maximization literature	Lack of practical experimentation
Zohu et al.'s Method	Provides approximation bound	Considers propagation distance
	Approximation bound has been improved from $(1 - \frac{1}{e})$ to 0.857	Experimentation lacks of large datasets
SKIM	Provides approximation bound	This methodology consists of parameters. So, proper tunning of these parameter may be an issue
	Introduce the concept of 'influence oracle'	
	Experimentation with large datasets	
TIM+, IMM	Provides approximation guarantee	These methodologies consists of parameters. So, parameter tunning may be an issue
	Highly scalable compared to the existing methods	
Stop-and-Stare	Provides approximation guarantee	These methodologies consists of parameters. So, parameter tunning may be an issue
	Uses asymptotically minimum number of samples	
	Even faster than IMM	

Table 5 continued

Name of the algorithm	Advantages	Disadvantages
Nguyen's Method	One of the proposed methods provides approximation guarantee	Methods are not scalable
	First study on the BIM Problem	
BCT	Provides approximation guarantee	Proposed methodology consists of many parameters. To get the effective performance parameters need to be tuned
	Proposed a generalized framework for the SIM and related problems	

6.2 Heuristic solutions

Algorithms of this category do not provide any approximation bound on the influence spread but have better running time and scalability. Here, we will describe the heuristic solution methodologies from the literature.

- **Random heuristic** For selecting seed set by this method, randomly pick k nodes of the network and return them as seed set. In Kempe et al.'s [67] experiment, this method has been used as a baseline method.
- **Centrality-based heuristics** Centrality is a well-known measure in network analysis, which signifies how much importance a node has in the network [45,76]. There are many centrality-based heuristics proposed in the literature for SIM Problem like *Maximum Degree Heuristic* (MDH) (select k highest degree nodes as seed node), High Clustering Coefficient Heuristic (HCH) (select k nodes with the highest clustering coefficient value) [115,134], High Page Rank Heuristic [11] (select k nodes with the highest page rank value), etc.
- **Degree discount heuristic** (DDH) This is basically the modified version of MDH and was proposed by Chen et al. [19]. The key idea behind this method is following for any two nodes $u, v \in V(G)$, $(uv) \in E(G)$ and u has been selected as a seed set by MDH, and then, during the counting the degree of v, the edge (uv) should not be considered. Hence, due to the presence of u in the seed set, the degree of v will be discounted by 1. This method is also named as *Single Discount Heuristic* (SDH). Experimental results of [19] show that DDH can achieve better influence spread than MDH.
- **SIMPATH** This heuristic was proposed by Goyal et al. [55] for solving SIM Problem under LT Model. SIMPATH works based on the principal of CELF (discussed in Sect. 6.1). However, instead of using computationally expensive Monte Carlo simulations for estimating influence spread, SIMPATH uses path enumeration techniques for this purpose. This algorithm has a parameter (η) for controlling trade off between influence spread and running time. The reported results conclude that SIMPATH outperforms other heuristics, such as MDH, Page Rank and LDGA with respect to information spread.
- **SPIN** Narayanam et al. [89] studied SIM Problem and λ Coverage Problem as a co-operative game and proposed a *Shapley Value-Based Discovery of Influential Nodes* (SPIN) Algorithm, which has the running time of $\mathcal{O}(t(n+m)\mathcal{R} + n\log n + kn + k\mathcal{R}m)$, where t is the cardinality of the sample collision set being considered for the computation of shapley value. This algorithm has mainly two steps. The first one is to generate a rank

list of the nodes based on the shapley value and then to choose top-k of them and to return as seed set. The reported results show that SPIN constantly outperforms MDH and HCH.

- **LDAG** Chen et al. [21] developed this heuristic for solving SIM Problem under LT Model. Influence spread in a *Directed Acyclic Graph* (DAG) is easy to compute. Hence, for computing the influence spread in general social networks, they introduced a *Local Directed Acyclic Graph* (LDAG)-based influence model, which computes local DAGs for each node to approximate influence spread. After constructing the DAGs, Basic Greedy algorithm proposed by Kempe et al. [67] can be used to select the seed nodes. The reported results show that LDAG constantly outperforms DDH or Page Rank heuristic.
- **IRIE** Jung et al. [62] proposed this heuristic based on influence ranking (IR) and influence estimation (IE) for solving SIM Problem under IC and its extension IC-N (independent cascade with negative opinion) Model. They developed a global influence ranking like belief propagation approach. If we select top-k-nodes, then there will be an overlap in influence spread by each node. For avoiding this shortcoming, they integrated a simple *influence estimation* technique to predict additional influence impact of a seed on the other node of the network. The reported results show that IRIE can achieve better influence spread compared to MDH, Pagerank, PMIA, etc., heuristics. However, IRIE has less running time and memory consumption.
- **ASIM** Galhotra et al. [46] designed this highly scalable heuristic for SIM Problem. For each node $u \in V(G)$, this algorithm assigns a score value (the weighted sum of the number of simple paths of length at most d starting from that node). ASIM has the running time of $\mathcal{O}(kd(m+n))$, and its idea is quite similar to the SIMPATH Algorithm proposed by Goyal et al. [55]. The results show that ASIM takes less computational time and consumes less memory compared to CELF++ and TIM, while achieving the comparable influence spread.
- **EaSyIm** Galhotra et al. [47] proposed *opinion cum interaction* (OCI) model, which considers negative opinions as well. Based on the OCI Model, they formulated the *maximizing effective opinion* problem and proposed two fast and scalable heuristics, namely Opinion Spread Influence Maximization (OSIM) and EaSyIm having the running time of $\mathcal{O}(k\mathcal{D}(m+n))$ for this problem, where $\mathcal{D}$ is the diameter of the graph. Both the algorithms work in two phases. In the first phase, each node is assigned with some score based on the contribution on influence spread for all the paths starting at that node. The second step is concerned with the node processing step. The nodes with the maximum score value are selected as seed nodes. The reported empirical results show that OSIM and EaSyIm can achieve better influence spread compared to TIM$^+$, CELF++ with less running time.
- **Cordasco et al.'s** [31,33] **method** Later Cordasco et al. proposed a fast and effective heuristic method for selecting the target set in a undirected social network [31,33]. This heuristic produces optimal solution for *trees*, *cycles* and *complete graphs*. However, for real-life social networks, this heuristic performs much better than the other methods available in the literature. They extended this work for directed social networks as well [32].

There are several other studies also, which focused on developing heuristics. Nguyen et al. [92] proposed an efficient heuristic for solving BIM Problem. Wu et al. [136] developed a two-stage stochastic programming approach for solving SIM Problem. In this study, instead of choosing a seed set of size exactly k, their problem is choosing a seed set of size less than or equal to k.

Table 6 Heuristic solutions for SIM Problem

Name of the algorithm	Proposed by	Complexity	Model
SIMPATH	Goyal et al. [55]	$\mathcal{O}(kmn\mathcal{R})$	LT
SPIN	Narayanam et al. [89]	$\mathcal{O}(t(n+m)\mathcal{R}+n\log n+kn+k\mathcal{R}m)$	IC and LT
LDGA	Chen et al. [20]	$\mathcal{O}(n^2+kn^2\log n)$	MIA
IRIE	Jung et al. [62]	–	IC and IC-N
ASIM	Galhotra et al. [46]	$\mathcal{O}(kd(m+n))$	IC
EaSyIm	Galhotra et al. [47]	$\mathcal{O}(k\mathcal{D}(m+n))$	OI

Now, the studies related to heuristic methods will be summarized here. Centrality-based heuristics (CBHs) consider the topology of the network only and hence obtained influence spread in most of the cases is quite less compared to that of other state-of- the-art methods. However, DDH performs slightly better than other CBHs, as it puts a little restriction on the selection of two adjacent nodes. The application of SIMPATH for seed selection is little advantageous, as it has a user-controlled parameter η to balance the trade-off between accuracy and running time. SPIN has the advantage, as it can be used for solving both Top-k node problem and λ-coverage problem. MIA and PMIA have the better scalability compared to Basic Greedy. As LDAG works based on the principle of computation of influence spread in DAGs, it is seen to be faster. As various heuristics are experimented with different benchmark data sets, drawing a general conclusion about the performance will be difficult. Here, we have summarized some of the important algorithms for solving SIM and related problems, as presented in Table 6. In the case of the algorithms for which complexity analysis has not been done in the paper, the corresponding column is left empty in the table. Pros and cons of the heuristic solutions for the SIM and related problem are summarized in Table 7.

6.3 Metaheuristic solution approaches

Since early seventies, metaheuristic algorithms had been used successfully to solve optimization problems arisen in the broad domain of science and engineering [141,142]. There is no exception for solving SIM Problem as well.

- Bucur et al. [12] solved the SIM Problem using a *genetic algorithm*. They demonstrated that with a simple genetic operator, it is possible to find out an approximate solution for influence spread within feasible run time. In most of the cases, influence spread obtained by their method was comparable with that of the Basic Greedy algorithm proposed by Kempe et al. [67].
- Jiang et al. [61] proposed a *simulated annealing*-based algorithm for solving the SIM Problem under IC Model. The reported results indicate that their proposed methodology runs 2–3 times faster compared to the heuristic methods existing in the literature.
- Tsai et al. [122] developed the *Genetic New Greedy Algorithm* (GNA) for solving SIM Problem under IC Model by combining the genetic algorithm with the new greedy algorithm proposed by Chen et al. [19]. Their reported results conclude that GNA can give 10% more influence spread compared to the genetic algorithm.
- Gong et al. [51] proposed a *discrete particle swarm optimization algorithm* for solving SIM Problem. They used the degree discount heuristic proposed by Chen et al. [19] to initialize the seed set and *local influence estimation (LIE) function* to approximate the

Table 7 Advantages and disadvantages of Heuristic solutions for SIM Problem

Name of the algorithm	Advantages	Disadvantages
Random Heuristic	Highly scalable	Does not provide approximation guarantee
	Easy to implement	Poor solution quality
Centrality-Based Heuristics	Highly scalable	Does not provide approximation guarantee
	Easy to implement	Poor solution quality
Degree Discount Heuristic (DDH)	Highly scalable	Does not provide approximation guarantee
	Easy to implement	Seed set quality is not good enough to obtain acceptable level of influence
	Seed set quality is better than random and centrality-based heuristics	
SIMPATH	Scalability has been improved significantly.	Does not provide approximation guarantee
	Seed set quality is also better	Works only in LT Model of diffusion
		Algorithm comprises of parameters and its effective performance will depend on the proper tuning of the parameter
SPIN	First study of the SIM Problem under Co-Operative game theoretic setting	Does not provide approximation guarantee
		Quality of the seed set is dependent on how many samples are used to compute the 'shapley value'
LDGA	Scalability has been improved significantly	Does not provide approximation guarantee
		Works only in LT Model of diffusion
IRIE	Also works in the presence of negative influence	Does not provide approximation guarantee
	Scalability has been improved significantly	
ASIM	Scalability has been improved significantly	Does not provide approximation guarantee
EaSyIm	Considers the presence of negative influence	Does not provide approximation guarantee
Cordasco et al.'s Method	Provides optimal seed set for trees and cycles	Does not provide approximation guarantee
	Scalability is not enough to work with large-scale datasets	

two-hop influence. They introduced the *network specific local search* strategy also for fast convergence of their proposed algorithm. The reported results conclude that this methodology outperforms the state-of-the-art CELF++ with less computational time.

After that, several studies were also carried out in this direction [86,109,128,145]. The main advantages of these methods that are the computational time requirement for selecting seed

nodes are acceptable. However, these methods do not provide any approximation guarantee. Though there are a large number of metaheuristic algorithms [140], only a few had been used for solving SIM Problem. Hence, the use of metaheuristic algorithms for solving SIM Problem and its variants has been largely ignored. Next, we have described the community-based solution methodologies for SIM Problem.

6.4 Community-based solution approaches

Most of the real-life social networks exhibit a community structure within it [29]. A community is basically a subset of nodes, which are densely connected among themselves and sparsely connected with the other nodes of the network. In recent years, *community-based solution framework* (**CBSF**) has been developed for solving SIM Problem.

- Wang et al. [130] proposed the *community-based greedy algorithm* for solving SIM Problem. This method consists of two steps, namely detecting communities based on information propagation and selecting communities for finding influential nodes. This algorithm could outperform the degree discount and random heuristic.
- Chen et al. [23,24] developed a CBSF for solving SIM Problem and named it CIM. By exploiting the community structure, they selected some candidate seed sets for each community and from the candidate seed sets, they have selected the final seed set for diffusion. CIM could achieve the better influence spread compared to some state-of-the-art heuristic methods, such as CDH-Kcut, CDH-SHRINK and maximum degree.
- Rahimkhan et al. [102] proposed a CBSF for solving SIM Problem under LT Model and named it ComPath. They used Speaker-Listener Label Propagation Algorithm (SLPA) proposed by Xie et al. [137] for detecting communities and then identified the most influential communities and candidate seed nodes. From the candidate seed set, they selected the final seed set based on the intra distance among the nodes of the candidate seed set. ComPath could outperform CELF, CELF++, maximum degree heuristic, maximum pagerank heuristic and LDGA.
- Bozorgi et al. [10] developed a CBSF for solving SIM Problem under LT Model and named it INCIM. Like ComPath, INCIM also uses the SLPA Algorithm for detecting the communities. They proposed an algorithm for selecting seed, which computes the influence spread using the algorithm developed by Goyal et al. [55]. INCIM could outperform some state-of-the-art methodologies like LDGA, SIMPATH, IPA (a parallel algorithm for SIM Problem proposed by [71]), high pagerank and high-degree heuristic.
- Shang et al. [111] proposed a CBSF for solving SIM Problem and named it CoFIM. In this study, they introduced a diffusion model, which works in two phases. In the first phase, the seed set $\mathcal{S}$ was expanded to the neighbor nodes of $\mathcal{S}$, which would be usually allocated into different communities. Then, in the second phase, influence propagation within the communities was computed. Based on this diffusion model, they developed an incremental greedy algorithm for selecting seed set, which is analogous to the algorithm proposed by Kempe et al. [67]. CoFIM could achieve the better influence spread compared to that of IPA, TIM+, MDH and IMM.
- Recently, Li et al. [81] proposed a community-based approach for solving the SIM Problem, where the users have a specific geographical location. They developed a social influence-based community detection algorithm using spectral clustering technique and a seed selection methodology by considering community-based influence index. The reported results show that this methodology is more efficient than many state-of-the-art methodologies, while achieving almost the same influence spread.

Table 8 Advantages and disadvantages of different community-based solution methodologies

Name of the algorithm	Advantages	Disadvantages
Wang et al.'s Method	First study toward the community-based solutions for the SIM Problem	Does not provide any approximation guarantee
CIM	Scalability is more compared to many centrality-based heuristics	Does not provide any approximation guarantee
		There are few parameters. Effective performance of this algorithm depends upon the proper tuning of the parameters
ComPath	Scalability is more compared to many centrality-based heuristics, CELF, and LDGA	Does not provide approximation guarantee
		Works only for LT Model
INCIM	Scalability has been improved compared to many centrality-based heuristics, LDGA, and Simpath	Does not provide approximation guarantee
		Works only for LT Model
CoFIM	High scalability	Does not provide approximation guarantee
	Better influence spread compared to many other community-based methods	
Li et al.'s Method	Considers 'geographical preference' in influence maximization queries	Does not provide approximation guarantee

Recently, there are many studies on the influence maximization problem by exploiting the community structure considering the links of the network may be unknown initially [132], non-submodular influence function [2], etc. It is important to note that except the methodology proposed by Wang et al. [130], all these methods are basically heuristics. However, these methods use community detection of the underlying social network as an intermediate step to scale down the SIM Problem into community level. Pros and cons of these methods are summarized in Table 8. There are a large number of algorithms available in the literature for detecting communities [15,44]. Among them, which one should be used for solving SIM Problem? How is the quality of community detection and influence spread related? This questions are largely ignored in the literature.

6.5 Miscellaneous

In this section, we have described some solution methodologies of SIM Problem, which are very different from the methodologies discussed till now. Also, each solution methodology presented here is different from another. It is reported in the literature that in any information diffusion process, less than 10% nodes are influenced beyond the hop count 2 [50]. Based on this phenomenon, recently, Tang et al. [116,117] developed a hop-based approach for SIM Problem. Their methodology also gives a theoretical guarantee on influence spread. Ma et al. [87] proposed an algorithm for SIM Problem, which works based on the heat diffusion process. It could produce better influence spread compared to Basic Greedy algorithm. Goyal et al. [53] developed a data-based approach for solving SIM Problem. They introduced the

credit distribution (CD) model that could grip the propagation traces to learn the influence flow pattern for approximating the influence spread. They showed that SIM Problem under CD Model is NP-hard and the reported results show that this model can achieve even better influence spread compared to IC and LT Models with less running time. Lee et al. [77] introduced a query-based approach for solving SIM Problem under IC Model. Here, the query is: What should be the seed set for activating all the users of a given set $\mathcal{T}$? This methodology is intended for maximizing the influence of a particular group of users, which is the case in *target-aware viral marketing*. Zhu et al. [146] introduced the CTMC-ICM diffusion model, which is basically the blending of IC Model with *Continuous Time Markov Chain*. They studied the SIM Problem under this model and came up with a new centrality metric *Spread Rank*. Their reported results show that the seed nodes selected based on spread rank centrality can achieve better influence spared compared to the traditional distance-based centrality measures, such as *degree*, *closeness* or *betweenness*. Wang et al. [127] proposed the methodology Fluidspread, which works based on fluid dynamic principle and can reveal the dynamics of diffusion process. Kang et al. [63] introduced the notion of diffusion centrality for selecting influential nodes in a social network.

7 Summary of the survey and future research directions

Based on the survey of the existing literature presented in Sects. 3 through 6, we have summarized the current research trends and given future directions in this section.

7.1 Current research trends

- **Practicality of the problem** Most of the current studies are focused on the practical issues of the SIM Problem. One of the major applications of social influence maximization is viral marketing. So, in this context, influencing a user will be beneficial, only if he/she will be able to influence a reasonable number of other users of the network. Recent studies, such as [93,96] along with the node selection cost, also consider *benefit* as another component in the SIM Problem. Recently, Banerjee et al. [6] studied the influence maximization problem under the utility driven influence diffusion model. As, this direction is completely new, many interesting works can be done.
- **Scalability** Starting from Kempe et al.'s [67] seminal work, scalability remains an important issue in this area. To reduce scalability problem, instead of using Monte Carlo simulation-based spread estimation, recently Borgs et al. [9] introduced reverse reachable set-based spread estimation. After this work, all the popular algorithms for SIM Problem, such as TIM, IMM and TIM+, use this concept as an influence spread estimation technique for improving scalability.
- **Diffusion probability computation** TSS problem assumes that influence probability between any pair of users is known. However, this is a very unrealistic assumption. Though there were some previous studies in this direction, people tried to predict influence probability using machine learning techniques [125].

Though since the last one and half decades or so, the *TSS Problem* had been studied extensively from both theoretical and applied contexts, still to the best of our knowledge, some of the corners of this problem are either not or partially investigated. Here, we have listed some future research directions from both problem specification and solution methodology points of view.

7.2 Future directions

Further research may be carried out in future in and around of TSS Problem of social networks, in the following directions:

7.2.1 Problem specific

- As on-line social networks are formed by the rational agents, incentivization is required, if a node is selected as a seed node. For practical applications, it is also important to consider what benefit will be obtained (e.g., how many other non-seed nodes becoming influenced through that node, etc.) by activating that node. At the same time, for influence propagation of time-sensitive events (where influencing one person after an event does not make any scene such as political campaign before election, viral marketing for a seasonal product, etc.), consideration of diffusion time is also important. To the best of our knowledge, there is no reported study on TSS Problem considering all three issues: *cost, benefit, and time*.
- Most of the studies done on SIM Problem, and its variants are under either IC or LT diffusion model. However, recently, some other diffusion models have also been recently developed, such as Independent Cascade Model with Negative Opinion (IC-N) [22], Opinion Cum Interaction Model (OI) [47] and Opinion-based Cascading Model (OC) [143] which consider negative opinions. SIM Problems and its different variants can also be studied under these newly developed diffusion models.
- Most of the studies done on SIM Problem consider that the underlying social network is static including influence probabilities. However, this is not a practical assumption, as most of the social networks are time varying. Recent studies on SIM Problem started considering temporal nature of the social network [120,148]. As this has just started, there is a lot of scope to work in TSS Problem in time-varying social networks.
- In real-world social networks, users have specific topics of choice. So, one user will be influenced by another user, if both of them have similar choices. Keeping 'topic' into consideration the spread of influence can be increased, which is known as *topic aware influence maximization*. Recent studies on influence maximization consider this phenomenon [18,83]. SIM Problem and its variants can be studied in this settings as well.

7.2.2 Solution methodology specific

- Among all the variants of TSS Problem in social networks described in Sect. 3, it is surprising to see that only SIM Problem is well studied. Hence, solution methodologies developed for SIM Problem can be modified accordingly, so that they can be adopted for solving other variants of SIM Problem as well.
- One of the major issues in the solution methodology for SIM Problem is the scalability. It is important to observe that the social network used in the Kempe et al.'s [67] experiment had 10,748 nodes and 53,000 edges, whereas the recent study of Nguyen et al.'s [96] has used a social network of with 41.7×10^6 nodes and 1.5×10^9 edges. From this example, it is clear that the size of the social network data sets is increasing day by day. Hence, developing more scalable algorithms is extremely important to handle large data sets.
- From the discussion in Sect. 6.3, it is understood that though there are many evolutionary algorithms, only genetic algorithm, artificial bee colony optimization and discrete particle

swarm optimization algorithms have been used till date for solving SIM Problem. Hence, other meta-heuristics, such as *ant colony optimization* and *differential evolution* can also be used for this purpose.
- There are many solution methodologies proposed in the literature. However, which one to choose in which situation and for what kind of network structure? For answering this question, by taking all the proposed methodologies from the literature, a strong experimental evaluation is required with benchmark data sets. Recently, Arora et al. [3] has done a benchmarking study with 11 most popular algorithms from the literature, and they have found some contradictions between their own experimental results and reported ones in the literature. More such benchmarking studies are required to investigate these issues.
- Most of the algorithms presented in the literature are serial in nature. The issue of scalability in SIM Problem can be tackled by developing distributed and parallel algorithms. To the best of the authors' knowledge, except dIRIEr developed by Zong et al. [149], there are no distributed algorithms existing in the literature. Recently, a few parallel algorithms have been developed for SIM Problem [71,135]. So, this is an open area to study the SIM Problem and its variants under parallel and distributed settings.
- Most of the solution methodologies are concerned with the selection of the seeds in one move, before the diffusion starts. In this case, if any one of the selected seeds does not perform up to expectation, then the number of influenced nodes will be less than expected. Considering this case, recently the framework of multiphase diffusion has been developed [35,57]. Different variants of this problem can be studied in this framework.

8 Concluding remarks

In this survey, at first we have discussed the SIM Problem and its different variants studied in the literature. Next, we have reported the hardness results of the problem. After that, we have reported major research challenges concerned with the SIM Problem and its variants. Subsequently, based on the approach, we have classified the proposed solution methodologies and discussed algorithms of each category. At the end, we have discussed the current research trends and given future directions. From this survey, we can conclude that SIM Problem is well-studied, though its variants are not and there is a continuous thirst for developing more scalable algorithm for these problems. We hope that presenting three dimensions (variants, hardness results and solution methodologies all together) of the problem will help the researchers and practitioners to have a proper understanding of the problem and better exposure in this field.

Acknowledgements The authors want to thank Ministry of Human Resource and Development (MHRD), Government of India, for sponsoring the project: E-business Center of Excellence under the scheme of Center for Training and Research in Frontier Areas of Science and Technology (FAST), Grant No. F.No.5-5/2014-TS.VII.

References

1. Ackerman E, Ben-Zwi O, Wolfovitz G (2010) Combinatorial model and bounds for target set selection. Theor Comput Sci 411(44–46):4017–4022

2. Angell R, Schoenebeck G (2017) Dont be greedy: leveraging community structure to find high quality seed sets for influence maximization. In: International conference on web and internet economics. Springer, pp 16–29
3. Arora A, Galhotra S, Ranu S (2017) Debunking the myths of influence maximization: an in-depth benchmarking study. In: Proceedings of the 2017 ACM international conference on management of data. ACM, pp 651–666
4. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM, pp 65–74
5. Balogh J, Bollobás B, Morris R (2010) Bootstrap percolation in high dimensions. Comb Probab Comput 19(5–6):643–692
6. Banerjee P, Chen W, Lakshmanan LV (2019) Maximizing welfare in social networks under a utility driven influence diffusion model. In: Proceedings of the 2019 international conference on management of data. ACM, pp 1078–1095
7. Banerjee S, Mathew R (2018) An inapproximability result for the target set selection problem on bipartite graphs. arXiv preprint arXiv:1812.01482
8. Bazgan C, Chopin M, Nichterlein A, Sikora F (2014) Parameterized approximability of maximizing the spread of influence in networks. J Discrete Algorithms 27:54–65
9. Borgs C, Brautbar M, Chayes J, Lucier B (2014) Maximizing social influence in nearly optimal time. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms. SIAM, pp 946–957
10. Bozorgi A, Haghighi H, Zahedi MS, Rezvani M (2016) Incim: A community-based algorithm for influence maximization problem under the linear threshold model. Inf Process Manag 52(6):1188–1199
11. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117
12. Bucur D, Iacca G (2016) Influence maximization in social networks with genetic algorithms. In: European conference on the applications of evolutionary computation. Springer, pp 379–392
13. Campbell WM, Dagli CK, Weinstein CJ (2013) Social network analysis with content and graphs. Linc Lab J 20(1):61–81
14. Centola D (2010) The spread of behavior in an online social network experiment. Science 329(5996):1194–1197
15. Chakraborty T, Dalmia A, Mukherjee A, Ganguly N (2017) Metrics for community analysis: a survey. ACM Comput Surv (CSUR) 50(4):54
16. Charikar M, Naamad Y, Wirth A (2016) On approximating target set selection. In: LIPIcs-Leibniz international proceedings in informatics, vol 60. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
17. Chen N (2009) On the approximability of influence in social networks. SIAM J Discrete Math 23(3):1400–1415
18. Chen S, Fan J, Li G, Feng J, Kl Tan, Tang J (2015) Online topic-aware influence maximization. Proc VLDB Endow 8(6):666–677
19. Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 199–208
20. Chen W, Wang C, Wang Y (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. ACM, pp 1029–1038
21. Chen W, Yuan Y, Zhang L (2010) Scalable influence maximization in social networks under the linear threshold model. In: 2010 IEEE 10th international conference on data mining (ICDM). IEEE, pp 88–97
22. Chen W, Collins A, Cummings R, Ke T, Liu Z, Rincon D, Sun X, Wang Y, Wei W, Yuan Y (2011) Influence maximization in social networks when negative opinions may emerge and propagate. In: Proceedings of the 2011 SIAM international conference on data mining. SIAM, pp 379–390
23. Chen Y, Chang S, Chou C, Peng W, Lee S (2012) Exploring community structures for influence maximization in social networks. In: Proceedings of the 6th SNA-KDD workshop on social network mining and analysis held in conjunction with KDD12 (SNA-KDD12), pp 1–6
24. Chen YC, Zhu WY, Peng WC, Lee WC, Lee SY (2014) Cim: community-based influence maximization in social networks. ACM Trans Intell Syst Technol (TIST) 5(2):25
25. Cheng S, Shen H, Huang J, Zhang G, Cheng X (2013) Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In: Proceedings of the 22nd ACM international conference on information & knowledge management. ACM, pp 509–518
26. Chopin M, Nichterlein A, Niedermeier R, Weller M (2012) Constant thresholds can make target set selection tractable. Springer, Berlin, pp 120–133

27. Chopin M, Nichterlein A, Niedermeier R, Weller M (2014) Constant thresholds can make target set selection tractable. Theory Comput Syst 55(1):61–83
28. Cicalese F, Cordasco G, Gargano L, Milanič M, Vaccaro U (2014) Latency-bounded target set selection in social networks. Theor Comput Sci 535:1–15
29. Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111
30. Cohen E, Delling D, Pajor T, Werneck RF (2014) Sketch-based influence maximization and computation: scaling up with guarantees. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. ACM, pp 629–638
31. Cordasco G, Gargano L, Mecchia M, Rescigno AA, Vaccaro U (2015a) A fast and effective heuristic for discovering small target sets in social networks. In: Combinatorial optimization and applications. Springer, pp 193–208
32. Cordasco G, Gargano L, Rescigno AA (2015b) Influence propagation over large scale social networks. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. ACM, pp 1531–1538
33. Cordasco G, Gargano L, Rescigno AA (2016) Active spreading in networks. In: ICTCS, pp 149–162
34. Cowan R, Jonard N (2004) Network structure and the diffusion of knowledge. J Econ Dyn Control 28(8):1557–1575
35. Dhamal S, Prabuchandran K, Narahari Y (2016) Information diffusion in social networks in two phases. IEEE Trans Netw Sci Eng 3(4):197–210
36. Diestel R (2005) Graph theory. 2005. Grad Texts in Math 101
37. Domingos P, Richardson M (2001) Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 57–66
38. Downey RG, Fellows MR (2012) Parameterized complexity. Springer, Berlin
39. Downey RG, Fellows MR (2013) Fundamentals of parameterized complexity, vol 4. Springer, Berlin
40. Downey RG, Fellows MR, Regan KW (1998) Parameterized circuit complexity and the W hierarchy. Theor Comput Sci 191(1–2):97–115
41. Dreyer PA, Roberts FS (2009) Irreversible k-threshold processes: graph-theoretical threshold models of the spread of disease and of opinion. Discrete Appl Math 157(7):1615–1627
42. Epasto A, Mahmoody A, Upfal E (2017) Real-time targeted-influence queries over large graphs. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017. ACM, pp 224–231
43. Feige U, Goemans M (1995) Approximating the value of two power proof systems, with applications to max 2sat and max dicut
44. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174
45. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
46. Galhotra S, Arora A, Virinchi S, Roy S (2015) Asim: a scalable algorithm for influence maximization under the independent cascade model. In: Proceedings of the 24th international conference on world wide web. ACM, pp 35–36
47. Galhotra S, Arora A, Roy S (2016) Holistic influence maximization: combining scalability and efficiency with opinion-aware models. In: Proceedings of the 2016 international conference on management of data. ACM, pp 743–758
48. Garey MR, Johnson DS (2002) Computers and intractability, vol 29. W. H. Freeman, New York
49. Gionis A, Terzi E, Tsaparas P (2013) Opinion maximization in social networks. In: Proceedings of the 2013 SIAM international conference on data mining. SIAM, pp 387–395
50. Goel S, Watts DJ, Goldstein DG (2012) The structure of online diffusion networks. In: Proceedings of the 13th ACM conference on electronic commerce. ACM, pp 623–638
51. Gong M, Yan J, Shen B, Ma L, Cai Q (2016) Influence maximization in social networks based on discrete particle swarm optimization. Inf Sci 367:600–614
52. Goyal A, Bonchi F, Lakshmanan LV (2010) Learning influence probabilities in social networks. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp 241–250
53. Goyal A, Bonchi F, Lakshmanan LV (2011a) A data-based approach to social influence maximization. Proc. VLDB Endow. 5(1):73–84
54. Goyal A, Lu W, Lakshmanan LV (2011b) Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 20th international conference companion on world wide web. ACM, pp 47–48

55. Goyal A, Lu W, Lakshmanan LV (2011c) Simpath: an efficient algorithm for influence maximization under the linear threshold model. In: 2011 IEEE 11th international conference on data mining (ICDM). IEEE, pp 211–220
56. Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: Proceedings of the 13th international conference on world wide web. ACM, pp 491–501
57. Han K, Huang K, Xiao X, Tang J, Sun A, Tang X (2018) Efficient algorithms for adaptive influence maximization. In: Proceedings of the VLDB endowment, vol 11, no 9
58. Harant J, Pruchnewski A, Voigt M (1999) On dominating sets and independent sets of graphs. Comb Probab Comput 8(6):547–553
59. Heidari N (2016) Modeling information diffusion in social networks. arXiv preprint arXiv:1603.02178
60. Ienco D, Bonchi F, Castillo C (2010) The meme ranking problem: maximizing microblogging virality. In: 2010 IEEE international conference on data mining workshops (ICDMW). IEEE, pp 328–335
61. Jiang Q, Song G, Cong G, Wang Y, Si W, Xie K (2011) Simulated annealing based influence maximization in social networks. In: AAAI, vol 11, pp 127–132
62. Jung K, Heo W, Chen W (2012) Irie: scalable and robust influence maximization in social networks. In: 2012 IEEE 12th international conference on data mining (ICDM). IEEE, pp 918–923
63. Kang C, Kraus S, Molinaro C, Spezzano F, Subrahmanian V (2016) Diffusion centrality: a paradigm to maximize spread in social networks. Artif Intell 239:70–96
64. Karp RM (1972) Reducibility among combinatorial problems. In: Complexity of computer computations. Springer, pp 85–103
65. Kasprzak R (2012) Diffusion in networks. J Telecommun Inf Technol 99–106
66. Ke X, Khan A, Cong G (2018) Finding seeds and relevant tags jointly: for targeted influence maximization in social networks. In: Proceedings of the 2018 international conference on management of data. ACM, pp 1097–1111
67. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 137–146
68. Kempe D, Kleinberg JM, Tardos É (2005) Influential nodes in a diffusion model for social networks. In: ICALP, vol 5. Springer, pp 1127–1138
69. Kempe D, Kleinberg JM, Tardos É (2015) Maximizing the spread of influence through a social network. Theory Comput 11(4):105–147
70. Khuller S, Moss A, Naor JS (1999) The budgeted maximum coverage problem. Inf Process Lett 70(1):39–45
71. Kim J, Kim SK, Yu H (2013) Scalable and parallelizable processing of influence maximization for large-scale social networks? In: 2013 IEEE 29th international conference on data engineering (ICDE). IEEE, pp 266–277
72. Kimura M, Saito K (2006) Tractable models for information diffusion in social networks. In: Knowledge discovery in databases: PKDD 2006, pp 259–271
73. Kimura M, Saito K, Nakano R, Motoda H (2009) Finding influential nodes in a social network from information diffusion data. In: Social computing and behavioral modeling, pp 1–8
74. Klasing R, Laforest C (2004) Hardness results and approximation algorithms of k-tuple domination in graphs. Inf Process Lett 89(2):75–83
75. Kortsarz G (2001) On the hardness of approximating spanners. Algorithmica 30(3):432–450
76. Landherr A, Friedl B, Heidemann J (2010) A critical review of centrality measures in social networks. Bus Inf Syst Eng 2(6):371–385
77. Lee JR, Chung CW (2015) A query approach for influence maximization on specific users in social networks. IEEE Trans Knowl Data Eng 27(2):340–353
78. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, pp 177–187
79. Leskovec J, Adamic LA, Huberman BA (2007a) The dynamics of viral marketing. ACM Trans Web (TWEB) 1(1):5
80. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007b) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 420–429
81. Li X, Cheng X, Su S, Sun C (2018a) Community-based seeds selection algorithm for location aware influence maximization. Neurocomputing 275:1601–1613
82. Li Y, Chen W, Wang Y, Zhang ZL (2013) Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, pp 657–666

83. Li Y, Zhang D, Tan KL (2015) Real-time targeted influence maximization for online advertisements. Proc VLDB Endow 8(10):1070–1081
84. Li Y, Fan J, Wang Y, Tan KL (2018b) Influence maximization on social graphs: a survey. IEEE Trans Knowl Data Eng 30:1852–1872
85. Liu B (2011) Social network analysis. In: Web data mining. Springer, Berlin, pp 269–309
86. Liu SJ, Chen CY, Tsai CW (2017) An effective simulated annealing for influence maximization problem of online social networks. Proc Comput Sci 113:478–483
87. Ma H, Yang H, Lyu MR, King I (2008) Mining social networks using heat diffusion processes for marketing candidates selection. In: Proceedings of the 17th ACM conference on information and knowledge management. ACM, pp 233–242
88. Maehara T, Suzuki H, Ishihata M (2017) Exact computation of influence spread by binary decision diagrams. In: Proceedings of the 26th international conference on world wide web, international world wide web conferences steering committee, pp 947–956
89. Narayanam R, Narahari Y (2011) A shapley value-based approach to discover influential nodes in social networks. IEEE Trans Autom Sci Eng 8(1):130–147
90. Nekovee M, Moreno Y, Bianconi G, Marsili M (2007) Theory of rumour spreading in complex social networks. Phys A 374(1):457–470
91. Nguyen H, Zheng R (2012) On budgeted influence maximization in social networks. arXiv preprint arXiv:1204.4491
92. Nguyen H, Zheng R (2013) On budgeted influence maximization in social networks. IEEE J Sel Areas Commun 31(6):1084–1094
93. Nguyen HT, Dinh TN, Thai MT (2016a) Cost-aware targeted viral marketing in billion-scale networks. In: IEEE INFOCOM 2016-the 35th annual IEEE international conference on computer communications. IEEE, pp 1–9
94. Nguyen HT, Thai MT, Dinh TN (2016b) Stop-and-stare: optimal sampling algorithms for viral marketing in billion-scale networks. In: Proceedings of the 2016 international conference on management of data. ACM, pp 695–710
95. Nguyen HT, Ghosh P, Mayo ML, Dinh TN (2017) Social influence spectrum at scale: near-optimal solutions for multiple budgets at once. ACM Trans Inf Syst (TOIS) 36(2):14
96. Nguyen HT, Thai MT, Dinh TN (2017) A billion-scale approximation algorithm for maximizing benefit in viral marketing. IEEE/ACM Trans Netw 25:2419–2429
97. Nichterlein A, Niedermeier R, Uhlmann J, Weller M (2010) On tractable cases of target set selection. In: Algorithms and computation, pp 378–389
98. Nichterlein A, Niedermeier R, Uhlmann J, Weller M (2013) On tractable cases of target set selection. Soc Netw Anal Min 3(2):233–256
99. Peleg D (2002) Local majorities, coalitions and monopolies in graphs: a review. Theor Comput Sci 282(2):231–257
100. Peng S, Zhou Y, Cao L, Yu S, Niu J, Jia W (2018) Influence analysis in social networks: a survey. J Netw Comput Appl 106:17–32
101. Raghavan S, Zhang R (2015) Weighted target set selection on social networks. Technical report, Working paper, University of Maryland
102. Rahimkhani K, Aleahmad A, Rahgozar M, Moeini A (2015) A fast algorithm for finding most influential people based on the linear threshold model. Expert Syst Appl 42(3):1353–1361
103. Raman V, Saurabh S, Srihari S (2008) Parameterized algorithms for generalized domination. Lect Notes Comput Sci 5165:116–126
104. Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 61–70
105. Saito K, Nakano R, Kimura M (2008) Prediction of information diffusion probabilities for independent cascade model. In: Knowledge-based intelligent information and engineering systems. Springer, pp 67–75
106. Saito K, Kimura M, Ohara K, Motoda H (2010) Selecting information diffusion models over social networks for behavioral analysis. In: Machine learning and knowledge discovery in databases, pp 180–195
107. Saito K, Ohara K, Yamagishi Y, Kimura M, Motoda H (2011) Learning diffusion probability based on node attributes in social networks. In: International symposium on methodologies for intelligent systems. Springer, pp 153–162
108. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH (2010) A high-resolution human contact network for infectious disease transmission. Proc Nat Acad Sci 107(51):22020–22025

109. Sankar CP, Asharaf S, Kumar KS (2016) Learning from bees: an approach for influence maximization on viral campaigns. PLoS ONE 11(12):e0168125
110. Shakarian P, Bhatnagar A, Aleali A, Shaabani E, Guo R (2015) The independent cascade and linear threshold models. In: Diffusion in social networks. Springer, pp 35–48
111. Shang J, Zhou S, Li X, Liu L, Wu H (2017) Cofim: a community-based framework for influence maximization on large-scale networks. Knowl Based Syst 117:88–100
112. Song X, Tseng BL, Lin CY, Sun MT (2006) Personalized recommendation driven by information flow. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 509–516
113. Sun J, Tang J (2011) A survey of models and algorithms for social influence analysis. In: Social network data analytics. Springer, Berlin, pp 177–214
114. Sun L, Huang W, Yu PS, Chen W (2018) Multi-round influence maximization. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. ACM, pp 2249–2258
115. Tabak BM, Takami M, Rocha JM, Cajueiro DO, Souza SR (2014) Directed clustering coefficient as a measure of systemic risk in complex banking networks. Phys A 394:211–216
116. Tang J, Tang X, Yuan J (2017) Influence maximization meets efficiency and effectiveness: a hop-based approach. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017. ACM, pp 64–71
117. Tang J, Tang X, Yuan J (2018) An efficient and effective hop-based approach for influence maximization in social networks. Soc Netw Anal Min 8(1):10
118. Tang Y, Xiao X, Shi Y (2014) Influence maximization: near-optimal time complexity meets practical efficiency. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data. ACM, pp 75–86
119. Tang Y, Shi Y, Xiao X (2015) Influence maximization in near-linear time: a martingale approach. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data. ACM, pp 1539–1554
120. Tong G, Wu W, Tang S, Du DZ (2017) Adaptive influence maximization in dynamic social networks. IEEE/ACM Trans Netw (TON) 25(1):112–125
121. Tovey CA (1984) A simplified np-complete satisfiability problem. Discrete Appl Math 8(1):85–89
122. Tsai CW, Yang YC, Chiang MC (2015) A genetic newgreedy algorithm for influence maximization in social network. In: 2015 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 2549–2554
123. Valente TW (1995) Network models of the diffusion of innovations
124. Valente TW (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18(1):69–89
125. Varshney D, Kumar S, Gupta V (2017) Predicting information diffusion probabilities in social networks: a Bayesian networks based approach. Knowl Based Syst 133:66–76
126. Wang C, Chen W, Wang Y (2012) Scalable influence maximization for independent cascade model in large-scale social networks. Data Min Knowl Disc 25(3):545
127. Wang F, Jiang W, Li X, Wang G (2017a) Maximizing positive influence spread in online social networks via fluid dynamics. Future Gener Comput Syst 86:1491–1502
128. Wang Q, Gong M, Song C, Wang S (2017b) Discrete particle swarm optimization based influence maximization in complex networks. In: 2017 IEEE congress on evolutionary computation (CEC). IEEE, pp 488–494
129. Wang T, Chen Y, Zhang Z, Xu T, Jin L, Hui P, Deng B, Li X (2011) Understanding graph sampling algorithms for social network analysis. In: 2011 31st international conference on distributed computing systems workshops (ICDCSW). IEEE, pp 123–128
130. Wang Y, Cong G, Song G, Xie K (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. ACM, pp 1039–1048
131. Weng J, Lim EP, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp 261–270
132. Wilder B, Immorlica N, Rice E, Tambe M (2017) Influence maximization with an unknown network by exploiting community structure. In: SocInf@ IJCAI, pp 2–7
133. Williamson DP, Shmoys DB (2011) The design of approximation algorithms. Cambridge University Press, Cambridge
134. Wilson C, Boe B, Sala A, Puttaswamy KP, Zhao BY (2009) User interactions in social networks and their implications. In: Proceedings of the 4th ACM European conference on computer systems. ACM, pp 205–218

135. Wu H, Yue K, Fu X, Wang Y, Liu W (2016) Parallel seed selection for influence maximization based on k-shell decomposition. In: International conference on collaborative computing: networking, applications and worksharing. Springer, pp 27–36
136. Wu HH, Küçükyavuz S (2017) A two-stage stochastic programming approach for influence maximization in social networks. Comput Optim Appl 69:1–33
137. Xie J, Szymanski BK, Liu X (2011) Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: 2011 IEEE 11th international conference on data mining workshops (ICDMW). IEEE, pp 344–349
138. Xu B, Liu L (2010) Information diffusion through online social networks. In: 2010 IEEE international conference on emergency management and management sciences (ICEMMS). IEEE, pp 53–56
139. Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: 2010 IEEE 10th international conference on data mining (ICDM). IEEE, pp 599–608
140. Yang XS (2010) Nature-inspired metaheuristic algorithms. Luniver Press, Oxford
141. Yang XS, Chien SF, Ting TO (2014) Computational intelligence and metaheuristic algorithms with applications. Sci World J 2014:425853
142. Yi H, Duan Q, Liao TW (2013) Three improved hybrid metaheuristic algorithms for engineering design optimization. Appl Soft Comput 13(5):2433–2444
143. Zhang H, Dinh TN, Thai MT (2013) Maximizing the spread of positive influence in online social networks. In: 2013 IEEE 33rd international conference on distributed computing systems (ICDCS). IEEE, pp 317–326
144. Zhang H, Mishra S, Thai MT, Wu J, Wang Y (2014) Recent advances in information diffusion and influence maximization in complex social networks. Oppor Mobile Soc Netw 37(1.1):37
145. Zhang K, Du H, Feldman MW (2017) Maximizing influence in a social network: improved results using a genetic algorithm. Phys A 478:20–30
146. Zhu T, Wang B, Wu B, Zhu C (2014) Maximizing the spread of influence ranking in social networks. Inf Sci 278:535–544
147. Zhu Y, Wu W, Bi Y, Wu L, Jiang Y, Xu W (2015) Better approximation algorithms for influence maximization in online social networks. J Comb Optim 30(1):97–108
148. Zhuang H, Sun Y, Tang J, Zhang J, Sun X (2013) Influence maximization in dynamic social networks. In: 2013 IEEE 13th international conference on data mining (ICDM). IEEE, pp 1313–1318
149. Zong Z, Li B, Hu C (2014) dirier: distributed influence maximization in social network. In: 2014 20th IEEE international conference on parallel and distributed systems (ICPADS). IEEE, pp 119–125
150. Zou CC, Towsley D, Gong W (2007) Modeling and simulation study of the propagation and defense of internet e-mail worms. IEEE Trans Dependable Secure Comput 4(2):105–118

Suman Banerjee is currently affiliated with the Department of Computer Science and Engineering, IIT Gandhinagar, India, as a research associate. Prior to this, he was a PhD student at Indian Institute of Technology Kharagpur, India, and for his PhD dissertation, he has worked on the problem of 'social influence maximization.' Prior to that, he has received his M.Tech and B.Tech degrees from National Institute of Technology, Durgapur and Government College of Engineering and Textile Technology, Serampore, India. His broad area of research includes probabilistic data management, computational social choice, and theoretical computer science. He is a student member of ACM.

Mamata Jenamani is currently working in the Department of Industrial & Systems Engineering as a professor. She obtained her PhD degree from IIT Kharagpur and worked as a Post-doctoral Researcher in Purdue University. In 2005, she won an Emerald/EFMD Outstanding Doctoral Research Award in the category of Enterprise Applications of Internet Technology for her PhD. She has more than twenty years of experience. She is a Principal Investigator of E-Business Centre of Excellence funded by MHRD. Her areas of interest include developing models for web data analysis, design of recommender system, web log analysis, user generated content analysis and social network analysis, ebusiness, information systems, recommender systems, and supply chain optimization. Her current projects include a number of projects in the areas such as ebusiness in general, auction, ICT in supply chain and urban sustainability with a focus on egovernance in association with National Institute of Rural Development and Panchayat Raj, MHRD, CSIR. Currently, she guides seven PhD scholars and three MS students. Her publications have appeared in the leading journals such as European Journal of Operational Research, Electronic Commerce Research and Application, Expert System with Application, Annals of Operations Research, and Applied Mathematical Modeling. She is rendering her services as a reviewer to leading journals in the area of ebusiness.

Dilip Kumar Pratihar received his BE (Hons.) and M. Tech. from REC (NIT) Durgapur, India, in 1988 and 1994, respectively. He obtained his PhD from IIT Kanpur, India, in 2000. He received University Gold Medal, A.M. Das Memorial Medal, Institution of Engineers (I) Medal, and others. He completed his postdoctoral studies in Japan and then in Germany under the Alexander von Humboldt Fellowship Programme. He is working as a Professor at IIT Kharagpur, India. His research areas include robotics, soft computing and manufacturing science. He has published more than 190 papers, mostly in various international journals. He has written a textbook on soft computing, co-authored another textbook on analytical engineering mechanics, edited a book on intelligent and autonomous systems, co-authored reference books on modeling and analysis of six-legged robots and modeling and simulations of robotic systems using soft computing. Recently, he has published another textbook named Soft Computing: Fundamentals and Applications. He has guided 19 PhD candidates. He is in editorial boards of 14 international journals. He has been elected as FIE and MIEEE.