

Improved Biogeography-based Optimization Algorithm for Social Influence Maximization

Chandrakant Mallick¹, Parimal Kumar Giri^{1,*}, Sagar S De², Sarojananda Mishra³

¹ Bhubaneswar and BPUT,
GITA Autonomous College, Dept. of CSIT, Odisha,
India

² S.N. Bose National Centre for Basic Sciences,
India

³ Dhenkanal,
Dept. of CSE, IGIT Sarang,
India

{ckmallick, parimal.6789, sagar.s.de, sarose.mishra}@gmail.com

Abstract. The influence maximization issue in social network analysis determines a small selection of seed nodes that effectively improves aggregated influence in a hierarchical propagation model. The process of seed identification and find the best-known seeds are an NP-hard task. In this paper we have proposed novel idea that adding a few links to a small structural change can significantly boost cumulative influence. Therefore, to accomplish coverage as an aim, we used multi-objective meta-heuristic optimization to find initial seeds while taking aggregated influence and time step into account. The suggested approach then determined the lowest number of required missing links against every uninfluenced node to sustain propagation for acceptable non-dominated seeds. Finally, end vertices for the links recommendation are identified using meta-heuristic optimization approach known as locally and globally tuned Biogeography-based optimization. The suggested links connect non-influenced components to the influenced component and allow further influence propagation.

Keywords. Influence maximization, link recommendation, information diffusion, biogeography-based optimization, multi-objective optimization.

1 Introduction

Over the last couple decades, online social networks (OSNs) have become the de facto

standard for communication and expression. Information propagation in a social network system is based on cascading phenomena, also known as word-of-mouth effectiveness. Instagram, Twitter, Facebook, and other online social networking networks have arisen as natural choices for product promotion, marketing, camping, and opinion polls by businesses [29].

Researchers from a variety of fields have been drawn attention towards the social network phenomena and its widespread implementation in real-world applications [24]. The scientific and commercial worlds have embraced social influence analysis and network modeling in recent decades

Social influence analysis, in general, one of the most difficult tasks in complex network analysis [20]. In a cascade propagation model, the *influence maximization (IM)* problem identifies a small subset (i.e., a size of k) of seed nodes that effectively maximizes the aggregated influence [25]. The *independent cascade model (IC Model)* and the *linear threshold propagation model (LT model)* are the most employed models [6].

In general, a constrained IM problem is a multi-objective optimization problem. As a multi-objective optimizer, De and Dehury [4]

suggested Non-dominated sorting biogeography-based optimization (NSBBO). NSBBO successfully solved the multi-objective influence maximization problem. For the identification of a non-dominated collection of seed nodes, the non-dominated sorting genetic algorithm II (NSGA-II) is used [7]. All non-dominated seeds are equally optimum, regardless of objective choice. However, the best-known seeds are sometimes unable to propagate influence throughout the network due to the network topology and used propagation model [25]. In this paper the basic objectives are to maximize aggregated influence and minimize propagation time.

According to Chen et al. [2], social influence is defined as a change in an individual's thoughts, feelings, attitudes, or behaviors as a result of interactions with other people or groups. A small structural adjustment, like adding a few links, can greatly increase collective effect, as we've learned from the literature review. As far as we know, not a single study has been conducted in the field of structural modification to boost influence. In this research, we first choose some of the possible seed sets that are distinct from the network's uninfluenced area. The least number of lost links required to maintain propagation was then determined for each node that was unaffected.

To calculate the number of extra links each non-influenced node needs to become active, we have developed a model. An approach known as *locally and globally tuned biogeography-based optimization (LGBBO)* [11] is used to depict the smallest number of nodes that are effective at spreading influence inside the non-influenced component of the network. The identified nodes are the missing links' one end. Our approach uses a few statistical calculations to identify another end node of these links. The suggested links allow for greater influence propagation by joining influenced and uninfluenced parts of the network. The advantage of the two-stage method is that we gradually identified and tried to activate the network's non-influenced portion rather than randomly establishing missing links throughout it. Numerous empirical studies on a range of benchmark social networks have demonstrated the effectiveness of the suggested technique in finding

seed nodes and recommending missing linkages to enhance influence propagation.

The rest of the paper is organized as follows. In Sections 2 and 3 provides information on the literature study. The current research study is highlighted in Subsection 3.1. Sections 4 and 5 presents the recommended strategy to link suggestion for influence maximization, as well as relevant empirical research. Section 6 concludes the study.

2 Backgrounds

This section briefly reviews the concepts of social influence, influence propagation, and the problem of influence maximization before moving on to multi-objective optimization.

2.1 Influence Propagation, Social Influence, Influence Maximization Problem

Social influence can be shown in a variety of ways throughout OSNs. When it comes to data mining and big data analysis, social influence is measured in a variety of ways e.g., diffusion modeling, recommender systems, and viral marketing. After the first exposure, Ryan-Gross's [30] looked into the delayed adaptation. They've displayed the percentage of operators who have had their first hearing and the percentage of operators who take hybrid seed in Fig. 1 for specific years.

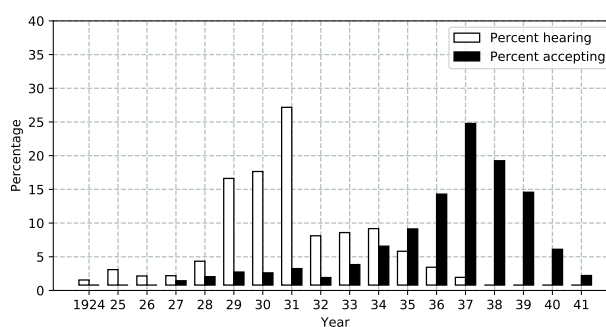


Fig. 1. Ryan-Gross study of hybrid seed corn delayed adaptation [30]

In 2003, at first Kempe et al. [20] proposed the influence maximization problem in the way of

algorithm. They have studied a social network represent a graph $G = (V, E)$, where V is the set of nodes (i.e., users), E is the set of (directed/undirected) edges in G (i.e., social links between users), and W represent a matrix of edge weight.

Each members $w_{i,j}$ in W denotes the weight of the edge between the node i and j , which represents the probability of node i activates node j . The goal of the IM issue is to find the initial group of users of number k who have the greatest effect in a graph G given a propagation model M . The method of knowledge diffusion among users determines the influence of any seed set.

Consider a graph $G = (V, E, W)$ be a social graph, a user set $S \subseteq V$, and a diffusion model M represent a stochastic process S for spreading information on G .

The influence spread of S , viz, $\sigma_{G,M}(S)$, is the expected number of users influenced by S (e.g., users who adopt the new product in viral marketing), provided $\sigma_{G,M}(\cdot)$ be a set of non-negative function such as $\sigma_{G,M} : 2V \rightarrow R \geq 0$ on any subset of users.

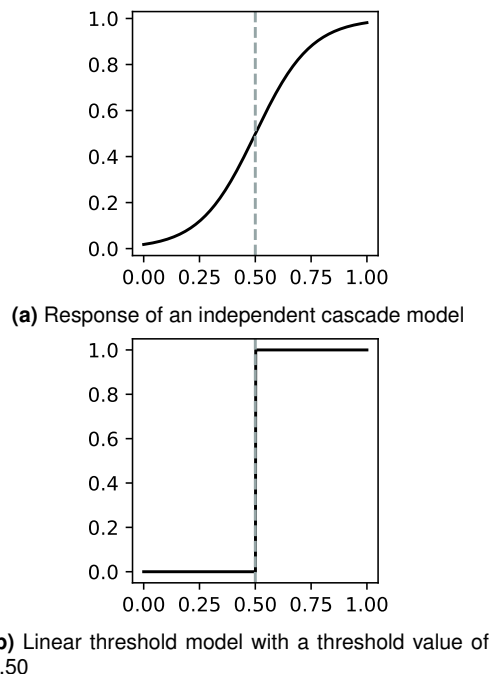
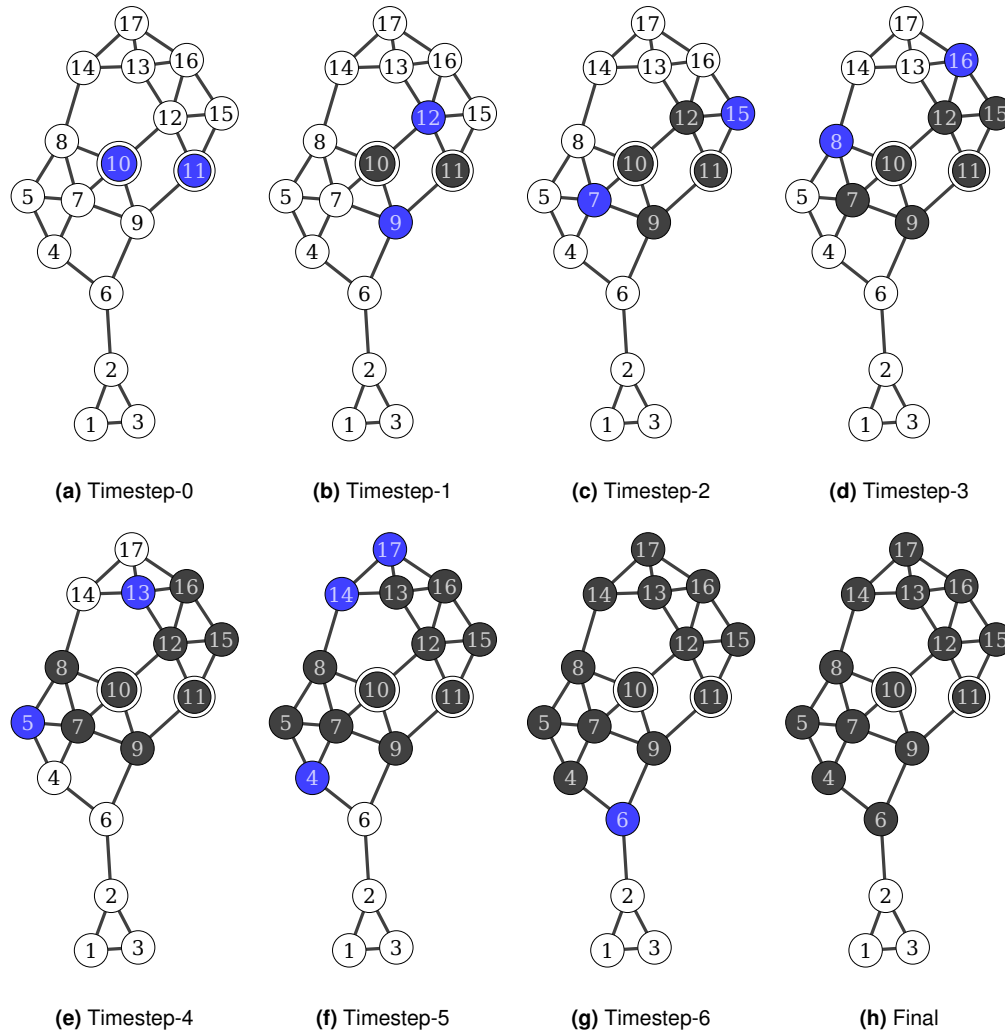


Fig. 2. Influence propagation response

Diffusion models have been developed in recent years to formulate the diffusion process and compute the influence spread [19, 32, 22, 21]. We focus on progressive diffusion models in this paper, which means that activated nodes cannot be deactivated. Recently, the economics and sociology communities have developed number of diffusion models [27]. Even though other diffusion models have been created in the literature, the Independent Cascade (IC) and Linear Threshold (LT) models are the ones that are most frequently applied for researching social influence maximisation issues. The primary characteristic of the independent cascade model is the independence of the influence distributed over each edge of the entire graph; in other words, activating one individual will stimulate another.

The dissemination of ideas and information can be simulated using this approach. Every edge $(u, v) \in E$ has an associated influence probability $p_{uv} \in [0, 1]$, that is, the probability that node u succeeds in its attempt to activate its neighbor node v . The threshold model can be used to explain how a new, untested product is adopted. Moreover, in the IC model every edge (u, v) was associated with an influence probability p_{uv} ; the same edges are this time assigned with influence weight $w_{uv} \in [0, 1]$ in the linear threshold model. This represents the degree to which a user u can influence another user v . Therefore the weights are normalized so that the sum of all incoming edges of each node is lower or equal to 1, i.e., $u \in \mathcal{N}_{in}(v) W_{uv} \leq 1$. For the independent cascade and linear threshold models, the idea of influence response is presented in Fig. 2.

Fig. 3 presents the cascading propagation analogy of the linear threshold model with a threshold value of 0.40, which shows the propagated influence in every time-step. Double-lined nodes (10, 11) are the seeds of the propagation. The blue-colored nodes are the newly influenced nodes at the respective time-step. In addition, two well-known models, there are other propagation models in the literature, such as time-aware, context-aware, and propagation with negative feedback [25, 21]. Based on the formalization of the influence spread, the influence maximization problem is expressed in Definition 2 as follows:

**Fig. 3.** Linear threshold propagation [5]

Consider a graph G be social graph, a diffusion model M with positive integer k , selects a subset S of k users as IM from V to maximize the influence spread $\sigma_{G,M}(S)$ based on seed set under the diffusion model M .

The influence function $\sigma_{G,M}(\cdot)$ appears to be strongly dependent on the diffusion process.

2.2 Multi-objective Optimization

Finding a group of prominent users in a social network is the challenge of influence maximization

(IM), which aims to maximize their combined impact [3, 38].

The primary goal of the traditional IM problem, given a limit on the size of the seed-set, is to maximize the total number of users who are influenced.

We present the Multi-Objective IM problem, a more sophisticated version of the IM problem that can handle numerous highlighted groups.

As a result, the objective function $f(x)$ is a list of optimization functions with k dimensions.

Mathematically the problem can be expressed as:

$$\begin{aligned} &\text{Optimize}_x \quad \vec{f}(x) = [f_1(x), f_2(x), \dots, f_k(x)]^T, \\ &\text{Subject to} \quad g_p(x) \leq 0, \quad \forall p = 1, 2, \dots, n, \\ &\quad \quad \quad h_q(x) = 0, \quad \forall q = 1, 2, \dots, m, \end{aligned} \quad (1)$$

where, $g(x)$ and $h(x)$ are n and m number of inequality and equality constraints. There is no one solution that concurrently optimizes all objectives in a non-trivial multi-objective optimization problem (MOP) [39]. In MOO, a solution x_1 is said to dominate another solution x_2 , if both conditions 1 and 2 are true:

1. In all objectives the solution x_1 is no worse than x_2 .
2. Similarly, the solution x_1 is better than x_2 in at least one objective.

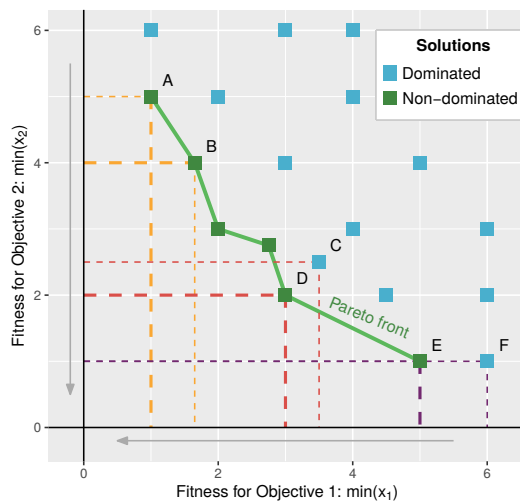


Fig. 4. Non-dominated solutions for a bi-objective minimization problem

In Fig. 4 demonstrates the Pareto front and non-dominated solutions concept. Here, the search space $\Omega \in \mathbb{R}^2$; $\vec{x} = [x_1, x_2]^T$ such that $x_1, x_2 \in [0, 6]$. The 2-objective functions are $f_1(\vec{x}) = \min(x_1)$ and $f_2(\vec{x}) = \min(x_2)$. A possible solution is represented by a square. The non-dominated or Pareto optimal solutions are represented by the green squares. The Pareto

front is the green line that connects Pareto optimal solutions. The cost for each objective function is indicated by dotted lines. The superiority of the objective is indicated by a bold dotted line.

In the last decade, many multi-objective optimizers have been introduced, notably the multi-objective genetic algorithm (MOGA) [10], the strength Pareto evolutionary algorithm 2 (SPEA2) [39], the non-dominated sorting genetic algorithm II (NSGA-II) [8], the multi-objective evolutionary algorithm based on decomposition (MOEA/D) [35]. Coello Coello [3] and Zitzler et al. [38] provide comprehensive research in multi-objective optimization among these, the NSGA-II is one of the most extensively used multi-objective optimizers.

2.3 Biogeography-based Optimization

Biogeography-based optimization (BBO), one of the best optimization techniques, has been shown to be quicker and more dependable than the alternatives [12]. Because it imitates natural patterns like species distribution, the BBO algorithm is rapid and produces the best results. The literature review revealed that researchers provided a variety of BBO migration models with outstanding performance results. The success of BBO is based on global exploration, yet it is susceptible to being stuck in local optima and unable to conduct a global search [13]. It's difficult to strike a balance between exploration and exploitation because of the low population diversity and convergence rate [14].

A novel modified BBO algorithm called locally and globally tuned BBO (LGBBO), which outperforms several variants of the original methods, has been presented to overcome the drawbacks of the existing BBO. The model is presented in Eq. 2 as:

$$H_i(SIV) \leftarrow \alpha NN(H_i(SIV)) + (1 - \alpha) H_j(SIV). \quad (2)$$

The parameter α is known as a geographical system's immaturity index [11]. Which is inversely related to the system's invasion resistance. As a result, the habitat's nearest neighbour $NN(H_i(SIV))$ can be specified in Eq. 3 as:

$$NN(H_i(SIV)) \leftarrow H_{(i \leq r?1:i-r)}(SIV), \quad (3)$$

where r is the neighborhood's radius. The best nearest neighbour habitat may be discovered at $(i - r)$ since the HSI's are sorted in a certain way. When $i \leq r$, the best habitat has been selected as the best locally.

Algorithm 1: Proposed Migration Algorithm (LGBBO)

```

Initialize:  $E = 1, I = 1, m_{max} = 1$ ,
Population size ( $N_p$ ), Maximum iteration;
Generate a random set of habitats
 $H_1, H_2, \dots, H_{N_p}$ ;
while Termination Condition do
    Determine HSI (fitness),  $\lambda$ ,  $\mu$ ,  $p_{mut}$  and
    mutation rate  $m_i$  for each habitat  $\alpha$  HSI ;
    Generate a random number
     $rand \in (0, 1)$ ;
    for each habitat according to their fitness
    values do
        Select a habitat  $H_i(SIV)$ 
        probabilistically  $\propto \lambda_i$ ;
        if  $rand < \lambda_i$  and  $H_i(SIV)$  selected
        then
            Select  $H_k(SIV)$  which is locally
            best to  $H_i(SIV)$  using
             $NN(H_i(SIV))$ ;
            Select an habitat  $H_j(SIV)$ 
            probabilistically  $\propto \mu_j$ ;
            if  $rand < \mu_j$  and  $H_j(SIV)$ 
            selected then
                Generate a constant  $\alpha \in [0, 1]$ ;
                for each solution features
                (SIV's) do
                     $H_i(SIV) \leftarrow$ 
                     $\alpha NN(H_i(SIV)) + (1 -$ 
                     $\alpha)H_j(SIV)$ ;
                end
            end
        end
    end
    Select an  $SIV(H_i)$  based on mutation
    probability proportional to  $p_i$ ;
    if  $rand < m_i$  then
        Replace  $SIV(H_i)$  with randomly
        generated  $SIV$ ;
    end
end

```

Algorithm 1 shows the general pseudo-code of standard LGBBO [11] for D-dimensional numerical optimization problem. The maximum number of iterations or achieving the intended outcome of the problem to be solved serve as the LGBBO algorithm's termination criteria.

3 Related Works and Research Gap

Influence maximization problem is a probabilistic method was first introduced as an algorithmic problem by Domingos and Richardson [9, 28]. But, Kempe et al. [20] established a discrete optimization problem and then used a hill-climbing greedy approximation approach to guarantee the influence spread is within $(1 - 1/e)$ of the optimum influence. Kimura and Saito [22] developed shortest-path-based influence cascade models and efficient computer influence spread techniques. The 'Cost-Effective Lazy Forward' (CELF) model, presented by Leskovec et al. [23], reduces the number of evolution's by employing a greedy optimization method and the sub-modularity property of influence maximization. Gong et al. [15] have extends the work of Leskovec et al. and improved the CELF approach.

Chen et al. [2] proposed NewGreedy and MixedGreedy algorithms for the IC model with uniform probabilities. However, their performance is non-steady, and sometimes even worse than CELF. Zhou et al. in [37, 36] have enhanced the CELF model using upper bound based lazy forward (UBLF). For mining the top-k influential nodes, Wang et al. [33] suggested a novel community-based greedy algorithm. Barbieri et al. [1] studied social influence for modeling perspective. Guo et al. [18] investigated the influence maximization problem for item-based data. Instead of taking influence probabilities as input, Goyal et al. [17] suggest an alternative technique to influence maximization that directly employs historical data.

The integral influence maximization problem was studied by [36] in the context of repeated activations. [16] and [31] investigate the parallel problem of learning how available data changes probability. On the IM-CM problem, Yang and Liu [34] employed a multi-objective

discrete particle swarm optimization technique to identify impactful seeds with the lowest cost. De and Dehuri [4] proposed non-dominated sorting biogeography-based optimization (NSBBO) algorithm to solve the IM-CM problem. They have also proposed a multi-objective ranking strategy for efficiently utilize preexisting knowledge for faster convergence.

3.1 Current Research Gap

Modeling social networks for the purpose of increasing influence is a non-deterministic polynomial (NP-hard) issue [26]. To distinguish seed nodes, previous studies used approximation and evolutionary methodologies.

However, the set of detected seeds might only cover a portion of the network due to the architecture of the network and the propagation model. Fig. 5(a) depicts such an example. We tried to identify two seeds in this figure using the linear threshold propagation model. In this propagation model, a node can be activated with a threshold value of 0.4 (i.e., at least 40% of the nearby nodes must be activated for a node to become active). Out of a total of 17, the propagation can, in the best situation, reach up to 14 nodes.

The propagation model has an impact on the dark nodes in this illustration. White nodes, however, are unaffected by the aforementioned propagation. Figs. 14(a), 15(a) also shows parallel instances for two well-researched networks, Zachary's Karate Club and Dolphins Social Network, respectively. Therefore, it is conceivable to increase diffused influence even further by adding a few missing linkages.

The two networked components in Fig. 5(a) are linked by a bridge between nodes 2 and 6. In this figure, the bridge-connected vertex could have a maximum threshold value of 0.33, meaning that the non-affected component won't be affected. The fact that 0.33 is smaller than 0.4 prevents propagation in this situation. But it might be able to increase the reach of its effects by including a missing piece like 2-11. After adding a few missing links, Figs. 14(b), 15(b) demonstrates how the influence can spread even further. As far as we know, there hasn't been any effort made to

reorganize the network in a way that maximizes influence. We provide a method for locating seed nodes and suggest missing linkages to maximize influence in the section below.

4 Our Proposed Approach

Our suggested strategy entails several stages. Here, we have identified a number of seeds with the appropriate time-step to spread influence. Next, figure out how many lost links are necessary for influence to spread to the network's unaffected area. The very minimum set of gaps was then suggested. We have covered them in detail in the subsections below.

4.1 Identification of Initial Seeds

A problem involving efficient influence maximization has two or more optimization goals, namely to maximize aggregate influence and reduce influence propagation time. As a result, the problem can be expressed in its simplest form as a bi-objective optimization problem, where both objectives must be optimized simultaneously. Consider a social graph G , a diffusion model M , and an influence maximization problem identifies a set S of k nodes from seed V , where k a positive integer, as the set that maximize the influence spread $\sigma_{G,M}(S)$, as well as minimize the propagation time-step $\mathcal{T}(S)$ for influence propagation, i.e., $\mathcal{T}(S) = \arg \max \mathcal{T}(v) \forall v \in \sigma_{G,M}(S)$. The influence maximization problem is mathematically represented using eq. 4:

$$\begin{aligned} & \underset{S}{\text{Maximize}} \quad [\sigma_{G,M}(S), -\mathcal{T}(S)]^T, \\ & \text{Subject to} \quad |S| = k, \\ & \quad \quad \quad S \in V. \end{aligned} \quad (4)$$

In eq. 4 we have used $-\mathcal{T}(S)$ as in optimization maximize $\mathcal{T}(S)$ is equivalent to minimize $-\mathcal{T}(S)$.

The majority of the time, social networks do not provide at least one choice that concurrently maximizes both goals. So, when choosing such optimization, a balance becomes essential. The aggregated influence objective has a higher priority in this specific challenge than maximization.

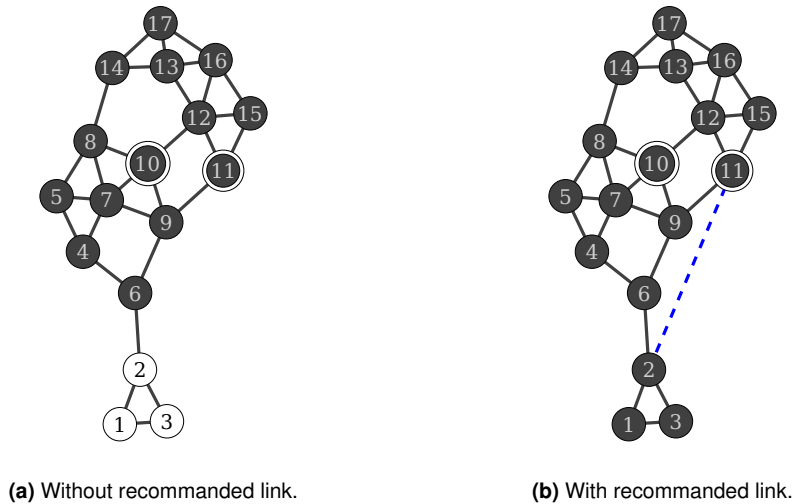


Fig. 5. Influence coverage considering structural modification [5]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0

Fig. 6. Binary encoding of the seeds (10, 11) presented in Fig. 3

Any multi-objective solver can be used to solve the influence maximization problem because it is a bi-objective optimization problem. We have employed NSGA-II in this work to identify seeds. With this method, numerous non-dominated seed sets are found together with the matching node coverage.

4.2 Solution Encoding

Encoding of the solutions is still one of the basic requirements of NSGA-II. The two basic types of encoding systems used in evolutionary algorithms are *Binary encoding* and *real coded encoding*.

The binary encoding of a network solution employs a n bit string, where n is the network's size. The string's index i denotes the network's i th node. A 1 bit indicates that the node is a seed node, while a 0 indicates that it is not.

The binary encoding for the solutions (seeds are 10, 11) reported in Fig. 3 is shown in Fig. 6.

The network size and the binary string size in this encoding are same. The strategy is therefore

1	2
10	11

Fig. 7. Real coded encoding of the seeds (10, 11) presented in Fig. 3

Parent 1	$\sqrt{(3)}$	$\sqrt{(2)}$	$\sqrt{(1)}$	$\sqrt{(5)}$	$\sqrt{(6)}$
Parent 2	$\sqrt{(2)}$	$\sqrt{(4)}$	$\sqrt{(1)}$	$\sqrt{(7)}$	$\sqrt{(5)}$
Offspring 1	$\sqrt{(3)}$	$\sqrt{(2)}$	$\sqrt{(1)}$	$\sqrt{(7)}$	$\sqrt{(5)}$
Offspring 2	$\sqrt{(2)}$	$\sqrt{(4)}$	$\sqrt{(1)}$	$\sqrt{(5)}$	$\sqrt{(6)}$
Parent 3	$\sqrt{(3)}$	$\sqrt{(2)}$	$\sqrt{(8)}$	$\sqrt{(5)}$	$\sqrt{(7)}$
Parent 4	$\sqrt{(2)}$	$\sqrt{(3)}$	$\sqrt{(1)}$	$\sqrt{(6)}$	$\sqrt{(4)}$
Offspring 3	$\sqrt{(3)}$	$\sqrt{(2)}$	$\sqrt{(8)}$	$\sqrt{(6)}$	$\sqrt{(4)}$
Offspring 4	$\sqrt{(2)}$	$\sqrt{(3)}$	$\sqrt{(1)}$	$\sqrt{(5)}$	$\sqrt{(7)}$

Fig. 8. Duplicate solutions generated after crossover. In the illustration, although all 4 parents are unique, after crossover they produce duplicate solutions i.e., offspring 1 and offspring 4

ineffective when dealing with huge social networks. It's possible that the representation will use up all of the RAM.

The collection of essential players, however, is merely maintained in its current state in the real coded encoding skim.

They may therefore offer model solutions regardless of the size of the network. Less memory

is used thanks to the approach. In Fig. 7 the real coded encoding for the solutions presented in Fig. 3 have been shown.

Although evolutionary approaches pose a few more difficulties, real coded encoding is better suited for large-scale network optimization problems. They are listed below:

1. The crossover and mutation operations may produce duplicate solutions. Figs. 8 and 9 depicts the situations for the crossover and for the mutation respectively.

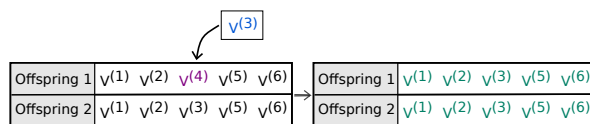


Fig. 9. A situation where duplicate solutions are generated after mutation. In Offspring 1, mutation happens to $v^{(4)}$. After the mutation $v^{(4)}$ is replaced with $v^{(3)}$. Therefore, Bothe the solutions (Offspring 1 and Offspring 2) become same after the mutation

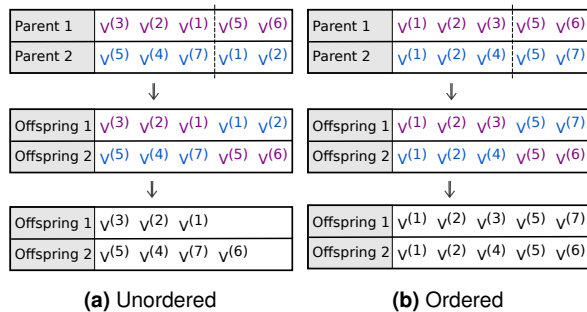


Fig. 10. An unordered set, shrink due to crossing. Using an ordered set, the problem was eliminated

2. A node occurring more than once in a solution may result in a shrinking of the set size (i.e., $|S|k$). Duplicate generation is potentially less likely in an ordered set. Nevertheless, the ordering does not ensure that the duplicate actor will be entirely eliminated. Fig. 10(a) illustrates how a crossover can reduce the size of the solution and Fig. 10(b) shows the same set reduces the duplicate generation on an ordered set. Fig. 11 illustrates the shrink of set size due to mutation.

4.3 Link Recommendation

For the sake of the link recommendation towards influence maximization, the first and most important thing is to make a few non-influenced nodes active so that the newly activated nodes can further propagate influence. Thus identifying these set of nodes is essential. Finding the amount of new links necessary to convert a non-influenced node into an active node is crucial for recommending a minimum number of links.

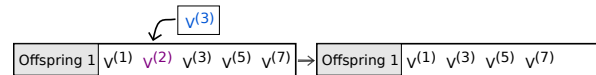


Fig. 11. Shrink due to mutation. In Offspring 1 mutation happens to $v^{(2)}$. $v^{(2)}$ is replaced with $v^{(3)}$. As no duplicate elements can present in a solution the solution size become less after the mutation

A node can be made active by adding a few extra links to a few other nodes that are already active. As a result, we have calculated the number of additional links needed to activate each node in the network's non-influenced segment.

As per the working principle of the linear threshold propagation model, a node v can be activated iff the given threshold is greater than or equal to $|\mathcal{N}_A(v)|/|\mathcal{N}(v)|$. Here $\mathcal{N}(v)$ and $\mathcal{N}_A(v)$ denote neighboring nodes of v and activated neighboring nodes of v respectively. Let to make a non-influenced node v active, $|\mathcal{N}_{A+}(v)|$ additional activated nodes needs to be connected with v . Here, $\mathcal{N}_{A+}(v)$ denotes set of additional active nodes that are to be linked to make node v active. Thus, after adding $|\mathcal{N}_{A+}(v)|$ additional links new number of activated neighbouring nodes of v will be $(|\mathcal{N}_A(v)| + |\mathcal{N}_{A+}(v)|)$ and new number of neighbouring node of v will be $(|\mathcal{N}(v)| + |\mathcal{N}_{A+}(v)|)$. Therefore eq. 5 becomes valid after adding $|\mathcal{N}_{A+}(v)|$ new links. Solving eq. 5 we obtain eq. 4 for computing the number of additional links required for every non-influenced node v to make them active.

For the partial influence propagation presented in Fig. 5(a), Table 1 shows the computation of the number of additional links required to make every non-influenced node to active.

Table 1. Computation of number of additional links required for each non-influenced node presented in Fig. 5(a)

v	$\mathcal{N}(v)$	$ \mathcal{N}(v) $	$\mathcal{N}_A(v)$	$ \mathcal{N}_A(v) $	Threshold	$ \mathcal{N}_{A+}(v) $
1	[2,3]	2	[]	0	0.4	$\frac{0.4 \times 2 - 0}{1 - 0.4} = \lceil 1.33 \rceil = 2$
2	[1,3,6]	3	[6]	1	0.4	$\frac{0.4 \times 3 - 1}{1 - 0.4} = \lceil 0.33 \rceil = 1$
3	[1,2]	2	[]	0	0.4	$\frac{0.4 \times 2 - 0}{1 - 0.4} = \lceil 1.33 \rceil = 2$

Table 2. Potential nodes from the non-influenced segment for the Dolphin social network presented in Fig. 15(a)

Nodes	Coverage	Timestep	Links required
(46, 52)	36	11	13 (46: 8, 52: 7)
(15, 38)	39	13	17 (15: 9, 38: 8)
(34, 38)	39	15	15 (34: 7, 38: 8)
⋮			

Here each non-influenced node v to be linked with $|\mathcal{N}_{A+}(v)|$ number of active nodes. The LT propagation has to run and be checked for convergence after activating a single node. This strategy, however, might necessitate a significant number of new links. An evolutionary approach, however, performs better at locating a small number of nodes from the network's non-influenced region. After influence propagation starts from the identified set of nodes, they must be able to impact the non-influence segment as much as possible. The recommended links will have this group of nodes as one of their ends.

In order to make the detected nodes active, connect them to the active node's required number of times (i.e., the total of the links that are required to activate these nodes). Since ranking data is accessible, we have employed the locally globally tuned biogeography-based optimization technique [11, 12] to identify this collection of nodes. Table 2 shows propagation statistics of few potential nodes against Fig. 15(a). The bold row in this table denotes the optimal solution. Because it needs fewer new links, maximum coverage, and a brief time stamp. The optimum combination in this scenario is (46, 52), as it requires the fewest (13 in this example) extra links to cover all 36 remaining nodes.

Finding the other end of the linkages within the active nodes is necessary after collecting the needed set of nodes from the network's non-influenced region. It might not always be possible to create a new link to an active node. Undoubtedly, linking an inactive node to an active node decreases the active node's threshold. In this situation, the propagation stops much before the expected propagation. Therefore for the activated node, it is essential to check whether a link can be added or not? Let the desired node was activated at time-step T_x . Therefore the ratio of activated neighboring nodes and the number of neighboring nodes at $T_x - 1$ time-step must be greater than or equal to the supplied threshold value. Thus the activation requirement can be checked using eq. 5:

$$\frac{|\mathcal{N}_A^{(T_x-1)}(v)|}{|\mathcal{N}(v)| + 1} \geq \text{Threshold}, \quad (5)$$

where, $\mathcal{N}_A^{(T_x-1)}(v)$ is the activated neighboring nodes of v at timestep $T_x - 1$. Number of neighboring nodes after adding a missing node at time $T_x - 1$ is $|\mathcal{N}(v)| + 1$.

The computation of the new threshold following the establishment of a link to an unaffected node is shown in the table 3. The bold threshold in this table denotes that values can be added to the relevant node after the link. The computed value is shown in the table's "Threshold after" column. Since nodes 7 and 12 require a minimum threshold value of 0.40 to be activated, creating a link from them is not possible. Other nodes than 7 and 12 are eligible for taking into account when creating linkages. As the seed nodes, nodes 10 and 11 will always have a threshold value of 1.

The required number of links in this experiment (1) is lower than the necessary number of seeds (2). The missing link can therefore be usually

Table 3. Computation of threshold after adding a link to a non-influenced node

v	Timestep	$\mathcal{N}_A(v)$	$\mathcal{N}(v)$	$ \mathcal{N}_A(v) $	$ \mathcal{N}(v) $	Threshold before	Threshold after
4	5	[5,7,9]	[5,6,7,9]	3	4	$3/4 = 0.75$	$3/(4+1) = \mathbf{0.60}$
5	4	[7,8]	[4,7,8]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
6	6	[4,9]	[2,4,9]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
7	2	[10,9]	[4,5,8,9,10]	2	5	$2/5 = 0.40$	$2/(5+1) = 0.33$
8	3	[7,10]	[5,7,10,14]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
9	1	[10,11]	[6,7,10,11]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
10	0	$\square$	[7,8,9,11,12]	0	5	1.0	1.0
11	0	$\square$	[9,10,12,15]	0	4	1.0	1.0
12	1	[10,11]	[10,11,13,15,16]	2	5	$2/5 = 0.40$	$2/(5+1) = 0.33$
13	4	[12,16]	[12,14,16,17]	2	4	$2/3 = 0.67$	$3/(4+1) = \mathbf{0.60}$
14	5	[8,13]	[8,13,17]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
15	2	[11,12]	[11,12,16]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
16	3	[12,15]	[12,13,15,17]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
17	5	[13,16]	[13,14,16]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$

linked to either of the seeds without the need for any prior calculation (as the seed nodes having minimum timestep value). The suggested links in this scenario are therefore (2, 10) or (2, 11).

5 Experimental Setup and Result Analysis

The experimental study uses three different use cases that might occur during the identification of seeds and recommended links. This study uses three different social networks namely 1) an experimental network used throughout this article, 2) the interaction network of Zachary's karate club, and 3) Dolphins social networks. In this study, the presentation of the network follows a few basic conventions. They are: 1) Circle represents a node, 2) A line between two circles indicate a link. 3) White colored nodes are non-influenced nodes, 4) Influenced nodes are Gray-colored, 5) double bordered nodes are the seeds of the propagation, 6) a blue node indicates the node becomes active in the current timestep, and 7) blue dashed lines are recommended links. Throughout the study, we have used the linear threshold propagation model with a threshold value of 0.40.

Fig. 12 presents the multi-objective seeds identification for the network presented in Fig. 5. X-axis of the plot shows the total influence

coverage after the end of the propagation. Y-axis of the plot shows the required timestep to meet the maximum possible coverage. Blue dots are the non-dominated solutions, whereas black dots are the dominated solutions. The solid blue line is the Pareto front.

However, as total influence coverage in terms of the number of nodes has a higher priority as compared to the timestep, solutions (9, 12) and (10, 11) are the best among all other Pareto optimal solutions. In this study, we have chosen seeds (10, 11).

Fig. 12(a) shows the maximum coverage of seeds without a recommended link. In this scenario without the recommended links the best seeds able to cover 14 nodes out of 17 nodes. The influence propagation took a maximum of 6 timesteps. Fig. 12(b) shows the propagation after adding the recommended link (2, 11). After adding the recommended link (2, 11) the influence covers complete networks. The additional propagation completes in 8 timesteps.

Fig. 13(a) shows the limited propagation with best seeds (10, 12) without the recommended link. Figs. 13(b) and 13(c) show the influence propagation with recommended link (2, 11). In Fig. 13(b) node 2 become active at time-step 7 after adding link (2, 11) and at time-step 8 (Fig. 13(c)) nodes 1, 3 become active. In this

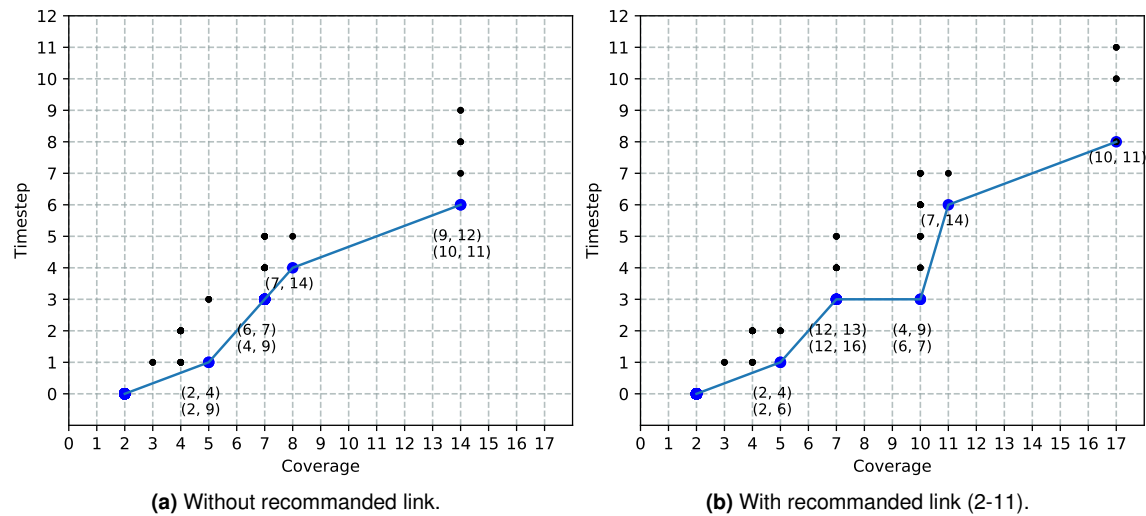


Fig. 12. Identification of initial non-dominated solutions for the network presented in Fig. 5

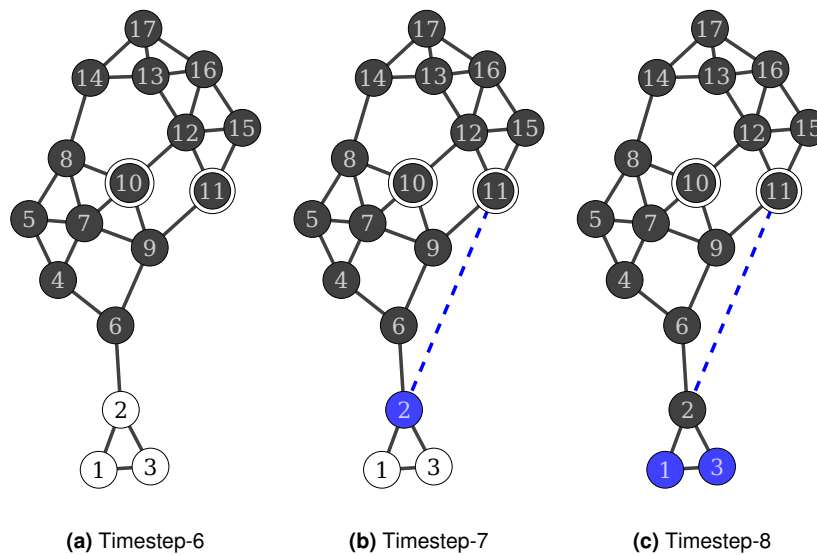


Fig. 13. Linear threshold propagation with recommended link on experimental network

particular case additional time-step is required to maximize influence.

Fig. 14 presents the influence propagation on the interaction network of Zachary's karate club.

Table 4 shows the initial seeds and their respective influence coverage and time-step. In this table the bold value represents the best solution.

Seeds (1, 34) is the only non-dominated solution present in this table as it covers maximum with the lowest time-step.

Fig. 14(a) and 14(b) presents the influence propagation without recommended link and with recommended link for the seeds (1, 34) respectively. In this particular experiment, only a single link (5, 34) is capable of propagating influence

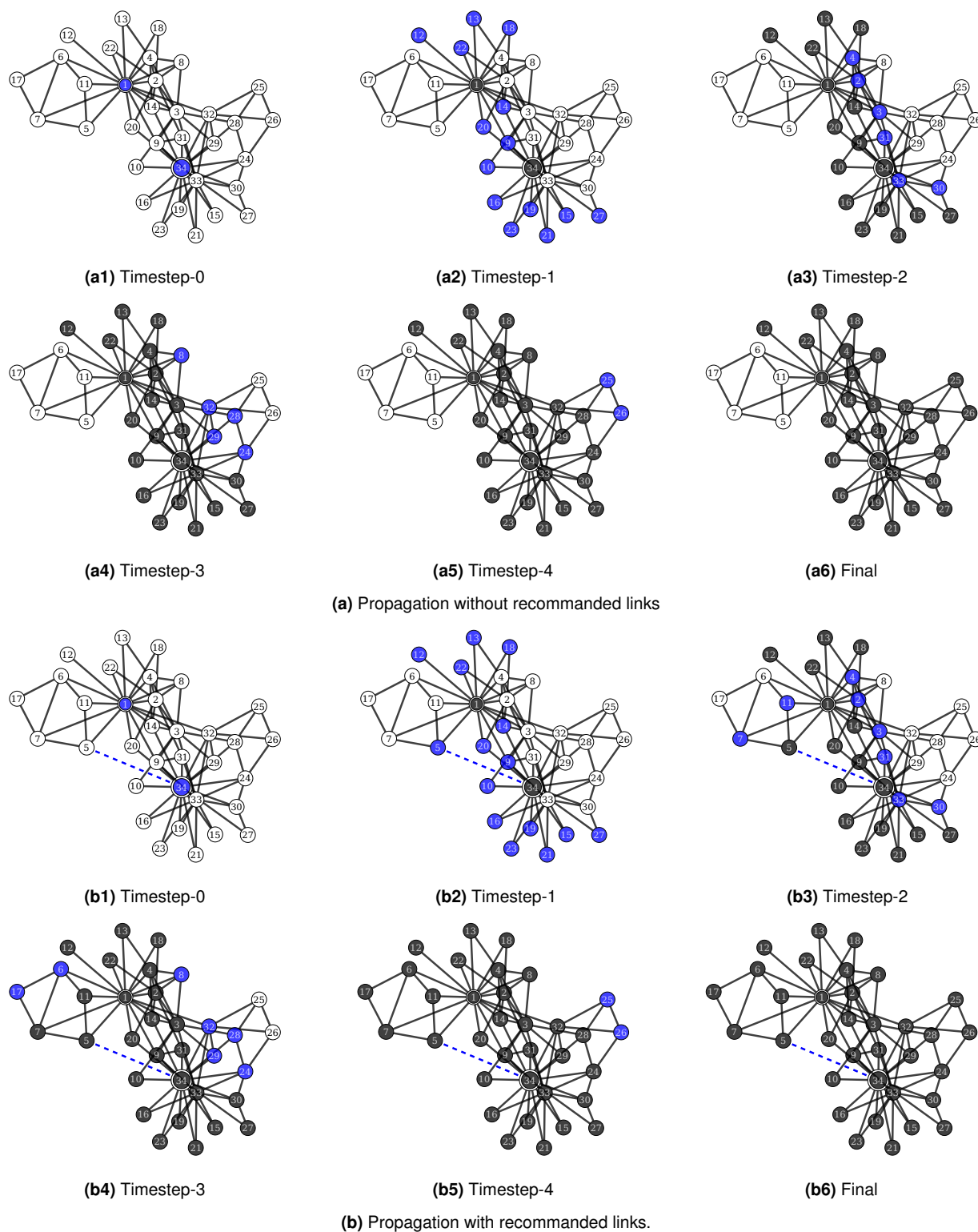
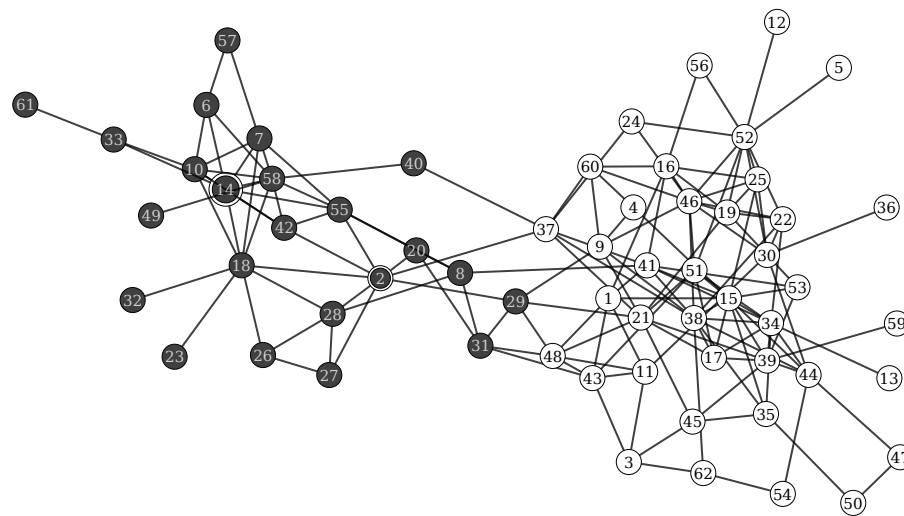
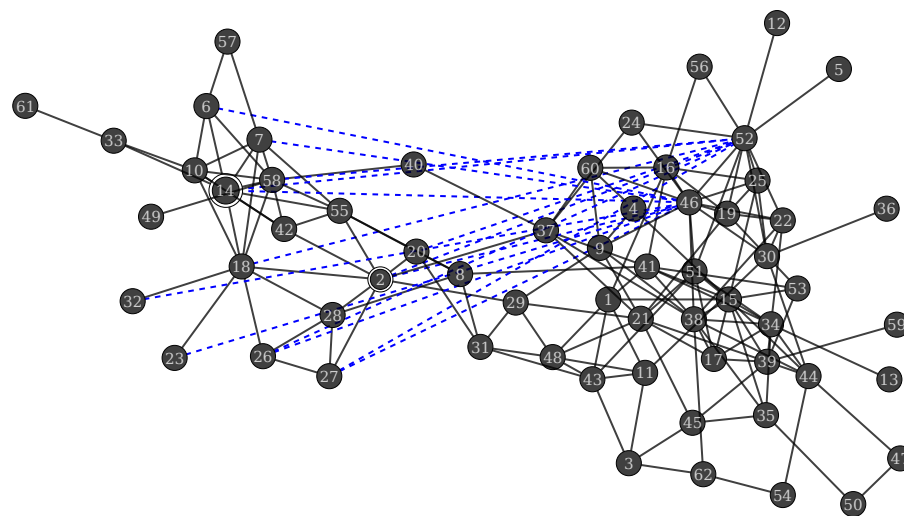


Fig. 14. Influence propagation on the network of Zachary's karate club under linear threshold model. A threshold value 0.4 is used



(a) Influence without recommended links



(b) Influence with recommended links

Fig. 15. Influence maximization on Dolphins social network under linear threshold model. A threshold value 0.4 is used as a parameter

Table 4. Seeds with their Coverage and Timestep

Seeds	Coverage	Timestep	Seeds	Coverage	Timestep	Seeds	Coverage	Timestep
(1, 34)	29	4	(4, 34)	29	6	(13, 34)	29	10
(1, 33)	29	5	(2, 33)	29	7	(18, 34)	29	10
(2, 34)	29	6	(8, 34)	29	8	(20, 33)	29	10

throughout the network. However, the added link doesn't demand any additional time-steps.

For the Dolphins social network the best seeds for LT propagation with a 0.4 threshold are (2, 14) and (2, 58). These seeds can influence 23 nodes out of 62 nodes in 11 timesteps. In this study, we have chosen (2, 14) as the candidate seeds. The propagation for seeds (2, 14) is presented in Fig. 15(a).

In Fig. 15(a) the non-influenced section contains 39 nodes. For these 39 nodes, LGBBO has been executed and determined potential nodes set that maximally influence the remaining 39 nodes. Table 2 presents some solutions. We have chosen solution (46, 52) as they can influence the remaining 39 nodes with the lowest timestep. Most importantly the solution requires less number of missing links i.e. 15 to make them active.

Fig. 15(b) shows the influence propagation after adding the recommended links.

6 Conclusion

In today's world, maximizing one's influence in social networks has huge ramifications. Identifying a set of seed nodes that are suitable for influence maximization, on the other hand, is extremely difficult. The identification process is considered NP-hard.

Alongside that due to network structure and given propagation model, sometimes it is not possible to propagate influence that covers the entire network or at least a significant portion of the network. However, we observed that a little structural modification by adding a few missing links further allows information propagation in the non-influenced segment of the network. Determining the number of missing links required to establish smooth propagation is hard. Effective

recommendation of such links is computationally NP-hard.

In the view of maximizing propagation, our proposed two-stage approach identifies seeds and few missing links with a minimum time-step for propagation. In the first stage, we have invoked a multi-objective optimizer specially NSGA-II to find out the initial set of seeds along with coverage and time-step for the convergence. In the second stage, we have statistically calculated the number of missing links that are required to convert the status of the non-influence nodes.

Using LGBBO, we have determined suitable nodes that maximally propagate influence in the non-influenced segment. Lastly, we have checked and established links to the active nodes. In the empirical analysis section, we have shown three use cases to validate the proposed strategy. The empirical analysis showed a good improvement in influence maximization after using the recommended seeds and links.

References

1. Barbieri, N., Bonchi, F., Manco, G. (2013). Topicaware social influence propagation models. DOI: 10.1007/s10115-013-0646-6.
2. Chen, W., Wang, Y., Yang, S. (2009). Efficient influence maximization in social networks. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 199–208. DOI: 10.1145/1557019.1557047.
3. Coello-Coello, C. A. (2001). A short tutorial on evolutionary multiobjective optimization. International Conference on evolutionary multi-criterion optimization, Springer, pp. 21–40. DOI: 10.1007/3-540-44719-9_2.

4. **De, S. S., Dehuri, S. (2019).** Multi-objective biogeography-based optimization for influence maximization-cost minimization in social networks. *International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making*, Springer, pp. 11–34.
5. **De, S. S., Dehuri, S. (2020).** Multi-objective biogeography-based optimization for influence maximization-cost minimization in social networks. *Biologically Inspired Techniques in Many-Criteria Decision Making: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (BITMDM-2019)*, Springer, pp. 11–34.
6. **De, S. S., Dehuri, S., Cho, S.-B. (2017).** Divide-and-conquer approach for revealing the non-dominated solutions in multi-objective optimization problem. *2017 International Conference on Information Technology (ICIT)*, IEEE, pp. 143–151.
7. **Deb, K., Agrawal, S., Pratap, A., Meyarivan, T. (2000).** A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *International conference on parallel problem solving from nature*, Springer, pp. 849–858. DOI: 10.1007/3-540-45356-3_83.
8. **Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. (2002).** A fast and elitist multiobjective genetic algorithm: Nsga-ii. DOI: 10.1109/4235.996017.
9. **Domingos, P., Richardson, M. (2001).** Mining the network value of customers. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 57–66. DOI: 10.1145/502512.502525.
10. **Fonseca, C. M., Fleming, P. J., et al. (1993).** Genetic algorithms for multiobjective optimization: Formulation/discussion and generalization. *Icga, Citeseer*, Vol. 93, No. July, pp. 416–423.
11. **Giri, P. K., De, S. S., Dehuri, S. (2018).** A novel locally and globally tuned biogeography-based optimization algorithm. In *Soft Computing: Theories and Applications*. Springer, pp. 635–646.
12. **Giri, P. K., De, S. S., Dehuri, S. (2021).** Adaptive neighbourhood for locally and globally tuned biogeography based optimization algorithm. DOI: 10.1016/j.jksuci.2018.03.013.
13. **Giri, P. K., De, S. S., Dehuri, S., Cho, S. (2017).** Locally and globally tuned chaotic biogeography-based optimization algorithm.
14. **Giri, P. K., De, S. S., Dehuri, S., Cho, S. (2021).** Biogeography based optimization for mining rules to assess credit risk. DOI: 10.1002/isaf.1486.
15. **Gong, M., Yan, J., Shen, B., Ma, L., Cai, Q. (2016).** Influence maximization in social networks based on discrete particle swarm optimization. DOI: 10.1016/j.ins.2016.07.012.
16. **Goyal, A., Bonchi, F., Lakshmanan, L. V. (2010).** Learning influence probabilities in social networks. *Proceedings of the third ACM international conference on Web search and data mining*, pp. 241–250. DOI: 10.1145/1718487.1718518.
17. **Goyal, A., Bonchi, F., Lakshmanan, L. V. (2011).** A data-based approach to social influence maximization. DOI: 10.14778/2047485.2047492.
18. **Guo, J., Zhang, P., Zhou, C., Cao, Y., Guo, L. (2013).** Item-based top-k influential user discovery in social networks. *2013 IEEE 13th International Conference on Data Mining Workshops*, IEEE, pp. 780–787.
19. **Iacobucci, D. (1996).** Network models of the diffusion of innovations.
20. **Kempe, D., Kleinberg, J., Tardos, É. (2003).** Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. DOI: 10.1145/956750.956769.
21. **Kempe, D., Kleinberg, J., Tardos, É. (2005).** Influential nodes in a diffusion model for social networks. *International*

- Colloquium on Automata, Languages, and Programming, Springer, pp. 1127–1138. DOI: 10.1007/11523468.91.
22. **Kimura, M., Saito, K. (2006).** Tractable models for information diffusion in social networks. European conference on principles of data mining and knowledge discovery, Springer, pp. 259–271. DOI: 10.1007/11871637.27.
 23. **Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N. (2007).** Cost-effective outbreak detection in networks. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420–429. DOI: 10.1145/1281192.1281239.
 24. **Li, F., Du, T. C. (2017).** The effectiveness of word of mouth in offline and online social networks. DOI: 10.1016/j.eswa.2017.07.004.
 25. **Li, Y., Fan, J., Wang, Y., Tan, K.-L. (2018).** Influence maximization on social graphs: A survey. DOI: 10.1109/TKDE.2018.2807843.
 26. **Lu, Z., Zhang, W., Wu, W., Kim, J., Fu, B. (2012).** The complexity of influence maximization problem in the deterministic linear threshold model. DOI: 10.1007/s10878-011-9393-3.
 27. **Purba, K. R., Raja, D. A., Murugesan, K. (2022).** Influence maximization diffusion models based on engagement and activeness on instagram. DOI: 10.1016/j.jksuci.2020.09.012.
 28. **Richardson, M., Domingos, P. (2002).** Mining knowledge-sharing sites for viral marketing. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 61–70. DOI: 10.1145/775047.775057.
 29. **Ritzer, G., et al. (2007).** The Blackwell encyclopedia of sociology.
 30. **Ryan, B., Gross, N. (1950).** Acceptance and diffusion of hybrid corn seed in two iowa communities.
 31. **Saito, K., Nakano, R., Kimura, M. (2008).** Prediction of information diffusion probabilities for independent cascade model. International conference on knowledge-based and intelligent information and engineering systems, Springer, pp. 67–75. DOI: 10.1007/978-3-540-85567-5_9.
 32. **Valente, T. W. (2005).** Network models and methods for studying the diffusion of innovations.
 33. **Wang, Y., Cong, G., Song, G., Xie, K. (2010).** Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1039–1048. DOI: 10.1145/1835804.1835935.
 34. **Yang, J., Liu, J. (2017).** Influence maximization-cost minimization in social networks based on a multiobjective discrete particle swarm optimization algorithm.
 35. **Zhang, Q., Li, H. (2007).** MOEA/D: A multiobjective evolutionary algorithm based on decomposition. DOI: 10.1109/TEVC.2007.892759.
 36. **Zhou, C., Zhang, P., Guo, J., Guo, L. (2014).** An upper bound based greedy algorithm for mining top-k influential nodes in social networks. Proceedings of the 23rd International Conference on World Wide Web, pp. 421–422. DOI: 10.1145/2567948.2577336.
 37. **Zhou, C., Zhang, P., Guo, J., Zhu, X., Guo, L. (2013).** Ublf: An upper bound based approach to discover influential nodes in social networks. 2013 IEEE 13th International Conference on Data Mining, IEEE, pp. 907–916.
 38. **Zitzler, E., Laumanns, M., Bleuler, S. (2004).** A tutorial on evolutionary multiobjective optimization.
 39. **Zitzler, E., Laumanns, M., Thiele, L. (2001).** SPEA2: Improving the strength pareto evolutionary algorithm.

Article received on 13/02/2023; accepted on 13/01/2025.

**Corresponding author is Parimal Kumar Giri.*