Physica A 586 (2022) 126480

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

An improved influence maximization method for social networks based on genetic algorithm

Jalil Jabari Lotf, Mohammad Abdollahi Azgomi *, Mohammad Reza Ebrahimi Dishabi

Department of Computer Engineering, Miyaneh Branch, Islamic Azad University, Miyaneh, Iran

ARTICLE INFO

Article history:
Received 21 March 2021
Received in revised form 14 September 2021
Available online 5 October 2021

Keywords:
Influence maximization
Social networks
Network dynamicity
Diffusion model
Genetic algorithms

ABSTRACT

Over the recent decade, much research has been conducted in the field of social networks. The structure of these networks has been irregular, complex, and dynamic, and certain challenges such as network topology, scalability, and high computational complexities are typically evident. Because of the changes in the structure of social networks over time and the widespread diffusion of ideas, seed sets also need to change over time. Since there have been limited studies on highly dynamical changes in real networks, this research intended to address the network dynamicity in the classical influence maximization problem, which discovers a small subset of nodes in a social network and maximizes the influence spread. To this end, we used soft computing methods (i.e., a dynamic generalized genetic algorithm) in social networks under independent cascade models to obtain a dynamic seed set. We modeled several graphs in a specified timestamp through which the edges and the nodes changed within different time intervals. Attempts were made to find influential individuals in each of these graphs and maximize individuals' influences in social networks, which could thereby lead to changes in the members of the seed set. The proposed method was evaluated using standard datasets. The results showed that due to the reduction of the search areas and competition, the proposed method has higher scalability and accuracy to identify influential nodes in these snapshot graphs as compared with other comparable algorithms.

1. Introduction

Over the recent decade, there has been a great deal of research on social networks. The structure of these networks has been irregular, complex, and changing dynamically over time. A social network consists of a structure containing the individuals or organizations and the relationships between them. Based on the network type, the relationships between the constituents can represent friendship, co-working, scientific relations, followership, and so on. Social networks have some characteristics that make their analysis difficult. Some of these characteristics are as follows [1]:

- *Structural complexity:* A complex network might have a complex topology; for example, a food network has an interrelated topology and analyzing its structure is considered to be a challenging issue;

* Corresponding author.
E-mail addresses: jalil.jabari@m-iau.ac.ir (J. Jabari Lotf), azgomi@iust.ac.ir (M. Abdollahi Azgomi), mrebrahimy@m-iau.ac.ir (M.R. Ebrahimi Dishabi).

https://doi.org/10.1016/j.physa.2021.126480
0378-4371/

- *Network dynamicity:* A network might change over time so that some nodes/edges are likely to be created or omitted;
- *Multiple nodes/edges:* There can be different types of nodes or edges in a network; The incongruence of nodes/edges has led to a new research field entitled "analysis of heterogeneous complex networks"; and
- *Large scales:* Many complex networks have a lot of nodes, making most algorithms unusable.

As the Internet evolves throughout the world rapidly, social networks become more and more popular. Social networks connect lots of individuals to each other within a short period of time. These networks have evolved new relations between people. Via interacting in social networks, a lot of data are exchanged, ideas and knowledge are shared, and individuals affect each other. In fact, a vast amount of data have been prepared due to the development and diffusion of social networks. Analyses of these data can help individuals understand the structural and behavioral features of social networks and answer many economic, social, and sociological questions. Therefore, the social network analysis has received much attention as one of the sub-branches of computer science in recent years. Network sampling, network partitioning, anomaly detection, link prediction, influence maximization, and so on are among challenging topics in the social network analysis. Also, due to the limitations in the complexity of graph data (e.g., irregular, independent, and large-scale structures) in the social network analysis, most studies have recently used soft computing to address these limitations [2–6].

One of the important research issues in the field of social networks is to analyze individuals' influence and information diffusion. Today, many individuals with different goals have become members of social networks and do different tasks. On the other hand, these individuals affect each other, leading to information diffusion and influence spread in a society or in a social network. Finding influential individuals in a social network has many practical applications in marketing, politics, and even control of diseases. Marketing managers may be interested in finding influential individuals and offering them a discount on their purchases or free products, with the hope that these individuals will encourage their friends to buy these products. Political actors are also seeking to find influential individuals to spread their message. Also, diffusion models which are used to analyze the influence on people can also be used to investigate the spread of infectious diseases. Infected influential nodes are able to infect a larger portion of the population than less influential nodes. Therefore, public health officials can get better results in treating diseases by spreading inoculation to the influential nodes. The main goal of influence maximization is to find a subset of influential individuals who can maximize the influence in the network under a diffusion model.

The study of the influence maximization relies on the diffusion models and the algorithms for identifying the influential nodes in order to maximize the influence. It has been proven that the influence maximization is an NP-hard problem [7–9]. Influence maximization has been studied vastly in the field of social networks. However, most of recent works have focused on investigating static networks. Meanwhile, social networks have a very dynamic nature, evolve over time, and are changing rapidly over time. Thus, the results related to the maximization and estimation of individuals' influence in social networks can become invalidated soon. Methods presented for solving the influence maximization problem have often been less studied due to the changes in nodes and edges and the changes in influence probability values on edges over time. On the other hand, issues such as the time-consuming feature of the existing solutions or the low accuracy of recognizing influential individuals and their lack of scalability have still remained unsolved. Therefore, it is necessary to develop a solution for dynamic and evolving networks with lower costs and higher accuracy. In this study, soft computing has been used to propose a method that takes into account time limitations, scalability, and network structures and improves the speed and accuracy of recognizing influential individuals through reduction of computations.

The rest of this paper is organized as follows: in the second section, a description of the related work on influence maximization in social networking is reviewed. The motivations and aims are mentioned in the third section. In the fourth section, the problem statement and description hare given. The precise steps to implement each of influence recognition and information diffusion activities are expressed in the fifth section. Finally, in the sixth section, the implementation results are presented and suggestions for future work are mentioned.

2. Related work

The influence maximization problem was first posed by Domingos and Richardson [10,11] as an algorithm issue to find the set of useful customers in viral marketing. To model this issue, they used Markov's random fields. Although the method proposed by Domingos and Richardson could resolve the influence maximization problem, their proposed model could not identify the way users affect each other and influence spread methods clearly.

Kempe et al. [12,13] modeled the influence maximization problem as a discrete optimization problem for the first time. They proved that on the whole, influence maximization is categorized among NP-Hard problems, but they proposed two diffusion models called independent cascade model and linear threshold model, and also a greedy approximation algorithm with an approximation limit of $1 - 1/(e - \varepsilon)$ where e and ε are respectively the Euler's number and estimation accuracy. Since the diffusion process is done randomly through the independent cascade and linear threshold models, the expected activated nodes in the greedy algorithms could be estimated through the frequent simulation of the influence spread process. The strong point of the greedy algorithms, is due to their high accuracy in identifying and recognizing important and effective nodes, but it should be noted that they are not scalable. Some of the proposed methods carried out some reforms in implementing the greedy algorithm, and reduction of simulation time tried to maintain limitation

approximation and improve time complexity. Some of these methods are as follows: CELF [14], CELF++ [15], TIM [16], BCT [17], RIS [18], and MixedGreedy [19]. Although these methods could improve greedy algorithm run time considerably, they have had high time complexities, and thus could not be generalized in large-scale dynamic social networks.

In some previous works, heuristic methods were utilized to resolve the influence maximization problem. Some of these methods are independent of the diffusion model, such as centrality criteria based on graph structure like degree criterion, Betweenness criterion, and closeness criterion [20–22]. Another group of heuristic methods were designed specifically for a diffusion model and showed an appropriate efficiency in the proposed model, and some of them are IRIE [23], and EaSylm [24]. The strong point of the heuristic proposed methods is the improvement of run time and increasing scalability. However, previous studies have shown that node selection based on centrality can create inappropriate results [25]. To resolve such a problem, some studies [26–28] have proposed new criteria such as centrality based on social interactions, based on node neighbors and the surroundings of neighbors, timed interactions, topology relations, and community-based ones that are more precise. Although run time was improved and scalability was increased, the heuristic methods do not have any guarantee for recognition accuracy of the effective nodes.

A group of proposed methods has tried to maintain the accuracy of influential individuals' recognition to improve the algorithm run time through some reforms in the implementation method of greedy and heuristic algorithms and the reduction of simulation time. Some of these methods are as follows: MOP [29], GOW [30], GNA [31]. Wang and et al. [32] considered the efficiency and the probability of the transfer of information for influence maximization in addition to diffusion and proposed a greedy algorithm based on a dynamic prune. [33] regarded the role of time in influence spread and proposed several sets of influential users in different times in a way that each of them is known as the best set of influential users in a certain snapshot of time. Since the influence maximization problem was proposed by Kempe and et al., many types of research have been carried out in the field and most of them included networks in static states which means that a network is deemed as a snapshot of the dynamic network in a certain time. Therefore, offline calculation methods could be utilized for them while in the real-world, social networks have a highly dynamic and evolving nature [34]. Tang and et al. in [35] presented a compatible greedy algorithm based on a dynamic independent cascade model that showed a good performance in recognizing influential individuals. Khomami and et al. in [36] proposed an algorithm based on learning automates to resolve the influence maximization problem in dynamic social networks.

Genetic algorithm (GA) is an optimization method in evolutionary computation and computational intelligence, which has been considered in recent years to solve optimization problems in the influence maximization problem. Bucur et al. [37] solved the influence maximization problem using a simple genetic algorithm. They showed that by using a simple genetic operator, an approximate solution can be obtained for influence in a better run time than by using the previously developed greedy algorithms. In [38], by extending a previous genetic algorithm, they proposed a meta-heuristic algorithm to select a fixed-length subset that uses the information about the network structure. Tsai et al. [39] used the genetic algorithm to find a solution and a greedy algorithm to improve the solution obtained through the genetic algorithm in order to enhance the efficiency of the influence maximization problem. Zhang et al. [40] provided a genetic algorithm, which maintains the diversity of solutions through competition and evolution between populations, and the results indicate a similar performance as of the greedy algorithm.

The dynamic social network can be isolated into a series of snapshots where each of them is utilized to explain the network status at an indicated time. Zhuang et al. [41] study the influence maximization under dynamic networks where the changes can be only detected by periodically probing some nodes. in [42] a practical framework is proposed by only probing partial communities to explore the real changes of a network and minimizes the possible difference between the observed topology and the real network through several representative communities and provides a regulatory mechanism for enhancing influence maximization. Wang et al. [43] propose the Influential Checkpoints framework and a Sparse Influence Checkpoints framework to tackle the stream influence maximization querying processing. By expanding the classic influence maximization problem, a lot of studies have been dedicated to adaptive influence maximization problem for dynamic social networks, which aims to address the challenges in regards to the network topology changes, high-speed data transfer, scalability, and high computational complexity [44,45].

3. Motivations and aims

Almost all proposed GA-based algorithms used for influence maximization cannot handle the dynamic structure of social networks. In the dynamic structure, the size of the data set becomes larger and the problem space increases. Since by increasing the problem space, the genetic algorithm does not guarantee an optimal solution, the previously proposed algorithms are not efficient in a large-scale environment and the accuracy of identifying influential individuals decreases.

The aim has been to propose a method to create a logical search range and optimizes the problem space by pruning, reducing the number of candidate solutions significantly. By limiting the problem space, we reduce the limitations of the genetic algorithm and the run time and increase the speed of convergence compared to the existing methods. Also, given that the problem space changes over time, we give the proposed generalized genetic algorithm the ability to evaluate and select solutions in a dynamic space.

In summary, our innovation in this research are as follows:

- We present an approach based on the dynamic structure of social networks and along with a generalized genetic algorithm to maximize the influence.

- We propose a new approach to reduce the computational costs and maintain the optimization process in large-scale dynamic social networks.
- We propose a new mechanism for identifying strategic nodes and edges considering the changes in social networks over time.

Regarding what was pointed out about influence maximization in social networks, it is considered a novel area of research that has challenges such as network topology changes, high speed data transfer, scalability, dynamic diffusion models, and high computational complexity.

4. The proposed method

In most research projects, social networks were assumed to be static. Also, graph changes (changing nodes and edges) and changes in influence probability on edges have rarely been investigated. On the other hand, issues such as the time-consuming feature of the proposed solutions or the low accuracy of recognizing influential individuals and their lack of scalability have still remained unsolved. Given the fact that in the real world, social networks have a dynamic and evolving nature, in this section, based on a dynamic generalized genetic algorithm (DGA) for the influence maximization problem in a dynamic social network (the network structure changes over time), we have proposed a novel method using the soft computing methods (the dynamic generalized genetic algorithm) in order to match them with the dynamic and scalable social networks.

4.1. Preliminaries and problem statement

Considering the changes in the social network structures over time, the seed set should also change over time. In this paper, we have modeled several graphs within a certain timestamp in which the edges and the nodes change in different time intervals and the goal is to find the influential individuals in each of these graphs, leading to changes in the members of the seed set. Of course, the number of the seed set constituents for each of the snapshot graphs is fixed. Each of these snapshot graphs are represented by $G_t = (V_t, E_t)$, where V_t is the set of nodes and E_t is the set of edges in time-step t. Also, G_t ($t = 1,\ldots,T$) defines a snapshot graph in a certain timestamp. The most transparent changes occurring between two time points t_{i-1} and t_i are the omission and addition of edges and nodes.

Algorithm 1: The Proposed Algorithm

Input:
Dynamic Network G_t
Timestamps T
The size of seed nodes k
Probability of influence node λ_{ij}

Output:
Seed set S_t at $t = 1,2,3\ldots, T$

```
1  for t = 1 to T do
2      G ← Gt
3      for each v ∈ V do
4          C ← Calculate Network Centrality Measures
5      P ← Pruning and Listing (C)
6      Set St ← Proposed DGA (P)
7      Extraction and Saving
8  return Seed set St at t = 1,2, 3…, T
```

As Algorithm 1 shows, the proposed algorithm acts as follows: First, we prune G_t snapshot graph according to the network structure based on the centrality criteria described in Section 4.2 and list it. Then, according to line 6 of the proposed algorithm, the generalized genetic algorithm, described in Section 4.4, is implemented and the seed set of the snapshot graph G_t is extracted and stored in order to be used in G_{t+1}. For the next snapshot graphs, we update the list and repeat the previous procedures for the graph G_{t+1}. Finally, we reach a dynamic seed set that expresses a set with the highest influence in each snapshot.

4.2. Pruning and listing

Social networks usually contain a lot of nodes. Thus, graphs resulted from such networks are very big, and naturally, such big graphs have high processing costs. One of the resolutions to reduce processing costs is to prune and list the

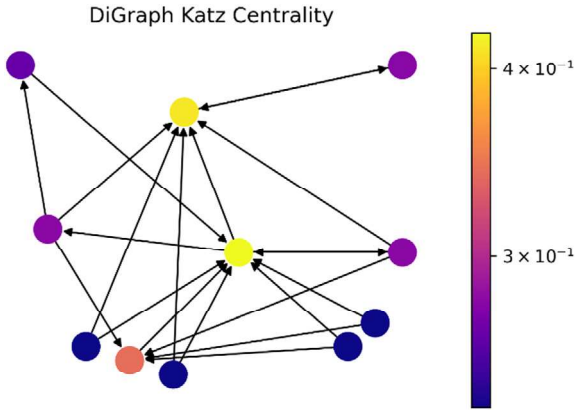


Fig. 1. Nodes with Katz centrality.

initial population appropriate with the graph. Using centrality criteria, we can recognize strategic nodes and edges in the network and start analyzing the network. In a directional graph and in order to spread information quickly within the network, we should look for nodes that are close to others to some extent. They can transfer the data to all nodes (directly or indirectly) and we can use closeness centrality criteria here. Closeness refers to the reverse of the average distance between a node and other nodes in the graph. The node with the highest amount of closeness has a higher access speed to other nodes and can send the data to all nodes within a short period of time. This criterion is calculated according to Eq. (1) as follows [20]:

$$C_c\left(n_i\right) = \left(\sum_{j=1}^{n} d_{ij}\right)^{-1} \tag{1}$$

where d_{ij} is the shortest route between the two nodes. In this study, we have used algorithm [46] with high scalability and accuracy to calculate closeness centrality. Betweenness centrality of a certain node in the network refers to the shortest routes between network nodes that cross a certain node. In fact, this criterion calculates how many numbers of network nodes need this node with less intermediaries for a more rapid connection. The higher betweenness of the node means that the node is strategic and has a high capability in transferring the data and such a criterion is calculated as Eq. (2) as follows [47]:

$$C_b\left(n_i\right) = \sum_{w \neq \acute{w} \in V} \frac{\delta_{w\acute{w}}(n_i)}{\delta_{w\acute{w}}} \tag{2}$$

where $\delta_{w\acute{w}}$ refers to the shortest routes that exist between the two nodes w and $\acute{w}$. Also, $\delta_{w\acute{w}}(n_i)$ is a number of shortest routes that cross node n_i too. Since the accurate calculation of betweenness centrality in very big networks is so costly, we used [48] to calculate approximate betweenness centrality. Degree centrality is the very first and most prevalent centrality criterion in a node. The higher degree of a node will lead to more nodes connected to it and there would be more probability of data transfer between the nodes to be efficient. Through studying real data in many social networks, it was found that the degree distribution of these networks follows power law. In power law distribution, a small number of nodes exist with high degrees which are considered as central nodes in most networks, but degree centrality solely depends on one-hop neighbors. Also, in the Katz centrality criterion, a node is important only when it is connected to other important nodes or if it has lots of neighbors. As centrality measures the number of direct neighbors of a node, Katz centrality measures the number of nodes that are accessible through a route that starts from a node. In other words, Katz centrality considers neighbors with high degrees, too. As it has been represented in Fig. 1, Katz centrality is known as a hybrid between degree and eigenvector and such a concept has been presented in Eq. (3) as follows [49]:

$$C_K\left(n_i\right) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \alpha^k \left(A^k\right)_{ij} \tag{3}$$

where, the coefficient α is known as a reduction coefficient that leads to Katz centrality, another type of eigenvector centrality.

We have carried out pruning and listing based on betweenness, centrality, closeness, Katz, root nodes, and seed set of snapshot graph G_{t-1} criteria. In fact, we reduce the number of snapshot graph nodes G_t by about 50% through this

activity. Also, to reduce the time complexity of the computation of each of the above criteria, we have utilized the methods proposed in [46,48].

4.3. Independent cascade diffusion model

We have investigated about influence maximization of individuals on the diffusion model based on time-discrete that means independent cascade diffusion model [50]. The Independent cascade diffusion method has been described in Algorithm 2. In this model, the whole social network is represented as a graph through which the nodes represent individuals and the edges show their relationships. After the selection of the initial activation nodes, the influence should start to diffuse within the whole network. The weight of edges represents a number between zero and one which shows their activation probability. When a node like V is activated, it can activate the neighboring nodes with a probability. This probability equal to edge weight represents a relation between node v and node u. This activation with the probability predetermined can only happen once and the sequence of activation efforts of the inactive neighbors is optional.

Algorithm 2: Independent Cascade Model

Input:
Dynamic Network G_t
The Set of Seed Nodes A_0
Number of Monte-Carlo simulations MC
Probability of influence node λ_{ij}

Output:
Average Number of Nodes Influenced

```
spread = 0
repeat:
    τ ← False    τ: Has The Propagation Ended?
    A ← A0       A: Active Nodes
    B ← A        B: the set of nodes activated
    while not τ do
        nextB ← ∅;
        for each n ∈ B do
            for each neighbour m of n , where m ∉ A, do
                with probability λij, add m to nextB
        B ←nextB
        A ←A ∪ B
        if B is empty then
            τ← True
    Spread. Append(Len(A))
until MC
return mean(Spread)
```

The success or failure of activation by node u could not affect the activation of the v node. The process continues to be administered until the time there is not any other possible activation. This process is independent because any node activation is independent from the activation trend. In this study, the influence probabilities were considered to be the same.

4.4. Generalized genetic algorithm

In this stage, we first produce several random responses for the influence maximization problem based on the outputs from the previous stage. This set of responses is called the initial population and each response is called a chromosome. During the next stage, these chromosomes are assessed and those chromosomes that can represent optimal responses for the influence maximization problem (our target) will have a better chance to be selected, compared to weaker responses. Then using the operators of the generalized genetic algorithm, we crossover the selected chromosomes and create a mutation. Finally, we aggregate the current population with the new population resulted from the crossover and mutation in the chromosomes. Then, we extract the optimal responses and continue the above stages for the predetermined repetition numbers. The basic structure of the generalized genetic algorithm has been represented in Algorithm 3.

Table 1
Generalized genetic algorithm parameters.

Parameter	Value
Population size	100
Mutation rate	0.3
Crossover rate	1.0
Number of generations	$100 * K$
Number of elites	2 is for $k = \{10, 20\}$ s

Algorithm 3: Proposed DGA

Input:
Dynamic network G_t
The size of seed nodes k
Set of Candidate Nodes
Set of Seed Previous
Output:
Seed set S_t

```
InitializationPopulation ()
    CreateInitialPopulation ()
    Evaluate ()
repeat
    Select Parents ()
    Crossover ()
    Mutation ()
    Replacement ()
    Evaluate ()
    Elitism ()
until Number of Generations
return Best Solution
```

Crossover: We use a crossover operator for recombining the two chromosomes. Here the aim is to produce better chromosomes. In the reproduction stage, new chromosomes (with different values from the major population) are not formed in the population. In this stage (and after the operations resulted from the crossover operator), new chromosomes are formed through data (gene) interaction among the chromosomes. According to Algorithm4 and in a crossover operator, we first choose two chromosomes out of the initial population as the parent based on the selection by the roulette wheel. Then we integrate them with each other and arrange them in an ascending sequence based on the number of nodes, and then extract their offspring alternatively. This method guarantees that the produced chromosomes do not entail any repetitive nodes.

Algorithm 4: Crossover

Input:
Initialization Population
Number of Crossovers
Output:
Crossover Population

```
repeat
    Chromosome A, B← Select Parents ()
    Chromosome M← Merge (Chromosome A, B)
    Sort (Chromosome M)
    Offspring's ←split (Chromosome M)
    Evaluate (Offspring)
until Number of Crossovers
return Crossover Population
```

Mutation: Mutation operator is one of the most important evolving processes for achieving the optimal response in genetic algorithms. In mutation operator, new data are randomly added to the search process in the genetic algorithm. Such an important feature helps the genetic algorithm to avoid being trapped in the local optimum.

Algorithm 5: Mutation

Input:
Initialization Population
Number of Mutation

Output:
Mutation Population

repeat
  Chromosome A← Select Random (Population)
  Select Node ← Mindescendants (Chromosome A)
  Remaining Nodes← Select Random (V – D)
  Exchange the node with the replacement
  Evaluate ()
until Number of Mutation
return Mutation Population

When the crossover and reproduction operations are repeatedly done on chromosomes in consecutive generations, the population of chromosomes or the candidate responses tend to become homogeneous. The mutation operator helps the genetic algorithm to increase diversity in the chromosome population or the candidate responses. According to Algorithm5, we first select the initial population based on the roulette wheel in the mutation operator and then substitute the node with the least offspring in the chromosome with a new node resulted from the difference between the total nodes and the total set of the offspring from the selected chromosome randomly. The complete list of the parameters of the generalized genetic algorithm is presented in Table 1. We calculated these parameters by testing different values.

4.5. Extraction and saving

Regarding the studies and based on the report [51], referring to the fact that in social networks such as Twitter that is a dynamic network, there exists about 9% of total relations changes within a month. For example, a usual user with 100 followers could add to the followers 10 percent while he/she can lose about 3 percent of the followers during one-month period. We list and save a node as a member of the seed set in the snapshot graph G_t in this stage and use it in the next snapshot graphs during the pruning and listing stages. For snapshot graphs G_{t+1} we do updating the list and repeat the previous stages for the upcoming snapshot graphs. The fitness function used for influence maximization is based on independent cascade diffusion model. Since the diffusion process is random, it is executed as a series of Monte Carlo repetitions and the average number of activated nodes are used as the fitness amount. Regarding the structure changes in social networks during the pass of time, the seed set also changes as time passes. Finally, we have a more efficient dynamic seed set than the previous methods for a certain timestamp considering the scalability and dynamicity of social networks.

5. Evaluation of the proposed method

To compare the proposed algorithm with other existing algorithms, we use a dataset presented in SNAP[1] according to Table 2 below. The simulated tests by Python were done on cloud service Colab and using GPU processors and the parameters in tests were arranged as follows. The number of simulation runs is 50 and for each candidate seed set, we run 100 Monte Carlo simulations to estimate the influence spread and the influence probability was considered to be equal to 0.5. These three parameters were the same in all tests. Also, the size of the seed set was adjusted to be between 10 and 50. The proposed algorithm was measured in comparison with other algorithms presented, regarding a varied number of nodes and edges in three different snapshots of t_a, t_b, and t_c within the same time-span in a way that $t_a < t_b < t_c$.

Stack Overflow dataset: users send their questions on the Stack Overflow website and receive responses from other users. Users may give opinions both in questioning and responding stages, and finally, a temporal network is gained within a certain time interval.

Wiki Talk dataset: this dataset is a temporal network representing Wikipedia users editing discussion pages of each other. The edge u to v during time t means that the user u has edited the discussion page of user v during time t.

[1] https://snap.stanford.edu/index.html.

Table 2
The characteristics of social networks.

Dataset	Stack overflow	Wiki talk
Nodes	1646338	1140149
Temporal edges	25405374	7833140
Edges in static graph	11370342	3309592
Timespan	2773 days	2320 days

Table 3
Run time (s) of different algorithms.

N_{seed}	HIGHDEG			SDISC			RIS			DGA		
	10	30	50	10	30	50	10	30	50	10	30	50
Stack Overflow	91.79 92.28	98.75 89.75	96.81 90.83	112.43 106.05	214.17 182.13	372.91 300.85	208.94 186.05	216.76 203.34	204.70 201.66	1986.48 1216.48	3745.96 2831.42	7993.94 5021.42
Wiki Talk	100.74 147.45	101.45 145.07	100.38 147.09	111.07 163.14	166.09 247.84	257.04 398.22	132.38 189.72	130.5 187.51	136.73 189.64	1527.54 2845.42	3327.54 4931.06	6827.54 9621.43

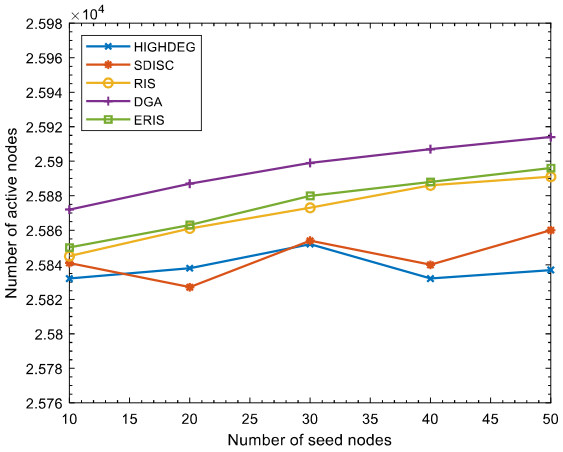
Several heuristic methods have been proposed to find appropriate solutions for the influence maximization problem. The ERIS algorithm proposed by Hautahi et al. [52] uses the concepts of the reverse influence sampling for calculating the exact solutions for influence in order to evaluate the empirical accuracy of greedy solutions in the influence maximization problem, The HIGHDEG is a greedy heuristic method that simply adds the nodes with the highest values of out-degree to the seed set [50]. Single discount (SDISC) is a refinement of the HIGHDEG proposed by Chen and et al. [22]. They considered the number of neighbors of the assessed node that are within the seed set as the new index, subtracted this value from the value of out-degree, and utilized the resulted value as the single degree. The Reverse Influence Sampling (RIS) was suggested by Brogs et al. [18]. It is a sketch-based method that, instead of using direct simulation of the influence processes, carries out reverse simulations to construct sketches comprised of a set of nodes for estimating the influence maximization problem effectively.

Fig. 2 represents the results of measurements carried out on the Stack Overflow dataset in three different snapshots of a certain timestamp and with different amounts of the seed set. As it can be seen in figures above, our proposed algorithm (DGA) has shown the greatest influence (the number of active nodes) in different time spans in the network compared to other methods such as the algorithm based on reverse influence sampling (RIS) posed by Brogs to avoid the limitations in traditional greedy algorithms to maximize the influence under independent cascade model and higher degree greedy heuristic algorithms and single-discount models. Also, as the size of the set of nodes and edges in our proposed algorithm increases, we encounter better results compared to other methods. As we compare the results in more detail, it can be seen that when the size of the seed set and the number of network nodes is low, there are not any clear differences between the results of our method and the results gained by other algorithms. But, as the number of nodes, edges, and the size of the seed set increase, the results of our proposed algorithm work better than other greedy heuristic algorithms. Of course, the greedy heuristic algorithms compared, have had the least time spans. But in big networks, these algorithms have had a relatively low accuracy in recognizing influential individuals due to a lack of conscious selection as the size of the seed set increases. In this way, our proposed algorithm could show a better performance compared with RIS algorithms. Results of figures *a*, *b*, and *c* in Fig. 2 showed that the proposed algorithm is more scalable and has had a higher accuracy in comparison with other simulation algorithms.

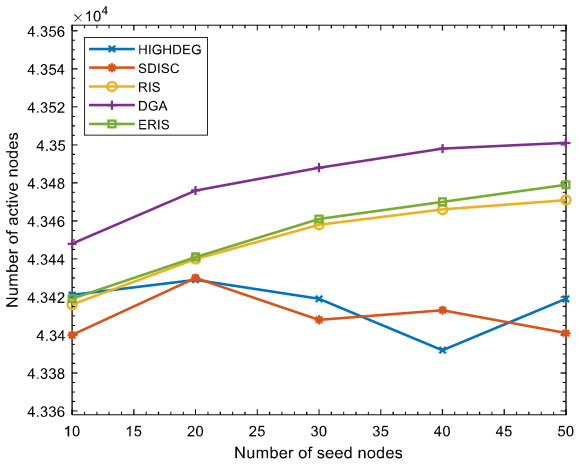
In Fig. 3 the influence spread on the wiki-talk dataset has been investigated for different numbers of nodes and edges. When the number of nodes and edges in a network increases, our proposed algorithm has shown better results compared to other algorithms studies due to having optimal search areas, competition and evolve of the initial population selected based on centrality criteria and also due to mutations carried out to avoid being trapped in local optimum. It should be noted that as the size of the seed set increases in the proposed algorithm due to the diversity of resolutions and competition between the nodes, influence spread increases but after a while and due to the network structure, this amount could not be increased considerably.

In terms of efficiency, we compared the run time of four algorithms, including HIGHDEG, SDISC, RIS and DGA with the number of seed sets 10, 30 and 50 in two different time stamps in the Stack Overflow and Wiki Talk datasets. The results are presented in Table 3

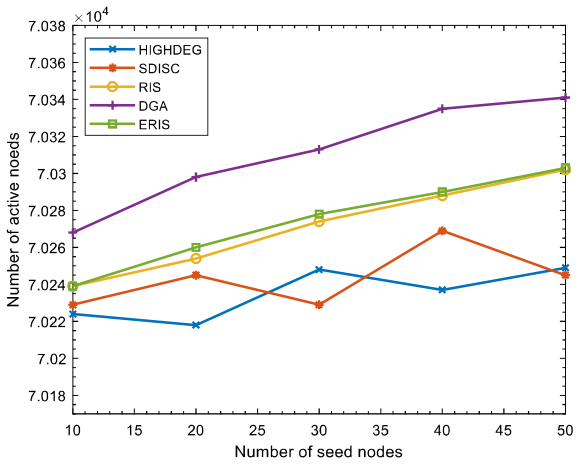
As shown in Table 3, the run time of the proposed algorithm is slower than that of other algorithms, but the accuracy of identifying influential individuals in the proposed algorithm is higher than in the other algorithms. Although the performance of the proposed algorithm is influenced by its parameters, but its results are much better than those of other simulated algorithms. In some cases, the proposed algorithm is affected by the network structure, although it provides a variety of solutions, but the optimization process becomes more complicated and the run time of the algorithm increases. In general, the proposed algorithm works better in terms of efficiency.



(a) the results of influence spread on the dataset in snapshot t_a

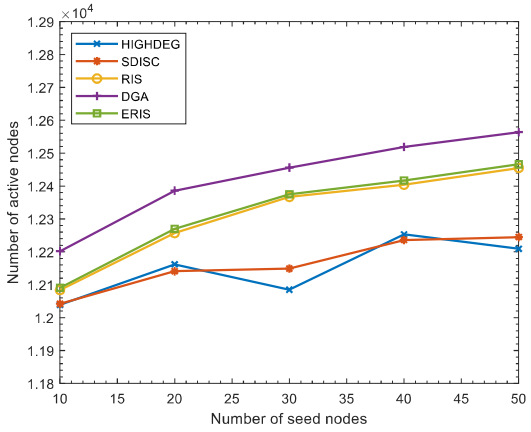


(b) the results of influence spread on the dataset in snapshot t_b

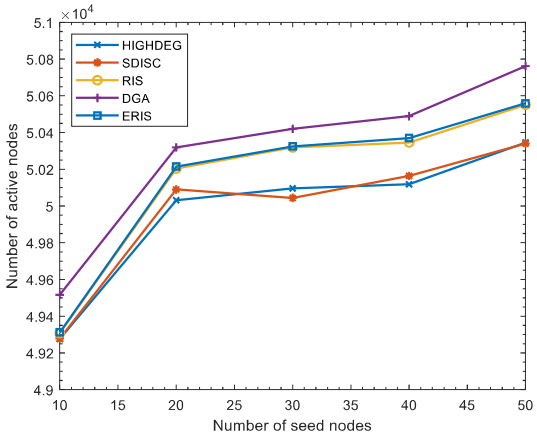


(c) the results of influence spread on the dataset in snapshot t_c

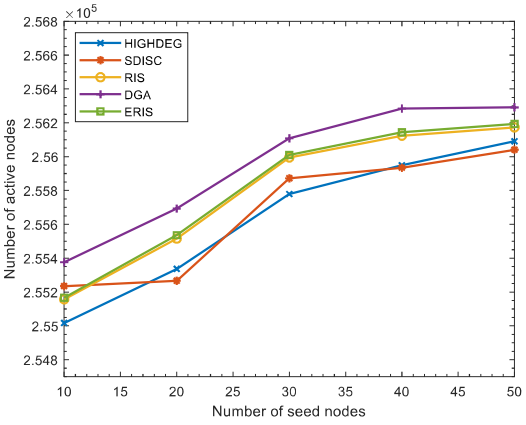
Fig. 2. Comparison of the numerical results on Stack Overflow datasets in a timestamp.



(a) the results of influence spread on the dataset in snapshot t_a



(b) the results of influence spread on the dataset in snapshot t_b



(c) the results of influence spread on the dataset in snapshot t_c

Fig. 3. Comparison of the numerical results on Wiki-Talk dataset in a timestamp.

6. Conclusions

Given the fact that most social networks have very big sizes and are dynamic and evolving, recognizing influential individuals for influence spread using the traditional methods will be a complex and time-consuming task. In this study, we have proposed a dynamic generalized genetic algorithm. This method identifies an optimal set of effective nodes in order to maximize the influence of individuals in social networks in each snapshot based on a certain timestamp. Using experimental results, we compared the performance of the proposed algorithm with sketch-based algorithms, greedy heuristic algorithms, and reverse influence sampling algorithms. Although the performance of the proposed algorithm was affected by its parameters, the results were much better than those of the simulated algorithms. But as the number of individuals recognized in an almost big network increased, the calculation costs for obtaining the centrality values used were high and the process was time-consuming. But we can use centrality values that have lower calculation costs (such as node degree). Also, it should be noted that due to the structure of social networks, it will be better to utilize centrality criteria such as "betweenness" and "closeness". Considering the recent studies conducted on community detection with deep learning in scalable networks, and in order to reduce computational costs and maintain the efficiency and accuracy of identifying the influential individuals on large-scale dynamic social networks in future work, we will focus on adjusting the structure of the initial population based on community detection with deep learning and new centrality criteria to improve the computational cost.

CRediT authorship contribution statement

Jalil Jabari Lotf: Investigation, Conceptualization, Methodology, Algorithm design, Software, Data curation, Experimental evaluations, Writing – original draft. **Mohammad Abdollahi Azgomi:** Investigation, Conceptualization, Supervision, Writing – review & editing. **Mohammad Reza Ebrahimi Dishabi:** Investigation, Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Not applicable

Code availability

All code for data analysis associated with the current submission is available from the corresponding author upon reasonable request.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] G. Chen, X. Wang, X. Li, Fundamentals of Complex Networks: Models, Structures and Dynamics, John Wiley & Sons, 2014.
[2] S. Banerjee, M. Jenamani, D.K. Pratihar, A survey on influence maximization in a social network, Knowl. Inf. Syst. 62 (9) (2020) 3417–3455.
[3] N.N. Daud, S.H. Ab Hamid, M. Saadoon, F. Sahran, N.B. Anuar, Applications of link prediction in social networks: A review, J. Netw. Comput. Appl. 166 (2020) 102716.
[4] X. Su, et al., A comprehensive survey on community detection with deep learning, 2021, arXiv preprint arXiv:2105.12584.
[5] S. Aghaalizadeh, S.T. Afshord, A. Bouyer, B. Anari, A three-stage algorithm for local community detection based on the high node importance ranking in social networks, Physica A 563 (2021) 125420.
[6] X. Ma, J. Wu, S. Xue, J. Yang, Q.Z. Sheng, H. Xiong, A comprehensive survey on graph anomaly detection with deep learning, 2021, arXiv preprint arXiv:2106.07178.
[7] S.S. Singh, K. Singh, A. Kumar, B. Biswas, Influence maximization on social networks: a study, Recent Adv. Comput. Sci. Commun. (Formerly: Recent Patents on Computer Science) 14 (1) (2021) 13–29.
[8] B. Chang, T. Xu, Q. Liu, E.-H. Chen, Study on information diffusion analysis in social networks and its applications, Int. J. Autom. Comput. 15 (4) (2018) 377–401.
[9] U. Can, B. Alatas, A new direction in social network analysis: Online social network analysis problems and applications, Physica A 535 (2019) 122372.

[10] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 61–70.
[11] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 57–66.
[12] D. Kempe, J. Kleinberg, É. Tardos, 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, 2003, pp. 137–146.
[13] D. Kempe, J. Kleinberg, É. Tardos, Influential nodes in a diffusion model for social networks, presented at the Proceedings of the 32nd international conference on Automata, Languages and Programming, Lisbon, Portugal, 2005, in: Lecture Notes in Computer Science, 3580, 2005, pp. 1127–1138.
[14] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2007, pp. 420–429.
[15] A. Goyal, W. Lu, L.V.S. Lakshmanan, CELF++: optimizing the greedy algorithm for influence maximization in social networks, in: Presented at the Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 2011.
[16] Y. Tang, X. Xiao, Y. Shi, Influence maximization: Near-optimal time complexity meets practical efficiency, in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, 2014, 75–86..
[17] H.T. Nguyen, M.T. Thai, T.N. Dinh, A billion-scale approximation algorithm for maximizing benefit in viral marketing, IEEE/ACM Trans. Netw. 25 (4) (2017) 2419–2429.
[18] C. Borgs, M. Brautbar, J. Chayes, B. Lucier, Maximizing social influence in nearly optimal time, in: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2014, pp. 946–957.
[19] J. Lv, J. Guo, H. Ren, Efficient greedy algorithms for influence maximization in social networks, JIPS 10 (3) (2014) 471–482.
[20] L.C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (3) (1978) 215–239.
[21] A. Landherr, B. Friedl, J. Heidemann, A critical review of centrality measures in social networks, Bus. Inf. Syst. Eng. 2 (6) (2010) 371–385.
[22] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Presented at the Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009, http://dx.doi.org/10.1145/1557019.1557047.
[23] K. Jung, W. Heo, W. Chen, Irie: Scalable and robust influence maximization in social networks, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 918–923.
[24] S. Galhotra, A. Arora, S. Roy, Holistic influence maximization: Combining scalability and efficiency with opinion-aware models, Presented at the Proceedings of the 2016 International Conference on Management of Data, San Francisco, California, USA, 2016.
[25] Y. Zhao, S. Li, F. Jin, Identification of influential nodes in social networks with community structure based on label propagation, Neurocomputing 210 (C) (2016) 34–44.
[26] F. Ullah, S. Lee, Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization, Physica A 486 (2017) 968–985.
[27] D. Cai, Z. Wang, N. Wang, D. Wei, A new method for identifying influential nodes based on DS evidence theory, in: 2017 29th Chinese Control and Decision Conference (CCDC), IEEE, 2017, pp. 4603–4609.
[28] J. Shang, S. Zhou, X. Li, L. Liu, H. Wu, CoFIM: A community-based framework for influence maximization on large-scale networks, Knowl.-Based Syst. 117 (2017) 88–100.
[29] J. b. Guo, F. z. Chen, M. q. Li, A multi-objective optimization approach for influence maximization in social networks, in: Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018, Springer, 2019, pp. 706–715.
[30] A. Şimşek, K. Resul, Using swarm intelligence algorithms to detect influential individuals for influence maximization in social networks, Expert Syst. Appl. 114 (2018) 224–236.
[31] S. Agarwal, S. Mehta, Social influence maximization using genetic algorithm with dynamic probabilities, in: 2018 Eleventh International Conference on Contemporary Computing (IC3), IEEE, 2018, pp. 1–6.
[32] N. Wang, J. Da, J. Li, Y. Liu, Influence maximization with trust relationship in social networks, in: 2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), IEEE, 2019, pp. 61–67.
[33] A. Mohammadi, M. Saraee, Finding influential users for different time bounds in social networks using multi-objective optimization, Swarm Evol. Comput. 40 (2018) 158–165.
[34] X. Wu, L. Fu, J. Meng, X. Wang, Maximizing influence diffusion over evolving social networks, in: Proceedings of the Fourth International Workshop on Social Sensing, ACM, 2019, pp. 6–11.
[35] G. Tong, W. Wu, S. Tang, D.-Z. Du, Adaptive influence maximization in dynamic social networks, IEEE/ACM Trans. Netw. 25 (1) (2016) 112–125.
[36] M.M.D. Khomami, A. Rezvanian, N. Bagherpour, M.R. Meybodi, Minimum positive influence dominating set and its application in influence maximization: a learning automata approach, Appl. Intell. 48 (3) (2018) 570–593.
[37] D. Bucur, G. Iacca, Influence maximization in social networks with genetic algorithms, in: European Conference on the Applications of Evolutionary Computation, Springer, 2016, pp. 379–392.
[38] P. Krömer, J. Nowaková, Guided genetic algorithm for the influence maximization problem, in: International Computing and Combinatorics Conference, Springer, 2017, pp. 630–641.
[39] C.-W. Tsai, Y.-C. Yang, M.-C. Chiang, A genetic newgreedy algorithm for influence maximization in social network, in: 2015 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2015, pp. 2549–2554.
[40] K. Zhang, H. Du, M.W. Feldman, Maximizing influence in a social network: Improved results using a genetic algorithm, Physica A 478 (2017) 20–30.
[41] H. Zhuang, Y. Sun, J. Tang, J. Zhang, X. Sun, Influence maximization in dynamic social networks, in: 2013 IEEE 13th International Conference on Data Mining, IEEE, 2013, pp. 1313–1318.
[42] M. Han, M. Yan, Z. Cai, Y. Li, X. Cai, J. Yu, Influence maximization by probing partial communities in dynamic online social networks, Trans. Emerg. Telecommun. Technol. 28 (4) (2017) e3054.
[43] Y. Wang, Q. Fan, Y. Li, K.-L. Tan, Real-time influence maximization on dynamic social streams, Proc. VLDB Endowment 10 (7) (2017) 805–816.
[44] N. Hafiene, W. Karoui, L.B. Romdhane, Influential nodes detection in dynamic social networks: A survey, Expert Syst. Appl. 159 (2020) 113642.
[45] W. Oueslati, S. Arrami, Z. Dhouioui, M. Massaabi, Opinion leaders' detection in dynamic social networks, Concurr. Comput.: Pract. Exper. 33 (1) (2021) e5692.
[46] E. Cohen, D. Delling, T. Pajor, R.F. Werneck, Computing classic closeness centrality, at scale, in: Presented at the Proceedings of the Second ACM Conference on Online Social Networks, Dublin, Ireland, 2014, http://dx.doi.org/10.1145/2660460.2660465.
[47] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1) (1977) 35–41.
[48] M. Riondato, E.M. Kornaropoulos, Fast approximation of betweenness centrality through sampling, Data Min. Knowl. Discov. 30 (2) (2016) 438–475.
[49] P. Bonacich, Simultaneous group and individual centralities, Social Networks 13 (2) (1991) 155–168.
[50] J.K. David Kempe, Éva Tardos, Maximizing the spread of influence through a social network, Theory Comput. 11 (2015) 105–147.
[51] S.A. Myers, J. Leskovec, The bursty dynamics of the twitter information network, in: Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 913–924.
[52] H. Kingi others, A numerical evaluation of the accuracy of influence maximization algorithms, Soc. Netw. Anal. Min. 10 (1) (2020) 70.