Influence Maximization based on a Non-dominated Sorting Genetic Algorithm

Elaf Adel Abbas

College of Information Technology, University of Babylon, elaf1982adil@gmail.com

Huda Naji Nawaf

College of Information Technology, University of Babylon, halmamory@itnet.uobabylon.edu.iq

Follow this and additional works at: <https://kijoms.uokerbala.edu.iq/home>

Part of the [Computer Sciences Commons](#)

Recommended Citation

Abbas, Elaf Adel and Nawaf, Huda Naji (2021) "Influence Maximization based on a Non-dominated Sorting Genetic Algorithm," *Karbala International Journal of Modern Science*: Vol. 7 : Iss. 2 , Article 5.

Available at: <https://doi.org/10.33640/2405-609X.2891>

This Research Paper is brought to you for free and open access by Karbala International Journal of Modern Science. It has been accepted for inclusion in Karbala International Journal of Modern Science by an authorized editor of Karbala International Journal of Modern Science.

University of
Kerbala

Influence Maximization based on a Non-dominated Sorting Genetic Algorithm

Abstract

Influence Maximization (IM) is a problem represented by a set of users who are specified in advance and are usually called the seed. The latter can influence their friends, who can in turn influence others and so on until it reaches the largest number of users within the network. This issue is of ultimate importance in a variety of fields. In the current study, a Non-dominated Sorting Genetic Algorithm II (NSGA-II) has been adopted in influence maximization to produce the so-called NSGAI based IM algorithm (NSGAI-IM). Principally, the population should be represented with individuals of variable lengths as the seed group, and the diffusion model should be designed so as to formulate its multi-objective function. In the context of individual representation, the nodes have been pseudo-randomly chosen using the centrality measures (based on high centrality nodes as degree, closeness, and eigenvector). As for the multi-objective function, increasing the coverage size of influence and decreasing the number of seed nodes as far as possible have been set as the conflicting objectives. Weighted Integration Cascade (WIC) has been suggested as an improved version of the Independent Cascade (IC) diffusion model. It has proven to be effective in the performance of the NSGAI-IM algorithm. In evaluating the proposed optimization model, two real-world social network datasets have been used: Facebook wall posts, and Digg networks. The algorithm showed promising results as it could relatively improve the solutions as compared with other methods, with an increased average of influential spread. Additionally, the WIC model has proven to be effective through the evaluation of the performance of the NSGAI-IM algorithm with other diffusion models.

Keywords

Social Networks, NSGA-II, Influence Maximization, Diffusion Model

Creative Commons License

This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Online social networking sites like Facebook, Google+, and Twitter have grown to be an integral part of people's daily lives and have changed the way people communicate and interact around the world [1]. Therefore, popular platforms have received a lot of attention from companies to promote their products and spread information, with the number of users increasing rapidly [2]. In the information-propagation process, the act of choosing a set of influential users to spread particular information (*i.e.* ideas, news, and trends) and thereby affect the behavior of other users [3] has attracted the attention of many researchers [4].

Online social networks allow users to interact, share, and collaborate with each other, which makes disseminating information between users easier. In today's world, online social networks are of use in a variety of disciplines [5,6], including marketing, healthcare, and political science due to their high ability to reach large audiences in a relatively shorter period of time. In this context, the goal of companies is to choose a small group consisting of the most influential users available on social networks, and let them spread their influence to as many potential customers as possible. This is considered an important aspect of viral marketing. From another point of view, it is possible to influence the results of democratic elections by disseminating misinformation or fake news through a group of influencers on social networks, thus influencing the decision of the voter [7].

The influence maximization summarizes the previous issue, which can be defined as a subset of users of size K that can be specified in advance, and are usually called seed or initial adopter [8]. The initial adopter can influence his friends, and these friends can in turn influence others, and so on. Therefore, the largest number of users is reached within the network. In fact, the influence maximization problem encompasses immense research challenges as follows:

- ❖ First, the way of modeling the process of information diffusion within the social network [9] has a great role in spreading the impact of any set of seeds in Influence Maximization (IM).
- ❖ Second, the high computational complexity to assess the spread of the influence of any seed group is due to the stochastic nature of information dissemination [10]. These theoretical results have

shown that traditional methods are rather difficult and tend to have limitations in terms of retrieving optimal seed sets, for which multiple stochastic and materialistic approaches have been proposed since that time.

- ❖ Third, novel features have recently been added to online social networks such as streaming content, topical analysis, and location-based services, to improve the performance by combining features such as topic information, time, and location [11].

To meet the challenge of complexity in computations, research studies have turned to evolutionary algorithms to find a wide range of solutions for combinatorial optimization problems. It has proven successful because they use more flexible research and improvement strategies, and they can adapt well to conditions associated with different situations of problems to be solved [12].

Consequently, the influence maximization problem could be considered as a single objective optimization issue, as most of the preceding researches only aimed to study the greater influence spread, without taking into consideration the number of seeds, which is left to the researcher. However, current research tends to address two interlinked concepts, namely studying the greatest spread of influence possible while using the smallest number of seeds for spreading this influence. Multi-Objective Evolutionary Algorithms (MOEA) have been proposed in this paper to formulate the influence maximization problem. MOEA has been a trade-off between conflicting objective functions with an Evolutionary Algorithm (EA), and produces a set of optimal solutions. Provided that the increase in influence and decrease in numbers of seed are two conflicting objectives, there should be a minimum number of seed nodes that have the best influence in the whole network. This paper has been structured as follows: Section 2 introduces the background and some relevant work concerning the influence maximization problem in social networks. The main materials and methods have been displayed in Section 3. The experimental results and discussion are presented in Section 4, and the concluding remarks of this work are given in Section 5.

2. Background and related works

The influence maximization problem has initially been introduced by Domingos and Richardson (2001),

suggesting a probabilistic solution. They considered the influence maximization as an algorithmic problem and suggested a Markov random modeling to obtain the probabilistic solution of the impact propagation process. The authors in Ref. [13] proposed a heuristic method for solving this issue, as well as many other researchers whose solutions are summarized in Table 1.

Kempe et al. proposed the HIGHDEG as a heuristic method for studying the influence of maximization as a combinatorial optimization problem. What has been focused on in the former works are the two most popular propagation models for the influence diffusion process, namely the Independent Cascade (IC) model and Linear Threshold (LT) model. Furthermore, a greedy algorithm has been proposed that could obtain an approximation ratio of $\left(1 - \frac{1}{e}\right)$ for realizing optimal solutions. Therefore, the Monte Carlo simulation running is repeated to achieve an acceptable approximate solution. Indeed, large scale social networks are time-consuming and may be difficult to apply [14].

The Cost-Effective Lazy Forward (CELf) algorithm is introduced by Leskovec et al., and it is one of the several methods proposed for speeding up the running time of the greedy algorithm, typically by using sub-

modularity of propagation. The results of the experiment indicated that CELf runs 700 times faster than a greedy algorithm, while at the same time achieving near-optimal solutions [15].

From another view, centrality measures like Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), and Eigenvector Centrality (EC) have been used to identify influences in the network. The Degree centrality might return suboptimal solutions because it does not take into account the global characteristics of networks but rather depends on the attributes of the nodes [16]. Since BC and CC are global measures and are based on the structural centralities of the network, these methods have certain computational complexity when used with large scale networks [17].

Furthermore, EC is based on iterative refinement centralities, in which the influential node is defined according to the importance and number of adjacent nodes directly related to this node, in addition to the importance of the node itself [18].

DegreeDiscount (SDISC) is another heuristic method to select seed nodes, the efficiency of seed set selection depends on network topology, they argued that when selecting a new node u as a seed, and then

Table 1
Overview of methods for related influence maximization algorithms.

Methods	Reference and Date	Scope	Short Description
Greedy	[14]: 2003	social networks	The influence of maximization as a combinatorial optimization problem
CELf	[15]: 2007	Real networks	Proposed to speed up the running time of the greedy algorithm by using sub-modularity of propagation.
Centrality measures	[16]: 2017	Real and artificial networks	It is easy to compute but might return suboptimal solutions.
	[17]: 2016		This method has certain computational complexity while based on the structural centralities of the network
	[18]: 2019		The method determines the importance of a node based on the importance of its neighbors.
	[19]: 2009		The qualification of seed set selection leans on network topology
Evolutionary algorithms	[8]: 2018	NetScienc NetGRQC NetHEP	Improves the efficiency of greedy algorithms in large scale social networks that are time-consuming
	[23]: 2018	wiki-Vote amazon0302	This algorithm uses the local optimal solutions to evolve towards the global optimal solution
MOEA	[25]: 2017	ego-Facebook ca-GrQc	This algorithm has two conflicting objective functions; maximizing the influence spread and minimizing the size of the individuals (number of seeds).
NSGA-II	[27]: 2018	WikiVote NetHEPT Epinions	It has taken conflicting objectives, namely the increase in coverage size of influence and the decrease of the propagation time
	[28]: 2019	Real networks	It is a multi-objective optimization method to formulate the spread cost is considered to be a goal for improving the evolutionary algorithms.

the prior knowledge has been used for the node selection [19].

Evolutionary algorithms are a potential candidate to solve the issue of influence maximization within a social network. For instance, the genetic algorithm has been suggested as a diffusion model. Therefore, it models the spread of information through static and dynamic social networks [20].

Concerning the new greedy algorithm presented by Tsai et al., it targets the improvement of the general greedy algorithm by utilizing the characteristics of the genetic algorithm in local research, so as to find the optimal solution [21].

As for the work presented in Ref. [22], it is considered to be one of the first works that use the genetic algorithm as a viable tool for solving the issue of improving the influence, particularly in the absence of any previous knowledge about the characteristics of the social network topology. The Degree-Descending Search (DDS) modern evolutionary algorithm in Ref. [8] has been introduced to improve the efficiency of greedy algorithms by evading repeated simulation that is time-consuming. The degree-descending search strategy has generated a seed set that affects spread commensurate to the degree centrality.

The study conducted by Silva et al. in Ref. [23] applied the genetic algorithm with different strategies to form the initial population, after which one individual has been chosen. It consists of nodes with a high central measure, such as Closeness, Betweenness, PageRank, and Degree, whereas the remaining individuals of the population have been chosen randomly. A linear threshold model has been used as a fitness function for measuring the influence propagation of every candidate individually. The experimental results indicated that the inclusion of an individual made up of nodes with high central quality achieved relatively better results than the case where no individual is included in the initial population.

Recently, several types of multi-objective optimization algorithms were proposed to solve a number of influence maximization issues within social networks. The authors in Ref. [24] investigated a set of key players in the whole network by applying multi-objective optimization for maximizing the two objective functions (Eigenvector centrality measure and distance measure) among key players.

Similarly in Refs. [25,26], a Multi-Objective Evolutionary Algorithm (MOEA) has also been proposed. However, it has rather conflicting objective functions so that it simultaneously maximizes the influence spread and minimizes the size of individual

(number of seeds) by returning the optimal solution. This approach has been evidence to be able to overcome and establish two heuristics methods: the High degree (HIGHDEG) and Singlediscount (SDISC). The approach has been applied onto (ego-Facebook, and ca-GrQc, soc-ePinions1) real-world social networks.

The NSGA-II algorithm has been suggested by the authors in Ref. [27], which is one of the most powerful multi-objective optimization algorithms to detect different seed sets that have the highest influence at various propagation times. It has taken into consideration two conflicting objectives, namely the increase in coverage size of influence and the decrease of propagation time, thereby making available the set of all candidate optimal solutions to support a decision-maker. This will straighten the trade-off between the two conflicting objectives. The results were obtained from the application of the NSGA-II algorithm in three real data sets (WikiVote, NetHEPT, Epinions), which prove to outperform previous works in terms of influence diffusion.

Finally, the multi-objective optimization method to formulate influence maximization problems has been presented by Guo et al. in Ref. [28] as well. In fact, the spread cost is considered to be among the goals that have been tackled for improving the evolutionary algorithms. Experimental results indicate the ability to find the optimal set of seeds in such a manner that they have the largest spread of influence with a minimum cost of the spread, thus being efficient and accurate. However, in their work, seed nodes that give more spread with less cost of activating are chosen, and the cost of user activating has been adopted as another goal in addition to the volume of spread. The work aimed to select the influencers who are with little user activation cost and large spread.

In our work, both the coverage size and the number of influencers are to be considered as the objectives of influence propagation. In fact, increasing coverage size and decreasing the number of influencers are two conflicting objectives that are considered in the selection of seed nodes. This employs an NSGA-II algorithm as a solution that introduces a multi-objective optimization problem.

3. Materials and methods

3.1. Diffusion models

Linear threshold (LT) and Independent Cascade (IC) are the basic diffusion models that have been used to simulate the manner of interactions and information

spread among users within the social network. In both models, the node in a social network has either one of two states: active or inactive. The active state of the node is when the node has already been influenced, while the inactive state is when it has not been affected yet [29].

As a matter of fact, if the node's state is inactive and most of its neighboring nodes are active, the chance of it becoming active increases consequently. In general, the contagion proceeds in discrete time steps in both models. The linear threshold model is presented by Kempe et al., in Ref. [14]. In this model, there is a weight for each edge between the two nodes. The weight value expresses the level of influence of one node on another. For instance, when W_{ij} is the weight of the edge from i to j , then the value of W_{ij} shows the level of the effect i to j . Therefore, the sum of the weight of input edges must be equal or less than one. In order for the node to be influenced, the sum of the weighted input of the influenced neighbors must exceed a certain threshold. Initially, each node in the network has a threshold with a uniformly random value between zero and one, and this threshold has to be less than the total summation of weights of input edges required to influence the node.

Randomly, a set of active nodes S has been chosen in time t . At the time $t+1$, the node will be active when the total weight of the adjacent active nodes is greater than the threshold for this node [30]. Concerning the independent cascade, a set of active nodes S at the time t is chosen randomly, where each edge in the network holds the probability of diffusion. Whenever the node is active in time $t+1$, it tries to activate its neighbors with a certain probability propagation P , and for one time only. The probability P_{ij} is the probability of i influencing j , meaning that the edges of the network can diffuse information with the probability P , but not with a probability $1-P$ [31].

In literature, the researchers have tried to develop IC in terms of edge probability instead of setting the probability with a constant value for all edges. In reality, the probability of diffusion should be different between any different pair of nodes. Therefore, the Weighted Cascade (WC) model has been suggested to compensate for the deficiency, according to the degree of the neighbor of becoming an active node [14]. In other words, the same probability value is assigned for all edges of active node j ; where the probability of edge (i, j) is equal to $p(i, j) = \frac{i}{\text{in-degree}(j)}$. Thus, the probability value for any edge between the neighbor j and active node i is the same value apart from the neighbor.

The Persuasiveness-weighted cascade model has been proposed to improve the IC models [30]. This model depends on the degree of the node in addition to the degree of its neighboring nodes. The propagation probability between the two nodes i and j could be calculated as shown in Equation (1):

$$P_{ij} = \frac{d_i}{\sum_{k \in N_j} d_k} \quad (1)$$

where the d_i represents the degree of node i ; $\sum_{k \in N_j} d_k$ is the sum of the degree of nodes that are neighbors of node j . This model shows that the node with a high degree is of more influence than a node with a lower degree. The symbols used throughout the paper are listed in Table 2.

In the current work, the new Weight Integration Cascade (WIC) model is proposed with the understanding that the interaction of a node is as important as its relationships. The user activity and degree can be merged in one measure that can locally define the propagation probability between any two nodes. According to that, the propagation probability P_{ij} between nodes i and j could be defined as follows:

$$P_{ij} = \alpha \frac{d_i}{\sum_{k \in N_j} d_k} + \beta \frac{I_{ij}}{\sum_{k \in N_j} I_{kj}} \quad (2)$$

The first term of Equation (2) denotes the degree of the active node i divided by the sum of degrees of the neighbor nodes that have a direct edge to i , while the second term represents the past interactions between node i and j divided by the sum interactions of node j with its other neighbor nodes. α and β are the constant value between (0, 1).

Supposing the following scenario; consider a sub-graph $G' \subseteq G$ that have inactive node j , and assuming that the persuasiveness of each neighbor is represented by the degree of the neighbor. The total degree of node j 's neighbors is 10, and the degrees of neighbors nodes are y, x, z , and i are 3, 1, 2, 4, respectively. Conversely, we have the probabilities are set as 0.3, 0.1, 0.2, 0.4,

Table 2
Symbol description.

Symbol	Description
P_{ij}	The propagation probability between nodes i and j
d_i	The degree of node i
d_k	The degree of node k
N_j	Set of neighbors of i that have direct relationship with i
I_{ij}	The no. interactions from node i to node j
I_{kj}	The no. interactions from node k to node j
α and β	The constant values between (0, 1)

Fig. 1. Subgraph $\tilde{G}$

which reflect the degree of each neighbor. Meanwhile, the number of times interaction between j and neighbors node y , x , z , and i are 2,1,2 and 5. The total interaction of node j 's neighbors is 10 as shown in Fig. 1(a). The probabilities as 0.2, 0.1, 0.2, 0.5 reflect the interaction of each neighbor. Which means that the probability of propagating information from i to j is higher than the probability of information propagating from other neighbor nodes that are shown in Fig. 1(b). Hence, the fixation of the influence probability over each edge is not feasible in the real world scenario. Next, the focus will be on developing Multi-objective evolutionary algorithms to select seeds for efficient spreading.

3.2. NSGA-II algorithm

Usually, companies choose the influencing people within social networks and pay them money to encourage them to publish targeted information in that social network. In literature, several methods have been investigated to find seed nodes in the influence maximization field in order to achieve the highest prevalence [28].

The NSGA-II algorithm is the most powerful of the multi-objective evolutionary algorithms [27,32]. In this algorithm, a random parent population is initially created, after which a quick non-dominated sorting approach is employed to get the Pareto front which has the sets of solutions that are non-dominant to each other [23,33]. The non-domination Pareto front can be classified into multi-levels, where level 1 is considered to be the best level, followed by level 2 and so on. Thus, each individual in the population has been assigned a fitness value according to its level.

Essentially, the NSGA-II uses a binary tournament selection, then creates an offspring population by using crossovers at one point and mutation operators. At first, the population combines the offspring and parent populations to create a mating pool. It is followed by sorting the population based on its non-domination [34]. Obviously, the procedure is somewhat different after the initial generation, so that the next population is opted from individuals in lower Pareto fronts to create a new generation every time. If the number of individuals on the first level front is less than the size of the population, then the individuals from the next level are added and continue until it reaches the size of the population. Yet, whenever the number of individuals on the next level front is greater than that of need, then the crowded-comparison approach can be used. With respect to the previous approach, it computes the crowding distance between an individual and its neighbors.

The crowded-comparison approach can be added as a new parameter to select individuals in the next generation. Indeed, the selection process works by using fitness (ranking the level of Pareto fronts) and crowding distance of each individual so that if both individuals belong to the same level of Pareto fronts, the priority is for the individual that is situated in a lesser crowded region. With this approach, the selection process can be directed to the optimal Pareto front [34].

3.3. NSGA-II based IM algorithm (NSGAIIM)

In this work, the optimization of the influence maximization problem was formulated in form of a multi-objective evolutionary algorithm, according to

an improved model that is based on two opposing objectives. The first objective is to choose the smallest number of influential individuals and the second is to achieve the highest spread of information within the network. According to that, the NSGA-II algorithm had to be adapted to meet the goal and makes use of the additional information hidden within the network structure to be used when feasible. Some changes in the initial population were performed to impact the quality of the solution. A pool was created which contains the nodes whose centrality is high based on a threshold determined according to experience. This centrality is determined by one of the following: degree, closeness, or eigenvector. Thus, for each network to be analyzed, the three measures of centrality are applied and the pool contains each node that has a high centrality, where the same node could have a high degree, closeness, and eigenvector. The individuals in the initial population are randomly selected from this pool. The former algorithm has been explained in terms of individual's representative, evolutionary operators, and finally the objective function.

3.3.1. Individuals representation

In the proposed scenario, the individual that is a candidate solution can be represented as a set of nodes (seeds), and the size of each individual is the number of all nodes within the network. An individual is represented in binary encoding, if a node is within the seed set, then 1 is assigned to its position in the

representation of individual, otherwise it is assigned to 0. However, this representation may not be implemented easily when the network is very large in size. Therefore, it is convenient to keep only the seed number rather than its indices, as illustrated in Fig. 2.

In this context, the individuals are characterized by variable rather than fixed length, as the latter depends on the number of nodes that are chosen within the seed nodes.

3.3.2. Evolutionary operators

Crossover is an evolutionary operator used to produce a new offspring of a population whose features have been inherited from those of parents' specifications. In this work, the simple one-point crossover [35] has been applied with some modification to avoid repeating the node in the same offspring, as demonstrated in Fig. 3. In fact, repeating the node for the same individual is not accepted in this scenario. Fig. 3(a) illustrates the basic one-point crossover, where the offspring 1 has node 12, which is repeated twice, while Fig. 3(b) states the modified version of the simple one-point crossover that is proposed in this work. In brief, the node from one parent is chosen randomly to be a crossover point.

In this context, offspring 1 can be generated by selecting the index of the node that is equal to or less than that of the crossover point, and larger than the crossover point from parent 1 and parent 2; respectively. On the contrary, offspring 2 consists of the

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1

(a) Individual representation in binary encoded in network with size 21 nodes

9	12	15	20
---	----	----	----

(b) Individual representation to minimize the amount space of storage in network size 21 nodes

Fig. 2. The representation of individual.

(a) simple one-point crossover

(b) randomly chosen node from parent 1

Fig. 3. Crossover method.

indexes of the nodes that are larger than that of the crossover point from parent 1, and equal to or less than that point of parent 2, as shown in Fig. 3(b).

As for the mutation which is a second operator, the traditional way has been used including the addition, replacement, or removal of one node from an individual. In particular, the replacement or addition of a node in an individual is conditional on non-repetition so that the individual does not have to contain the same node more than once.

3.3.3. Fitness function

Concerning the multi-objective, the best candidate solution is representing a set of seeds, which have to meet two conditions, the first being the best diffusion (*i.e.* the set of seeds fulfills maximum influence on others), and the second is realizing this diffusion with a minimum number of members within that particular set. It is worth mentioning that the improved IC model explained in Section 3 has been implemented in modeling the propagation.

The whole framework of the proposed NSGAI-IM algorithm is presented in the following detailed steps:

1. Calculate the node centrality for each node of the network.
2. Create a pool containing nodes that exceed the threshold of centrality measure (the threshold is determined according to the nature of the network).

3. Generate individuals with variable sizes of central nodes from the pool of nodes to create the initial population.
4. Evaluate the population.
5. Select two individuals.
6. Apply crossover and mutation to create the offspring set.
7. Evaluate the offspring set.
8. Merge the population as well as the set of offspring into a one set.
9. Building Pareto front from the set using a non-dominated sorting algorithm.
10. Calculate the crowding distance of individuals in each front and select the best N individuals according to their Pareto front and crowding distance.
11. If the stop condition not satisfied, return to step 4.

The flowchart of the proposed NSGAI-IM algorithm is illustrated in Fig. 4.

A candidate solution has been generated as a variable size of nodes, and these nodes have been chosen randomly from this pool of nodes instead of randomly chosen from network nodes. For each candidate solution, a Weighted Integration Cascade (WIC) model proposed in section 3 was employed as the fitness function for evaluation. Because of the stochastic nature of the diffusion model, the fitness value is estimated empirically through repetition, by applying the diffusion models 100 times and then taking the mean.

Fig. 4. Flowchart for NSGAI-IM

4. Results and discussion

In this part, the experiments have been applied in two real-world datasets to evaluate the performance of the proposed method. Therefore, the dataset is described at first.

The first dataset is the Facebook wall posts network, which represents a subset of the posts that each user has posted to another user's wall over the period between 14 and 09–2004 and 22-01-2009. It consists of 46,952 nodes and 876,993 edges [36].

The Digg network is the second dataset that represents messages sent among users of the Digg Social news website over the period from 28 to 10–2008 to 13-11-2008. The datum comprises 30,398 nodes with 87,627 edges [37]. Table 3 represents a list of attributes used in the proposed method of implementing in both NSGAI-IM and NSGA-II, in order to obtain the optimal set of seed, which meets the maximum influence propagation with the fewest possible number of seed.

For evaluation purposes, a comparison of the performance of NSGAI-IM has been made with the NSGA-II and three of the following algorithms: high degree (Degree), degree discount (HIGHDEG), and single discount (SDISC), as demonstrated in Fig. 5. It is worth to mention that the previous algorithms have been replicated to measure their performance on the same datasets. Hence, all algorithms, including the proposed ones, have been encoded in Python language and executed on a PC equipped with Intel Core i7-2630 of 2.00 GHz and 4.00 GB Ram.

Fig. 5 shows the application of the algorithms to the Facebook-wall posts dataset in two experiments; the number of generations in the first is 100, whereas the second experiment has 500 generations, as presented in Fig. 5(a) and (b). The X-axis presents the average influence spread, whereas the Y-axis indicates the number of seed nodes in the candidate solution. The candidate solution is represented in green, while the red color represents the Pareto front found by NSGA-II-IM algorithm. The remaining colors represent the four methods.

The NSGA-II-IM in this experiment was seeded with the results of the NSGA-II and three heuristic algorithm (Degree, HIGHDEG, and SDISC). In Fig. 5(a), the evolution with 100 generation was able to outperform three algorithms in some size of seed and not in others. As for Fig. 5(b), the evolution with 500

Table 3

List of attributes.

Attributes	Values
Population size	100
Length of individual	Variable size in range [25,200]
Tournament selection size	2
No. generations	100 and 500

Fig. 5. Experimental results for the benchmark Facebook wall post network.

generation was able to outperform three algorithms in all sizes of seed.

Generally, the behavior of the NSGA-II-IM algorithm in both experiments is superior to the other algorithms including the baseline NSGA-II algorithm. However, Fig. 5(a) records the enhancement in performance of the NSGA-II-IM algorithm in terms of the maximization of the influence when having smaller individuals size. It is noted that the NSGA-II algorithm in Fig. 5(b) score progress in performance not only in terms of the maximization of influence, but also in its superiority to other algorithms such as high degree, degree discount, and single discount. The NSGA-II-IM algorithm scores progress at 100 generations over the NSGA-II algorithm for 500 generations in terms of the maximization of influence rate. It is worth mentioning that the proposed diffusion model (WIC) has been applied with each of the algorithms above to settle all conditions for comparison.

In particular, the NSGA-II algorithm excels the heuristics algorithms when the seed set (individual) is between 35 and 90, so that the black curve intersects with the other curves at other seed sets. In other words, this indicates that the algorithm is not better than heuristics algorithms with all seed sets. However, the NSGA-II algorithm needs 500 generations to completely outperform the heuristic algorithms with all sizes of seed groups. On the other hand, the proposed NSGAII-IM algorithm outperforms all methods with a percentage of 21% in spread of influence through only 100 generations. Properly configuring the first generation leads to faster convergence on the better Pareto fronts (red curve). Additionally, the role of the proposed diffusion model WIC had to be highlighted and compared to the WC model.

Fig. 6 demonstrates the performance of the NSGAII-IM algorithm using both models of diffusion. It is clear that the latter algorithm with the proposed diffusion WIC model which can improve the influence

Fig. 6. Influence spreads of NSGA-II-IM algorithm for wall post-Facebook dataset under two different diffusion.

spread from 10% to 21% with respect to the WC diffusion model.

A benchmark Digg network features a large network. The application of NSGA-II-IM algorithm is displayed in Fig. 7 on the Digg network. It is significant to state that the size of spreading is greater than in the Facebook network, besides the fact that the algorithm outperformed the other algorithms with an influence spread of around 18%.

Concerning the NSGA-II algorithm, it outperforms the other heuristics ones at some sizes of the individuals (often small individuals sizes), with the highest average impact spread of about 2000 affected nodes in the entire network. However, the NSGA-II-IM algorithm outperforms all algorithms to reach the highest an average spread of 5000 nodes. Again, the performance of the NSGA-II-IM algorithm has been evaluated through two different diffusion models, as is

the case with the Digg network. Obviously, the behavior of the NSGA-II-IM algorithm under the proposed diffusion model WIC showed an improved performance from 6% to 18% compared to the WC diffusion model, as can be noticed in Fig. 8.

5. Conclusion

The findings of this work highlight the importance of the method of selecting the initial seed from one side, and the diffusion model from the other side. It can be concluded that the NSGA-II-IM algorithm proved the ability to reliably outperform three heuristics (Degree, SDISC, and HIGHDEG) for the same size of the set of seed, with fewer generations, which means less time than required to reach sufficient solutions by the NSGA-II algorithm. This is mainly due to the carefully selected seeds for using the centrality

Fig. 7. Experimental results for the benchmark Digg network. The NSGA-II-IM in this experiment 100 generation was seeded with the results of the NSGA-II and three heuristic algorithm.

Fig. 8. Influence spreads of NSGA-II-IM algorithm for Digg dataset under two different diffusion.

measures, which in turn have a big role in the performance of the NSGA-II-IM algorithm. An important conclusion can be drawn that guiding the first generation duly leads to a faster convergence on better Pareto fronts. Particularly, the initial population can be assigned pseudo-random by selecting its individuals that have relatively higher centrality measures. However not necessarily the best, such as a degree, closeness, and eigenvector centrality. Considering the results obtained from applying the NSGA-II-IM for two real networks, the maximization of the influence spread is improved by 14%–21% on the existing heuristic algorithms with the Facebook wall post network, while the proposed approach demonstrated the improvement in influence spread from 4% to 18% with the Digg network.

In addition, it is clear that the role of the proposed propagation model (WIC) cannot be ignored, as it is credited with the superiority of the algorithm when fixing the other conditions. Finally, the nature of the network affects the influence propagation rate. The Digg network is denser in nature than the wall post-Facebook network, and thus the interactions between nodes are more in a shorter period of time, which affects the nature of the influence diffusion calculation.

Conflicts of interest

The authors have no conflicts of interest to disclose

References

- [1] L. Berkani, A semantic and social-based collaborative recommendation of friends in social networks, *Software Pract. Ex.* 50 (2020) 1498–1519.
- [2] A. Acquisti, C. Fong, An experiment in hiring discrimination via online social networks, *Manag. Sci.* 66 (2020) 1005–1024.
- [3] S. Debnath, D. Sarkar, P. Jana, Top-k Influential nodes identification based on activity behaviors in egocentric online social networks, *Comput. Intell. Patt. Recog.* 999 (2020) 463–475. Springer.
- [4] S. Tsugawa, K. Kimura, Identifying influencers from sampled social networks, *Phys. A Stat. Mech. its Appl.* 507 (2018) 294–303.
- [5] M.K. Alasadi, H.N. Almamory, Diffusion model based on shared friends-aware independent cascade, *IOP* 1294 (2019), 042006.
- [6] D. Mohapatra, A. Panda, D. Gouda, S.S. Sahu, A combined approach for k-seed selection using modified independent cascade model, *Comput. Intell. Patt. Recog.* 999 (2020) 775–782. Springer.
- [7] B. Wilder, Y. Vorobeychik, Controlling elections through social influence, *Proc. Int. Jt. Conf. Auton. Agents Multiagent Syst. AAMAS.* 1 (2018) 265–273.
- [8] L. Cui, et al., DDSE: a novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks, *J. Netw. Comput. Appl.* 103 (2018) 119–130.
- [9] J. Fan, Y. Wang, K. Tan, Influence maximization on social graphs: a survey, *IEEE Trans. Knowl. Data Eng.* 30 (2018) 1852–1872.
- [10] S. Chen, K. Tan, Online topic-aware influence maximization, *Proc. VLDB Endow.* 8 (2015) 666–677.
- [11] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, W. Jia, Influence analysis in social networks: a survey, *J. Netw. Comput. Appl.* 106 (2017) 17–32.
- [12] P. Krömer, J. Nowaková, Guided genetic algorithm for the influence maximization problem, *Lect. Notes Comput. Sci.* 10392 (2017) 630–641.
- [13] 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*, 2001, pp. 57–66.
- [14] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, *Proc. ACM SIGKDD Int. Conf. Knowl. Disco. Data Minin.* (2003) 137–146.

- [15] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, *Proc. ACM SIGKDD Int. Conf. Knowl. Disco. Data Minin.* (2007) 420–429.
- [16] Z.K. Bao, C. Ma, B.B. Xiang, H.F. Zhang, Identification of influential nodes in complex networks: method from spreading probability viewpoint, *Phys. A Stat. Mech. its Appl.* 468 (2017) 391–397.
- [17] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, *Phys. Rep.* 650 (2016) 63.
- [18] R.K. Behera, S.K. Rath, S. Misra, R. Damaševičius, R. Maskeliūnas, Distributed centrality analysis of social network data using mapreduce, *Algorithms* 12 (8) (2019) 161.
- [19] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, *Proc. ACM SIGKDD Int. Conf. Knowl. Disco. Data Minin.* (2009) 199–208.
- [20] M. Lahiri, M. Cebrian, The genetic algorithm as a general diffusion model for social networks, *Proc. Natl. Conf. Artif. Intell.* 1 (2010) 494–499.
- [21] C.W. Tsai, Y.C. Yang, M.C. Chiang, A genetic newgreedy algorithm for influence maximization in social network, *Proc. IEEE Int. Conf. Syst. Man, Cybern. SMC.* 1 (2015) 2549–2554.
- [22] D. Bucur, G. Iacca, Influence maximization in social networks with genetic algorithms, *Lect. Notes Comput. Sci.* 9597 (2016) 379–392.
- [23] A.R. da Silva, R.F. Rodrigues, V. da Fonseca Vieira, R. Xavier, Influence maximization in network by genetic algorithm on linear threshold model, *Lect. Notes Comput. Sci.* 10960 (2018) 96–109.
- [24] R.C. Gunasekara, K. Mehrotra, C.K. Mohan, Multi-objective optimization to identify key players in large social networks, *Soc. Netw. Anal. Min.* 5 (2015) 21.
- [25] D. Bucur, G. Iacca, A. Marcelli, G. Squillero, A. Tonda, Multi-objective evolutionary algorithms for influence maximization in social networks, *Lect. Notes Comput. Sci.* 10199 (2017) 221–233.
- [26] D. Bucur, G. Iacca, A. Marcelli, G. Squillero, A. Tonda, Improving multi-objective evolutionary influence maximization in social networks, *Proc. Int. Jt. Conf. Appl. Evol. Comput.* 10784 (2018) 117–124.
- [27] 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.
- [28] J. Guo, F. Chen, M. Li, A multi-objective Optimization approach for in fluence maximization in social networks, Springer, Singapore, *Proc. Int. Jt. Conf. Indu. Engin. Manag.* (2019) 706–715.
- [29] B. Saxena, P. Kumar, A node activity and connectivity-based model for influence maximization in social networks, *Social Network Analysis and Mining, Soc. Netw. Anal. Min.* 9 (2019) 40.
- [30] T. Hong, Q. Liu, Seeds selection for spreading in a weighted cascade model, *Physica A* 526 (2019) 120943.
- [31] A. You, M.A.Y. Be, I. In, Identifying highly influential nodes in multilayer networks based on global propagation Identifying highly influential nodes in multilayer networks based on global propagation, *Chaos: Int. J. Nonlinear Sci.* 30 (2020), 061107.
- [32] A. Pourdayaei, H. Mokhlis, H.A. Illias, S.H.A. Kaboli, S. Ahmad, Short-term electricity price forecasting via hybrid backtracking search algorithm and ANFIS approach, *IEEE Access* 7 (2019) 77674–77691.
- [33] M.I. Hlal, V.K. Ramachandaramurthya, S. Padmanaban, H.R. Kaboli, A. Pouryekta, T.A.R.B.T. Abdullah, T. Ab Rashid, NSGA-II and MOPSO based optimization for sizing of hybrid PV/wind/battery energy storage system, *nt, J. Power Electron* 10 (2019) 463–478.
- [34] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Trans. Evol. Comput.* 6 (2002) 182–197.
- [35] A. Malik, A study of genetic algorithm and crossover techniques, *Int. J. Comput. Sci. Mobile Comput.* 8 (2019) 335–344.
- [36] M. Weskida, Evolutionary algorithm for seed selection in social influence process, *IEEE/ACM Int. Conf. Adva. Soci. Net. Anal. Minin. (ASONAM).* (2016) 1189–1196.
- [37] M. Weskida, R. Michalski, Finding influentials in social networks using evolutionary algorithm, *J. Comput. Sci.* 31 (2019) 77–85.