Improving Multi-Objective Evolutionary Influence Maximization in Social Networks

Doina Bucur[1], Giovanni Iacca[2], Andrea Marcelli[3], Giovanni Squillero[3], and Alberto Tonda[4]

[1] EEMCS, University of Twente
Zilverling, 2027, 7500 AE Enschede, The Netherlands
{doina.bucur}@gmail.com
[2] Chair for Integrated Signal Processing Systems
RWTH Aachen University, 52056, Aachen, Germany
{giovanni.iacca}@gmail.com
[3] DAUIN, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino ,Italy
{andrea.marcelli,giovanni.squillero}@polito.it
INRA, UMR 782 GMPA
Avenue Lucien Brétignières, 78850 Thiverval-Grignon, France
{alberto.tonda}@inra.fr

Abstract. In the context of social networks, maximizing influence means contacting the largest possible number of nodes starting from a set of seed nodes, and assuming a model for influence propagation. The real-world applications of influence maximization are of uttermost importance, and range from social studies to marketing campaigns. Building on a previous work on multi-objective evolutionary influence maximization, we propose improvements that not only speed up the optimization process considerably, but also deliver higher-quality results. State-of-the-art heuristics are run for different sizes of the seed sets, and the results are then used to initialize the population of a multi-objective evolutionary algorithm. The proposed approach is tested on three publicly available real-world networks, where we show that the evolutionary algorithm is able to improve upon the solutions found by the heuristics, while also converging faster than an evolutionary algorithm started from scratch.

Keywords: influence maximization, social network, multi-objective evolutionary algorithms, seeding

1 Introduction

Social networks (SNs) are graphs that model a generic society and the internal flows of ideas. Nodes represent the actors, usually individuals, and edges their capability to transmit information. While the high-level structure is quite simple, a sharp definition of the rules that control the passage of information, together with the amount of data willfully provided by users, enable the creation of very precise models. The situation has been plainly acknowledged by scholars for years, and recently reached popular recognition [1,2,3].

In the simplest SN model, a directed edge $a \rightarrow b$ denotes that b is exposed to a and may be influenced by it. The rule used for determining whether the information is actually transmitted is called the *propagation model*. Scholars of social sciences studied a number of *probabilistic propagation models*: the edge $a \rightarrow b$ signifies that there is a given probability p for node a to influence node b. Common models include using fixed probabilities, probabilities inversely proportional to the number of edges directed to b, or other topological features [4].

Given the set of *seeds* nodes, that is, network nodes who broadcast a specific information, and a propagation model, the eventual set of influenced nodes $\mathcal{I}$ may be computed. Indeed, one of the most studied problem in SNs is "influence maximization" and consists in determining the optimal set of seed nodes for maximizing the final influence. Such a problem was initially formulated in [5], and it was later proven NP-hard for most propagation models [4].

As with many other NP-hard problem, evolutionary algorithms (EAs) where used to explore the vast search space of all possible subsets of nodes [6]. More recently, a multi-objective EA (MOEA), was used to maximize the influence $\mathcal{I}$ while concurrently minimizing the size of the seed set [7], providing users the necessary data to trade off between *budget* (the number of nodes that need to be influenced) and *effect* (the final influence over the whole network). This paper progresses on the same research line, showing how significant improvements can be attained by carefully initializing the initial population. The results obtained on three different social network graphs are compared against five state-of-the-art heuristics, and clearly demonstrate the efficacy of the proposed methodology.

The rest of the paper is organized as follows: Section 2 introduce the background and survey the related works; Section 3 details the proposed approach, while Section 4 reports an extensive experimental evaluation; finally, Section 5 concludes the paper.

2 Background and Related Work

In this section we first describe the models available in the literature for simulating the influence propagation, and the general formulation of the influence maximization problem; then, we briefly survey the existing methods for solving this problem, based on either ad hoc heuristics or computational intelligence algorithms.

2.1 Models for influence propagation and problem formulation

As most influence propagation models are stochastic, an approximate estimation of the global influence can be obtained empirically, by simulating the propagation process a given number of times: this approach, however, can be computationally expensive, especially when included in an optimization framework.

Furthermore, as the propagation of a message from a node to another may be modeled as a discrete event, propagation models are also time-discrete. As the receptiveness of users to incoming messages from the network differs, several models have been proposed: the most popular belong to the "Cascade" family [4], which views influence as being transmitted through the network in a tree-like fashion, where the seed nodes are the roots. In this work, we will use in particular the Independent Cascade (IC) model, whose pseudocode is given in Algorithm 1. IC was first studied in the marketing domain, modeling the effects that word-of-mouth communication has upon macro-level marketing [8]. Each newly "activated" node n will succeed in activating each inactive neighbor m with a fixed probability p, which is a global property of the system, equal for all edges $n \rightarrow m$ in G.

Algorithm 1 The **Cascade** family of propagation models. G is the network graph, S the set of "seed" nodes, and $p(n \rightarrow m)$ the probability that information will reach across a graph edge $n \rightarrow m$.

```
procedure CASCADE($G, S, p$)
    $A \leftarrow S$          ▷ $A$: the set of active nodes after the propagation ended
    $B \leftarrow S$          ▷ $B$: the set of nodes activated in the last time slot
    while $B$ not empty do
        $C \leftarrow \emptyset$
        for each $n \in B$ do
            for each direct neighbor $m$ of $n$, where $m \notin A$, do
                with probability $p(n \rightarrow m)$, add $m$ to $C$
            end for
        end for
        $B \leftarrow C$
        $A \leftarrow A \cup B$
    end while
    return the size of $A$
end procedure
```

In the classical problem of influence maximization, the goal is to optimize the seed set $\mathcal{S}$ given a budget $k = |\mathcal{S}|$ so that its eventual influence over the whole network is maximal. The influence of a seed set $\mathbf{I}$ is measured as the size of the set $\mathcal{I}$ of active nodes, obtained by the propagation model. Independently from the influence propagation model used, the problem has been proven to be NP-hard [4], and approximating the optimal solution by a factor better than $1 - \frac{1}{e}$ (roughly 63% approximation) is also NP-hard [4].

2.2 Existing solutions for influence maximization

Several heuristics have been presented to find good solutions to the influence maximization problem. *High degree* (HIGHDEG) is a greedy heuristic that simply adds nodes n to A in order of decreasing out-degree [4]. *Single discount*

(SDISC) is a refinement of HIGHDEG proposed by Chen et al. [9], using the idea that if a node n is already active and also there exists an edge $m \rightarrow n$, then, when considering whether to add node m to A, this edge should not be counted towards the out-degree of m. Other popular techniques include DISTANCE, that greedily adds to the set S select nodes in order of increasing average distance to other nodes in the network, following the intuition that being able to reach other nodes quickly translates into higher influence; *Generalized degree discount* (GDD) [10], a refinement of SDISC, which considers not only the direct neighbours of a node candidate to being a seed, but also nodes one level deeper in the graph; and *Cost-Effective Lazy Forward selection* (CELF), a greedy hill-climbing algorithm [11]. Several metaheuristics and optimization algorithms have also been applied to the problem, ranging from simulated annealing [12] to genetic algorithms [6]. In [7], a Multi-Objective Evolutionary Algorithm (MOEA) [13] was proposed for influence maximization, where the two considered objectives were (i) maximizing the influence of a seed set and (ii) minimizing the number of nodes in the seed set. Intuitively, this produced a Pareto front of candidate solutions, each one a different compromise. While the proposed methodology was shown to outperform both HIGHDEG and SDISC for all values of the budget k (number of seed nodes) on the considered case studies, the main drawback was the computational time required to reach satisfying solutions: millions of individual evaluations were necessary, each one consisting of multiple runs of an influence spread model. This observation provided the motivation for the present work, where we try to reduce the time consumption needed for the MOEA to converge in order to make the method applicable also in contexts with limited computing resources.

3 Proposed approach

To improve upon the work presented in [7] and overcome the aforementioned limitations due to the method time consumption, we introduce here a seeding mechanism. In particular, we show that by seeding the initial population of the MOEA with the results of computationally cheap heuristics, the number of individual evaluations required to reach satisfying solutions drops dramatically.

Another important difference with respect to the MOEA used in [7] concerns the algorithmic implementation: while that work was based a C++ open source customizable evolutionary tool [14], here we use *inspyred*[4], a Python open source framework for creating biologically-inspired computational intelligence algorithms, including evolutionary computation, swarm intelligence, and immunocomputing. *inspyred* provides easy-to-modify canonical versions of several bio-inspired algorithms, among which the MOEA NSGA-II [15], that we use in the experiments presented in this paper.

Individual representation and evolutionary operators were custom-designed for this specific application: for the problem at hand, a candidate solution is a

[4] `http://pythonhosted.org/inspyred/`

set of nodes of variable size, consisting of a subset of the set of nodes in the original network. Individuals are thus unordered sequences of unique integer node identifiers, representing the seeds of influence in the network.

As for the evolutionary operators, we used three problem-specific mutations (add, remove or replace one node in a set), and one crossover operator with a check that removes inconsistencies from the resulting individuals, ensuring that a specific node appears only once in each individual. The operators are always applied with uniform probability, while the parent individuals are selected through a tournament selection of size 2.

Finally, the fitness value of a candidate solution is a probabilistic metric of the number of nodes that are likely to be reached, starting from a given set of seeds of influence — according to the IC model of influence propagation (described in Section 2). Given the stochastic nature of both propagation models, the fitness estimation is empirical, and itself a stochastic process: repeated simulations of the network propagation model yield an extent to which the network is reached, and the final fitness value is the average of these fitness samples.

4 Experimental evaluation

4.1 Benchmarks

In order to assess the proposed improvements for influence maximization, we selected three case studies among social network graphs available in the Network Repository[5] and SNAP[6] databases. Two of the selected social networks, **ego-Facebook** and **ca-GrQc** were also considered in [7], while **soc-ePinions1** was not considered in the previous study. The selected benchmarks, with their respective features, are reported in Table 1.

Table 1. Main features of the case studies considered for the experimental evaluation of the proposed MOEA approach.

Name	**ego-Facebook**	**ca-GrQc**	**soc-ePinions1**
Nodes	4,039	5,242	75,879
Edges	88,234	14,496	508,837
Type of graph	undirected	undirected	directed
Nodes in largest WCC	4,039	4,158	75,877
Nodes in largest SCC	4,039	4,158	32,223
Average clustering coefficient	0.6055	0.5296	0.1378
Diameter	8	17	14

[5] `http://networkrepository.com/`

[6] `https://snap.stanford.edu/index.html`

4.2 Experimental results

In all the experiments, we consider the IC propagation model. The MOEA configuration used here is: $\mu = 2000$, $\lambda = 2000$, tournament selection of size $\tau = 2$, influence propagation model IC with $p = 0.05$, stop condition 500 generations.

Figures 1, 2 and 3 show the results for the social networks **ego-Facebook**, **ca-GrQc** and **soc-ePinions1**, respectively. Considering the first two networks, we observed that with respect to [7] the improved algorithm is able to find solutions that outperform the heuristic already during the first few generations, with less than 10,000 evaluations (compared to the almost 1,000,000 that were necessary to outperform both HIGHDEG and SDISC in the previous study). A similar trend was observed also on the third network, **soc-ePinions1**, for which we could not even run two of the heuristics, CELF and DISTANCE, due to their excessive time complexity (approximately one data point in ten hours).

In summary, while the seeding procedure requires running the heuristic once, the computational cost is roughly equivalent to just one generation of the MOEA. Furthermore, the proposed methodology is able to outperform even more refined heuristics which were not considered in [7], such as CELF. Finally, the new configuration also solved previously experienced issues with populating the higher part of the Pareto front (see [7] for more details).

5 Conclusions

In this paper, we introduced an improvement over a previously proposed multi-objective evolutionary approach for influence maximization in social networks.

A MOEA is tasked with finding the set of k seed nodes that, given a model of influence propagation, maximize the nodes reached in the network. As minimizing the value of k is also given as an optimization objective, the MOEA is able to find a Pareto front of compromises between number of seed nodes in the set and global influence in the graph. While the main weak point of the previously proposed approach was the time required to reach good solutions, in this paper we show how initializing the first generation properly leads to faster convergence on better Pareto fronts. The approach has been tested on three real-world social networks, and proved to be able to overcome also the state-of-the-art heuristics.

In future works, we aim to progress on this research line by combining the seeding mechanism proposed here with a surrogate model approach, in order to speed up even further the computations and make the method applicable also on larger networks.

Acknowledgments

This article is based upon work from COST Action CA15140 'Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice (ImAppNIO)' supported by the COST Agency.

References

1. Hersh, E.D.: Hacking the electorate: How campaigns perceive voters. Cambridge University Press (2015)
2. Kreiss, D.: Prototype politics: Technology-intensive campaigning and the data of democracy. Oxford University Press (2016)
3. Grassegger, H., Krogerus, M.: The data that turned the world upside down. Luettavissa: http://motherboard. vice. com/read/big-data-cambridge-analytica-brexit-trump. Luettu 28, 2017 (2017)
4. Kempe, D., Kleinberg, J., Éva Tardos: Maximizing the spread of influence through a social network. Theory of Computing 11(4), 105–147 (2015)
5. Richardson, M., Agrawal, R., Domingos, P.: Trust management for the semantic web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) The Semantic Web - ISWC 2003, Lecture Notes in Computer Science, vol. 2870, pp. 351–368. Springer (2003)
6. Bucur, D., Iacca, G.: Influence maximization in social networks with genetic algorithms. In: European Conference on the Applications of Evolutionary Computation. pp. 379–392. Springer (2016)
7. Bucur, D., Iacca, G., Marcelli, A., Squillero, G., Tonda, A.: Multi-objective evolutionary algorithms for influence maximization in social networks. In: European Conference on the Applications of Evolutionary Computation. pp. 221–233. Springer (2017)
8. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12(3), 211–223 (2001)
9. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 199–208. KDD '09, ACM, New York, NY, USA (2009)
10. Wang, X., Zhang, X., Zhao, C., Yi, D.: Maximizing the spread of influence via generalized degree discount. In: PloS one (2016)
11. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 420–429 (August 2007)
12. Jiang, Q., Song, G., Cong, G., Wang, Y., Si, W., Xie, K.: Simulated annealing based influence maximization in social networks. In: Burgard, W., Roth, D. (eds.) AAAI. AAAI Press (2011)
13. Coello, C.A.C., Van Veldhuizen, D.A., Lamont, G.B.: Evolutionary algorithms for solving multi-objective problems, vol. 242. Springer (2002)
14. Squillero, G.: MicroGP - an evolutionary assembly program generator. Genetic Programming and Evolvable Machines 6(3), 247–263 (2005)
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation 6(2), 182–197 (2002)

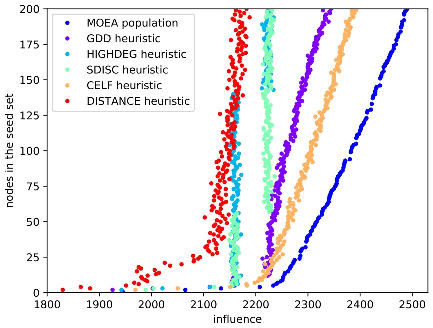

Fig. 1. Experimental results for the benchmark graph ego-Facebook. The MOEA in this experiment was seeded with the results of the GDD heuristic, and the evolution was able to outperform even the effective CELF.

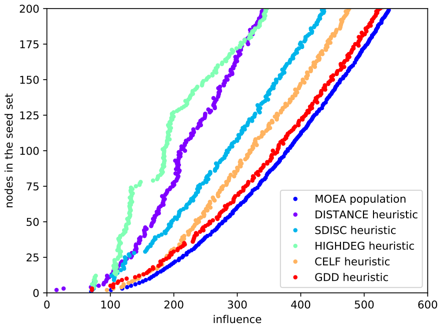

Fig. 2. Experimental results for the benchmark graph ca-GrQc. In this case, GDD was already the most performing heuristic of the group, but the MOEA seeded with the initial results was able to eventually find improvements over the initial approximation.

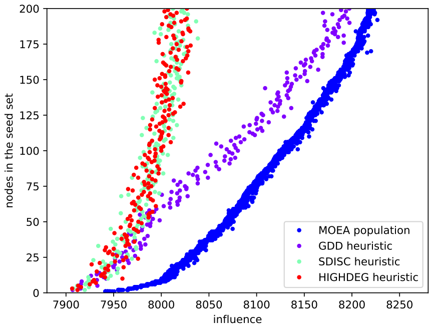

Fig. 3. Experimental results for the benchmark graph soc-ePinions1. The MOEA in this experiment was seeded with the results of the GDD heuristic. DISTANCE and CELF could not be run on this network due to their excessive time complexity.