Social Network Analysis and Mining (2021) 11:55
https://doi.org/10.1007/s13278-021-00765-9

ORIGINAL ARTICLE

The generalized influence blocking maximization problem

Fernando C. Erd[1] · André L. Vignatti[1] · Murilo V. G. da Silva[1]

Received: 13 February 2021 / Revised: 27 May 2021 / Accepted: 29 May 2021 / Published online: 13 June 2021

Abstract

Given a network N and a set of nodes that are the starting point for the spread of misinformation across N and an integer k, in the influence blocking maximization problem the goal is to find k nodes in N as the starting point for a competing information (say, a correct information) across N such that the reach of the misinformation is minimized. In this paper, we deal with a generalized version of this problem that corresponds to a more realistic scenario, where different nodes have different costs and the counter strategy has a "budget" for picking nodes for a solution. Our experimental results show that the success of a given strategy varies substantially depending on the cost function in the model. In particular, we investigate the cost function implicitly used in all previous works in the field (i.e., all nodes have cost 1), and a cost function that assigns higher costs to higher-degree nodes. We show that, even though strategies that perform well in these two diverse cases are very different from each other, both correlate well with simple (but different) strategies: greedily choose high-degree nodes and choose nodes uniformly at random. Furthermore, we show properties and approximations results for the influence function in several diffusion models .

Keywords Influence blocking maximization · Misinformation · Complex networks

1 Introduction

The spread of misinformation is not a new phenomenon; however, with the prevalence of social media this problem seems to have being gaining more momentum (Lazer et al. 2018). There is evidence that people tend to believe in information that matches their perception of social narratives and to discredit narratives that deconstruct that perception (Lewandowsky et al. 2012). In this way, social media, due to its structure and ways of disseminating information, could expand the circulation of misinformation. Vosoughi et al. (2018) show that in Twitter there is a 70% greater chance that fake news will be shared rather than real one.

Recently, several studies showed how the spread of misinformation has potential to influence the behavior of the society. Allcott and Gentzkow (2017) present an analysis of how misinformation may have affected the result of the 2016 United States presidential elections. Another example is the number of questionable sources on the main social platforms regarding the outbreak of COVID-19, as shown by Cinelli et al. (2020).

The algorithmic aspects of a problem originally from the field of "viral marketing" were investigated by Kempe et al. (2003). The proposed computational problem, known as influence maximization in networks, is the following. Given a network where a node corresponds to a person and an edge corresponds to the connection between two people, the goal is to select the best individuals to advertise a product, such that the information about that product reaches the largest number of people. From this problem, a line of research arose addressing the problem of finding a counter strategy for the spread of such influence (He et al. 2012). In our paper, we assume that we are dealing with the spread of misinformation and the counter strategy seeks to spread the correct information (or a counter narrative). This computational problem, called influence blocking maximization, is the following. Given a set of nodes as starting point for the spread of misinformation across the network and an integer k, the goal is to find k nodes for the spread of the correct

✉ André L. Vignatti
vignatti@inf.ufpr.br

Fernando C. Erd
fcerd@inf.ufpr.br

Murilo V. G. da Silva
murilo@inf.ufpr.br

[1] Federal University of Paraná, Curitiba, Brazil

information across the network so the reach of the misinformation is minimized .

In the previous work in this field (He et al. 2012; Arazkhani et al. 2019b, a; Wu and Pan 2017) (we discuss these works in detail in Section II), given k, the counter strategy is able to pick any set of nodes of size k for blocking the misinformation. We note that this scenario might be unrealistic, since choosing a node with very high degree might be much more expensive than a node of degree one, for example. In all previous works using models based on the independent cascade, the proposed strategies for choosing the set of k nodes for the counter strategy perform only marginally better than choosing nodes of high degree. (The algorithms are about 1% more effective than picking high-degree nodes.) So, in our work we generalize the problem so that different nodes might have different costs and the counter strategy has a "budget" k for finding a set of nodes such that the total cost of the nodes in the set stays within that budget. Note that the previous works in the literature fall into the particular case of our problem where the cost function assigns cost 1 to every node in the network.

In our paper, we investigate counter strategies in this generalized scenario using two distinct cost functions. Our experimental results show that the success of a given strategy varies substantially depending on the cost function in the model. The counter strategies used in our experiments are four node properties well known in the literature: betweenness centrality, percolation centrality, PageRank, and clustering coefficient. The two different cost functions that we compare are the uniform cost function, which is the cost function implicitly used in all previous works in the field, and the degree penalty cost function, which assigns a higher cost to a higher-degree node. This second cost function may be a more realistic since nodes of high degree in a network might be more expensive. In consonance with previous works, we show that for the uniform cost function the winning strategies (betweenness, percolation, and PageRank) correlate well with simply choosing high-degree nodes. In the degree penalty cost function, we show that the winning strategy (choosing nodes with high clustering coefficient) does not correlate at all with the previous strategy. Interestingly, we note that there is also a simple strategy in this scenario: picking nodes uniformly at random the solution. This article is an extended version of the Erd et al. (2020). In addition to the previous version, we explore properties regarding the influence function that ensures approximations for the most studied dissemination models, and we added experiments with different spreading probabilities.

The rest of this article is organized as follows: A brief review of recent studies is provided in Sect. 2. Section 3 presents the MCICM information dissemination model. The problem definition is described in Sect. 4. In Sect. 5, we show properties and approximations results for the proposed problem. The methodology used for our results is discussed in Sect. 6. Experimental results on some well-known datasets are reported in Sect. 7. Finally, Sect. 8 concludes the work.

2 Related work

The influence maximization problem in a network is the computational problem of finding a set of nodes of size k, for a given integer k that is part of the input, as the starting point for the spread of information in this network so that the maximum number of nodes is reached. There are two models widely used to simulate the dissemination of information in the network, namely the independent cascade (IC) model and the linear threshold (LT) model, both proposed by Kempe et al. (2003).

In this paper, we deal with a version of the influence maximization problem where there are two types of competing information being disseminated in the network, referred here as the misinformation and the correct information. The competitive version of the influence maximization problem is formally proposed by He et al. (2012). In this scenario, the input consists of a network with k specified nodes for the spread of the misinformation and the goal is to find k nodes for the spread the correct information so that number of nodes reached by the misinformation is minimized. Using a variation of the linear threshold model, called competitive linear threshold model (CLT), the authors show that the problem is submodular and monotonic, which guarantees an approximation of $1 - 1/e$ of the optimal solution using a hill climbing strategy. Also, they propose the CLDAG algorithm, based on the LDAG (Chen et al. 2010) algorithm which was previously used for the influence maximization problem.

The first work analyzing the problem with the independent cascade model in the competitive version is from Budak et al. (2011) who proposed the eventual influence limitation (EIL) problem, where the cascade of negative (false) information propagates alone in the network for a given number of steps, and only after that the cascade of positive information begins to spread through the network. Budak's main contribution is a proof that the function is submodular and monotonic over the campaign-oblivious independent cascade model (COICM). In addition, Budak showed that using the multi-campaign independent cascade model (MCICM) when the probabilities of positive and negative dissemination are arbitrary, they do not possess submodularity property, but when the probability of positive dissemination is 1 for all edges, the model can guarantee an approximation to the optimal solution.

In the MCICM, Arazkhani et al. (2019b) used a metric based on some centrality measures, such as degree, betweenness, and closeness, in order to choose the set of positive seed nodes. In a later study, Arazkhani et al. (2019a) combined the centralities in a pre-processing method to find the largest k communities using fuzzy clustering, which chooses a node with the highest degree, betweenness, or closeness of each community as being the positive seed. Regarding the dissemination taking place on the COICM, Wu and Pan (2017) used the structure of maximum influence arborescence (MIA) proposing two heuristics, CMIA-H and CMIA-O. In the same work, they consider the MCICM in the particular case where the probability of positive dissemination is 1 for all edges.

A variant of the influence maximization problem proposed by Kempe et al. (2003) considers costs for selecting each node in the network. For example, the budgeted influence maximization problem studied by Nguyen and Zheng (2013), each node v is associated with an arbitrary cost $c(v)$. The goal of such problem is to select a set S of nodes so that the cost of those nodes in S is at most a budget b, and S maximizes the spread of information though the network. In the budgeted competitive influence maximization problem proposed by Pham et al. (2019), the goal is to maximize the spread of one product over another, given a budget. Finally, the work (Leskovec et al. 2007) proposes the outbreak detection problem under the context of water distribution. The contamination starts at some point, and from the moment the contamination passes through a sensor, an alarm is triggered. The goal is to select the best sensor placement for monitoring the quality that respects the budget.

3 Diffusion model

In this work, we use the multi-campaign independent cascade model (MCICM) introduced by Budak et al. (2011), which is similar to the independent cascade model proposed by Kempe et al. (2003). In MCICM, given a directed or undirected graph $G = (V, E)$ there are two spreading cascades P and N representing the positive and negative cascades, respectively, two initial sets $S \subseteq V$ and $N_0 \subseteq V$ of positive and negative seeds, respectively. The negative seeds are the starting point for the misinformation, and the positive seeds are the starting point for the correct information. Each node assumes three different states: positive, negative, or inactive, and in the starting configuration, the nodes in S are set as positive, those in N_0 are set as negative and the rest of the nodes are set as inactive. In addition, each edge $(u, v) \in E$ has two weights, $w^+_{u,v}$ and $w^-_{u,v}$ in the range [0, 1], which denote the probabilities of u activating, respectively, positively or negatively the node v. The simulation occurs in discrete time steps, and if u is activated in step t by the cascade of P or N, it has only one chance to positively or negatively activate a neighbor v during the simulation. As a tiebreaker rule, if the P cascade and the N cascade in the same step t try to activate the same inactive node, the N cascade has preference for the activation. The step t finishes when all nodes activated during step $t - 1$ try to activate their inactive neighbors, and simulation ends in step t when no node is activated by the cascades.

Another model is the campaign-oblivious independent cascade model (COICM) also used by Budak et al. In such model, each edge has a single probability value, meaning that both positive and negative information has the same propagation probabilities, and the rest of the model is similar to MCICM. Although the MCICM is the main model dealt with in this article, some of our results in Sect. 5 refer to the COICM.

4 Problem definition

Let N_T be the set of negative nodes that is the outcome of an execution of the stochastic process of diffusion (in our case, dictated by the MCICM). The outcome N_T depends on the graph G, the probabilities w^+ and w^-, and the initial negative and positive seed sets N_0 and S. Thus, given an integer k, the probability that $|N_\mathrm{T}| = k$ depends on the same input variables, but in the notation we only make it explicit the dependence on the initial positive seed set S, writing $\Pr\big(|N_\mathrm{T}| = k \,\big|\, S\big)$.

Now, given the initial positive seed set S, the expected size of the negative nodes N_T is,

$$\mathbb{E}\big[|N_\mathrm{T}| \;\big|\; S\big] = \sum_{k=0}^{|V|} k \cdot \Pr\big(|N_\mathrm{T}| = k \,\big|\, S\big).$$

We can measure the impact of an initial positive seed set S by considering the difference between two scenarios, when the initial positive seed set is S, and when the initial positive seed set is empty. This is called the expected blocked negative influence of S and is formally defined as

$$\sigma(S) = \mathbb{E}\big[|N_\mathrm{T}| \;\big|\; \{\emptyset\}\big] - \mathbb{E}\big[|N_\mathrm{T}| \;\big|\; S\big],$$

and we want to maximize this quantity. We can now define the problem.

Problem 1 (Generalized Influence Blocking Maximization (GIBM)) Given a graph $G(V, E)$ with costs $c(v)$ for each $v \in V$, propagation probabilities w^+ and w^-, a negative seed set N_0, and a positive integer k, the GIBM problem aims to find the positive seed set S that maximizes $\sigma(S)$ where $\sum_{v \in S} c(v) \leq k$.

We note that in a real scenario, receiving the correct input parameters—e.g., detecting the negative seeds N_0, obtaining the propagation probabilities w^- and w_+, and defining the costs $c(v)$—is a problem on its own. Therefore, since our main focus in this work is the investigation of the algorithmic aspects of misinformation diffusion and blocking in networks, we assume that the algorithms in this paper receive the correct input.

4.1 Particular cases

In this work, we are interested in two particular cases for the cost function. The first is the case, which we call the uniform cost function, all nodes have the same cost, w.l.o.g., say, $c(v) = 1$ for each $v \in V$. This cost function has been used in previous works (He et al. 2012; Arazkhani et al. 2019b, a; Wu and Pan 2017) for the particular problem called influence blocking maximization. Let $\delta(v)$ be the degree of a node v. The second cost function investigated in our work $c(v) = \delta(v)$ for each $v \in V$, which we call the degree penalty cost function.

5 Properties and approximations

According to Nguyen and Zheng (2013), the budgeted influence maximization problem can be understood as the budget version of the problem influence maximization proposed by Kempe et al. (2003). The problem has as input a directed graph $G(V, E)$, a cost function $C : V(G) \rightarrow \mathbb{Z}^+$, and a budget $B \in \mathbb{Z}^+$. The cost function C assigns a non-uniform selection cost to each node of the network, which is the cost to be paid to select a given node for starting the spreading in the graph. The goal is to select a set of nodes within the budget that maximizes the spread of influence over the network. More formally, the goal is to select a set S where $\sum_{u\in S} C(u) \leq B$ such that, for any set S' where $\sum_{u\in S'} C(u) \leq B$,
$\sigma(S) \geq \sigma(S')$.

Nguyen and Zheng proposed a greedy algorithm where the selection criteria are the largest cost–benefit ratio of a node, whereas other studies propose greedy algorithms where a node is chosen if it maximizes the reach without considering the cost. However, when considering only the cost–benefit criteria, the algorithm has an unbounded approximation factor. Thus, the authors proposed a new algorithm, which chooses the maximum between two cases: the approach that considers the cost and the one that considers only the reach and ignores the cost. Such algorithm has a $(1 - 1/\sqrt{e})$ approximation factor for the budgeted influence maximization problem. This approximation result can be extended to the GIBM problem. The idea is that, given an input graph G for the GIBM problem, we build a graph G'' based on G that serves as the input for the budgeted influence maximization problem, such that the solution in G'' can be used for G in GIBM problem. Such result is presented in Theorem 1.

Theorem 1 *The GIBM problem has a* $(1 - 1/\sqrt{e})$*-approximation in the COICM.*

Proof Our approach is similar to the one by Budak et al. (2011) for the eventual influence limitation problem. The spreading process that occurs on an edge can be seen as a coin toss, where the positive and negative spread on an edge (u, v) occurs as coin tosses with probability of success, respectively, $w^+_{u,v}$ and $w^-_{u,v}$. As pointed out by Kempe et al. (2003) and Budak et al. (2011), it does not matter if the coins are tossed at the exact moment when a node u tries to activate its neighbor v, or if the coins are pre-tossed and their results are stored only to be used when u tries to activate v. We use the idea that all coins were previously tossed to create the graph G' below.

Let G be the input graph for the GIBM problem (see, e.g., Fig. 1). The first step is to create a graph $G'(V, E')$ where E' is the set of activated edges, previously determined by pre-tossed coins. The graph G' has the paths of both positive and negative dissemination, since the edge activation in the COICM indicates that it can send positive or negative information, whichever comes first. Given N_0, the set of nodes that are reachable from N_0 by the activation process is referred to as N'. Note that, in G', as we pre-tossed the coins, it is easy to the identify the nodes belonging to N' (see, e.g., Fig. 1).

Next, we create a graph G'' that represents where the positive spread arrive before the negative one. Let

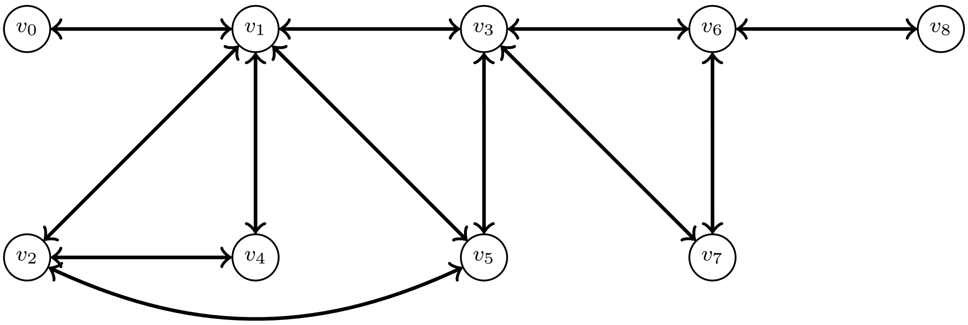


Fig. 1 An example of the input graph G for the GIBM problem, where $N_0 = \{v_0, v_5\}$. For simplicity, we suppose in this example that all edges become active by the pre-tossing activation process. So, the graph G' is equal to G and the set N' is all V except for those belonging to the set N_0

Table 1 The distances between the nodes from N'' to N', based on the example of Fig. 1

	N'						
N''	v_1	v_2	v_3	v_4	v_6	v_7	v_8
v_1	X	1	1	1	2	2	3
v_2	1	X	2	1	3	3	4
v_3	1	2	X	2	1	1	2
v_4	1	1	2	X	3	3	4
v_6	2	3	1	3	X	1	1
v_7	2	3	1	3	1	X	2
v_8	3	4	2	4	1	2	X

Table 2 The distances between nodes from N_0 to N', based on the example of Fig. 1. In this way, $|P(N_0, v)|$ is the minimum value of each column

N'							
N_0	v_1	v_2	v_3	v_4	v_6	v_7	v_8
v_0	1	2	2	2	3	3	4
v_5	1	1	1	2	2	2	3
$\|P(N_0, v)\|$	1	1	1	2	2	2	3

Table 3 This table shows the added edges in G'', based on the values of Tables 1 and 2 . For example, for node v_4, the shortest path from N_0 to it is 2 (i.e., the value $|P(N_0, v_4)|$, recovered from Table 2), such that all the shortest paths $P(u, v)$ with distance less than 2 (such distances are recovered from Table 1) are added in E'' —in this case, $v_1 \rightarrow v_4$ and $v_2 \rightarrow v_4$. This means that, when choosing v_1 or v_2 as positive seeders, the positive spread arrive before the negative in v_4

$v \in N'$	$\|P(N_0, v)\|$	Added paths in G'' (if $\|P(u, v)\| < \|P(N_0, v)\|, \forall u \in N''$)
v_1	1	X
v_2	1	X
v_3	1	X
v_4	2	$v_1 \rightarrow v_4, v_2 \rightarrow v_4$
v_6	2	$v_3 \rightarrow v_6, v_7 \rightarrow v_6, v_8 \rightarrow v_6$
v_7	2	$v_3 \rightarrow v_7, v_6 \rightarrow v_7$
v_8	3	$v_3 \rightarrow v_8, v_6 \rightarrow v_8, v_7 \rightarrow v_8$

$P(u, v)$ be the set of edges of the shortest paths from u to v in G'. Let $P(N_0, v)$ be the set of edges of the shortest paths from the closest node in N_0 to v in G'. Formally, $P(N_0, v) = \{P(u, v) : u = \arg\min_{u' \in N_0} |P(u', v)|\}$. Then, $G''(N'', E'')$ is defined as

$$N'' = V \setminus N_0$$
$$E'' = \left\{P(u, v) \ : \ |P(u, v)| < |P(N_0, v)| \quad \forall\, v \in N', u \in N''\right\}.$$

Intuitively, G'' adds the shortest paths from u to v in G' if such paths arrives before a node from N_0 to v. (See Tables 1, 2, and 3 for a step-by-step construction of E'' based on the example of Fig. 1. Also, Fig. 2 shows the final graph G'' for this example.) The idea is that G'' is the graph where the positive spread arrives before the negative.

Now, it is clear that solving the GIBM problem is equivalent to maximizing the number of nodes reachable from of an initial set S in G'', but the latter is precisely the budgeted influence maximization problem. □

A well-known result (Cornuejols et al. (1977), G. L. Nemhauser et al. (1978)) says that if the influence function is *submodular* and *monotonic* (see Definitions 1 and 2), then a general greedy procedure leads to a $(1 - \frac{1}{e})$ -approximation guarantee. Theorem 2 shows that it is not possible to use such greedy procedure on GIBM problem based on the MCICM because the submodularity property does not hold for this case.

Definition 1 Let S, T and U be sets, such that $S \subseteq T \subseteq U$ and $f : 2^U \rightarrow \mathbb{R}^+$. We say that f is *submodular* if $f(S \cup \{w\}) - f(S) \geq f(T \cup \{w\}) - f(T)$ for all $w \in U \setminus T$.

Definition 2 Let S and T be sets, such that $S \subseteq T$ and $f : 2^T \rightarrow \mathbb{R}^+$. We say that f is monotonic if $f(S) \leq f(T)$.

Theorem 2 *The influence function in GIBM problem is not submodular on the MCICM.*

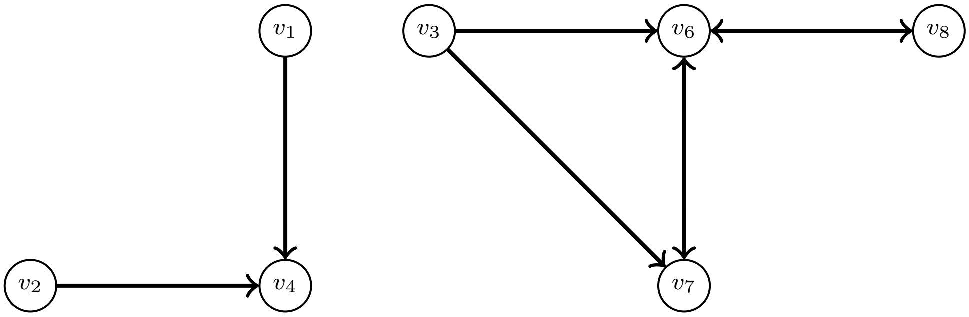


Fig. 2 G'' Graph, note that in this graph we want to maximize the reach, as it is the graph where the positive spread arrives before negative spread. So we want to choose the best set that is within a budget that increases the reach in that graph and that is the problem of budgeted influence maximization

Proof Our proof is similar to the one by Budak et al. (2011), but assuming the case where the negative spreading has preference over the positive.

We present an input where the submodularity property does not hold. Figure 3 shows the graph G used as the input graph for the problem, where $N_0 = \{v_1\}$. The first step is to create a graph $G^P(V, E^P)$ where E^P are the previously activated edges that send the positive information before the simulation starts (defined by the pre-tossed coins idea, as discussed before). Suppose G^P is the graph of Fig. 4, i.e., the only activated edges are (v_3, v_4) and (v_9, v_{10}). In a similar way, a graph $G^N(V, E^N)$ is created that contains only the previously activated edges that send the negative information; such graph is shown in Fig. 5.

In G^P, we see that v_3 and v_{10} are the only nodes available to add to the solution set S. If the problem input budget allows adding v_3 to S, then $f(\{v_3\}) = 1$, as the only saved node is v_4. Similarly, when $S = \{v_{10}\}$, $f(\{v_{10}\}) = 1$, as the only saved node is v_9. However, if the budget allows the choice of both nodes, i.e., $S = \{v_3, v_{10}\}$, then $f(\{v_3, v_{10}\}) = 3$, because $\{v_4, v_8, v_9\}$ are saved. Let $S = \emptyset$, $T = \{v_3\}$ and $v = v_{10}$, then the submodularity property holds if

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$
$$f(\emptyset \cup \{v_{10}\}) - f(\emptyset) \geq f(\{v_3\} \cup \{v_{10}\}) - f(\{v_3\})$$
$$1 - 0 \geq 3 - 1$$
$$1 \geq 2$$

but this contradicts the submodularity property. □

In addition, it is possible to extend the proof of He et al. (2012) to the GIBM problem, when treated on the competitive linear threshold (CLT) model. In this case, the influence blocking maximization problem is the uniform case of the GIBM problem we address here. Thus, we claim that the GIBM problem is also submodular, as long as the budget allows to choose the k nodes that maximizes the reduction of the negative spread, since in the proof of the submodularity

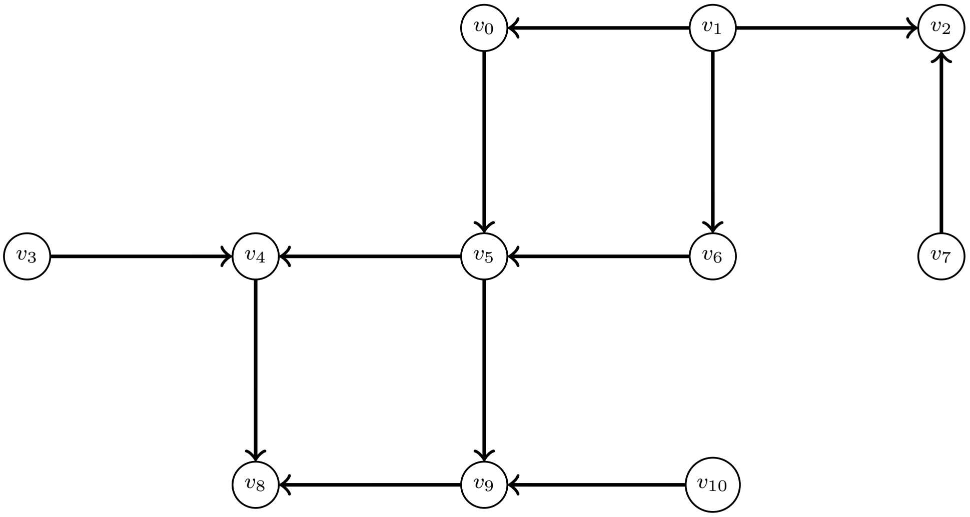


Fig. 3 The graph G used in the proof of Theorem 2, where $N_0 = \{v_1\}$

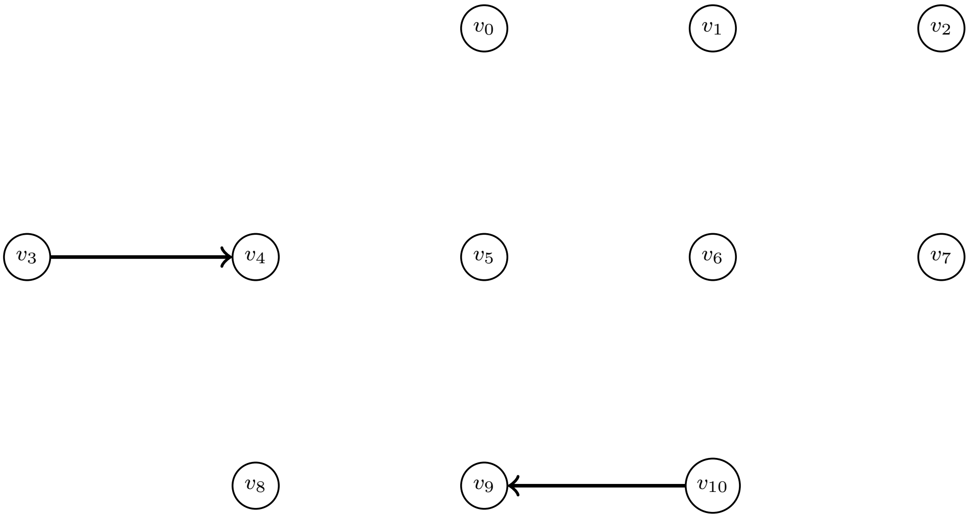


Fig. 4 Graph G^P with the pre-activated edges that send positive information

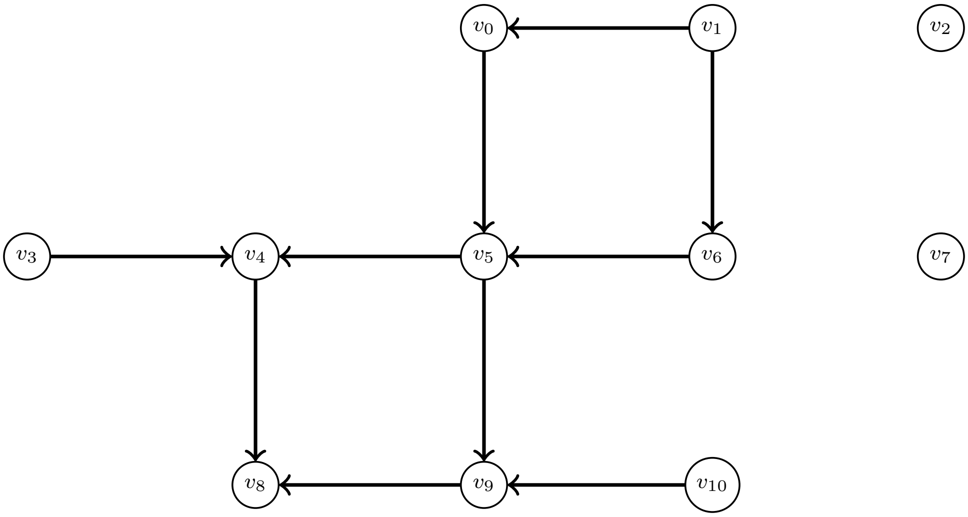


Fig. 5 Graph G^N with pre-activated edges that send negative information. Note that $N_0 = \{v_1\}$ reaches a large number of the nodes

the budget does not matter. Such facts lead to the statement of Theorem 3.

Theorem 3 *The GIBM problem has a* $(1 - \frac{1}{e})$*-approximation in the CLT model, if the budget allows selecting the nodes that optimize the dissemination.*

The proof of Theorem 3 is omitted as it is identical to the analogous statement proved in He et al. (2012).

6 Methodology

Our goal is to investigate several measures of centrality to be used as strategies to solve the problem under the uniform and degree penalty cost functions cases. In the former case, we mimic the (somewhat unrealistic) case that has already been considered in previous works, while in the latter we seek a more realistic scenario where the cost of a node is proportional to its degree. For both cost functions cases, we perform simulations on directed and undirected graphs, with the goal of analyzing the behavior between the two types of graphs.

If the input graph is not connected, we take into consideration only the largest connected component (resp. largest weakly connected component for directed graphs) of the graph. This is a common practice in the field since both the correct information and the misinformation cannot "jump" from one component to another. In the simulations, we analyzed three different scenarios for the probability of an information/misinformation being propagated. More precisely, we run experiments for probabilities w^+ and w^- as follows:

low spread: w^+ and w^- are chosen in the interval [0, 0.2]
normal spread: w^+ and w^- are chosen in the interval [0, 1].
high spread: w^+ and w^- are chosen in the interval [0.75, 1].

In all three cases, the values w^+ and w^- are chosen independently and randomly from the uniform distribution. The nodes of the negative seed set N_0 are positioned uniformly at random in the graph. Various sizes of N_0 are considered in the experiments (more details in Sect. 7).

For the experiments, we choose three real-world datasets, among which are two networks of citations (DBLP and CORA) and the Wikipedia Election dataset. The DBLP citation network (Ley 2012) is a dataset of scientific publications, such as papers and books, where a node represents a publication and an edge represent a citation, that is, there is an edge from A to B if paper A cites publication B. The original DBLP database has 12,590 nodes and 49,749 edges. The Cora dataset (Šubelj and Bajec 2013) contains more than 23,000 nodes and approximately 90,000 edges. Similar to DBLP, this dataset represents citations in articles from a platform of scientific articles, where the nodes are articles and the edges are citations between them. The Wikipedia Elections dataset (Leskovec et al. 2010) represents the English Wikipedia social network, of users who voted for and against each other in admin elections, nodes represent users and an edge represents a user who voted for another. All datasets originally describe directed graphs. The same datasets were used in the experiments on undirected graphs, but the direction of the edges was ignored. The choice to ignore the direction of the edges instead of using others originally undirected datasets allows a more straightforward comparison between the directed and non-directed cases. Table 4 shows the comparison of the three datasets after the preprocessing to find the giant component.

For each proposed scenario, we use the following network metrics as a counter strategy:

Clustering coefficient (Fagiolo 2007),
PageRank (Page et al. 1999),
Betweenness (Brandes 2004),
Percolation (Piraveenan et al. 2013).

In addition to measures above, we use two strategies for experiment control: choosing high-degree nodes first (greedily) and choosing nodes at random.

The percolation centrality requires weights for nodes reflecting a certain degree of "contamination," so we use this measure in our experiments in the following way. Let $d(v, N_0)$ be the distance from v to the nearest node in N_0. Thus, the percolation weight for a node v is defined as

$$\text{perc}(v) = \frac{1}{d(v, N_0) + 1}.$$

The idea is that the nodes initially in N_0 are 100% percolated (in this case, $d(v, N_0) = 0$), and as a node is further away from N_0, its percolation weight decreases.

The experiments were launched in an Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and 8 GB RAM. The scripts were implemented in Python 3.6.9 language. For graph manipulations, we use the NetworkX 2.3 library (Hagberg et al. 2008). The implementation of all the networks measures considered in this work is available in NetworkX as well.

7 Experiments and results

In this section, we evaluate the performance of different strategies for finding a solution for the GIBM problem. Since MCICM is a probabilistic model, we run repeated experiments for the spreading over the initial sets N_0 and S in order to obtain the average behavior. In each different scenario, we perform the simulation 1000 times to obtain the average of the positively and negatively contaminated sets.

Initially, we compare four different scenarios, as shown in Table 5. We remember that for each scenario there are three cases for the spreading probability.

Table 4 Dataset statistics.

Network	Giant component nodes	Giant component edges
CORA	23,166	89,157
DBLP	12,495	49,578
Wikipedia Election	7,066	100,721

7.1 Uniform cost function

In this section, we show and analyze the results obtained for the uniform cost function. For these experiments, we set the size of N_0 to be 1% of the number of nodes of each dataset, and we vary the parameter k (here the size of the output set S for positive seeds equals k) between 0.1%, 0.5%, 1%, 1.5%, and 2.0% of the number of nodes in each dataset. We analyze the undirected and directed cases. In each plot, we show the average results for the three datasets. The vertical axis we show the percentage of negatively contaminated nodes, therefore, the lower the values, the better the network metric works as a strategy for the problem. The absolute number of nodes as a function of the percentage is given in Table 6.

In Figs. 6, 7, and 8, we present the results for the uniform cost function on undirected graphs with, respectively, low, normal, and high spreading probability. In the case of low spreading probability (Fig. 6), we see that the percolation, betweenness, degree, and PageRank measures behave similarly, and also, they perform better than other strategies. In the setting of normal spreading probability (Fig. 7), we have a greater negative spreading compared to the case of low probability (which is expected, since the probability of spread is greater), but the quality of the strategies remains consistent with the previous case. In the case where the network is highly influenceable (Fig. 8), we notice that the degree centrality has a decreased performance compared to previous cases and with the betweenness, percolation, and PageRank centralities. Finally, the clustering and the random strategies present poor results in all three cases.

In the case of directed graphs, the number of positively influenced nodes decreases in comparison with the undirected version, as shown in Figs. 9, 10, and 11. In the low probability scenario (Fig. 9), there is a slight improvement in the quality of the degree centrality; however, we can consider that the percolation, degree, betweenness, and

Table 5 Experiments scenarios

Cost function	Graph type	Spreading probability
Uniform	Undirected	[0,0.2], [0, 1], and [0.75,1]
Uniform	Directed	[0,0.2], [0, 1], and [0.75,1]
Degree Penalty	Undirected	[0,0.2], [0, 1], and [0.75,1]
Degree Penalty	Directed	[0,0.2], [0, 1], and [0.75,1]

Table 6 Size of k (as a function of the % of $|V|$)

Network	0.1%	0.5%	1%	1.5%	2%
CORA	23	115	231	347	463
DBLP	12	62	124	187	249
Wiki	7	35	70	105	141

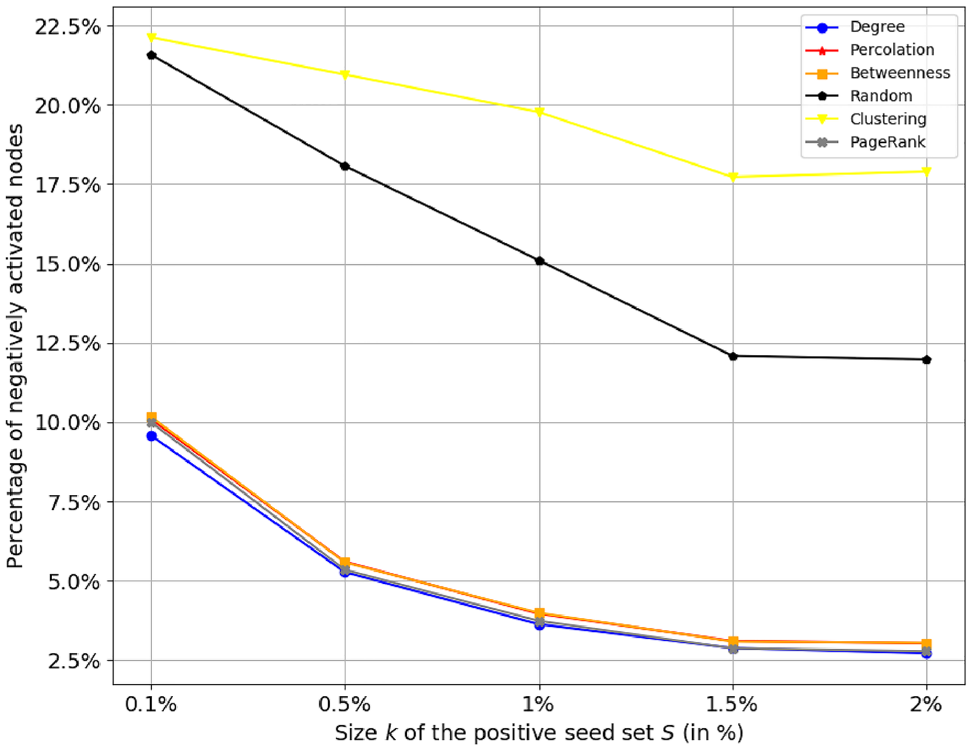


Fig. 6 Uniform cost function in undirected graphs with low spread

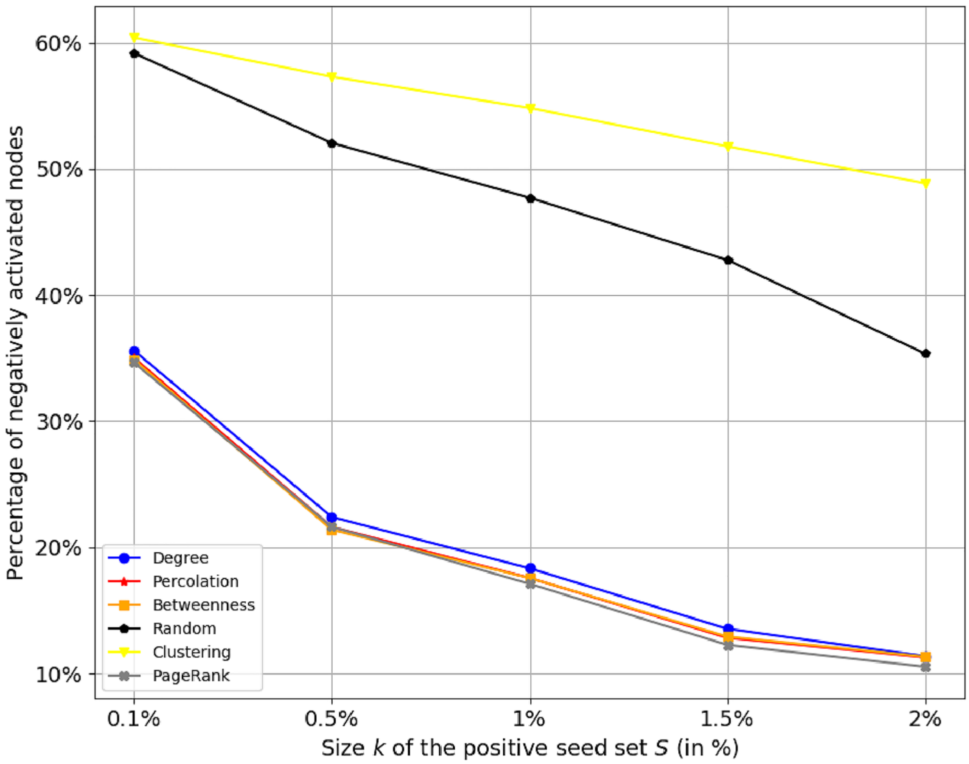


Fig. 7 Uniform cost function in undirected graphs with normal spread

PageRank strategies again have statistically equivalent performances within a small margin of error. With normal probability of spread (Fig. 10), PageRank's behavior apparently worsens in comparison with the previous cases. The last case for directed graphs on the uniform cost function (Fig. 11) is similar to the previous case (normal probability), with the only difference being the greater number of negatively influenced nodes.

Fig. 8 Uniform cost function in undirected graphs with high spread

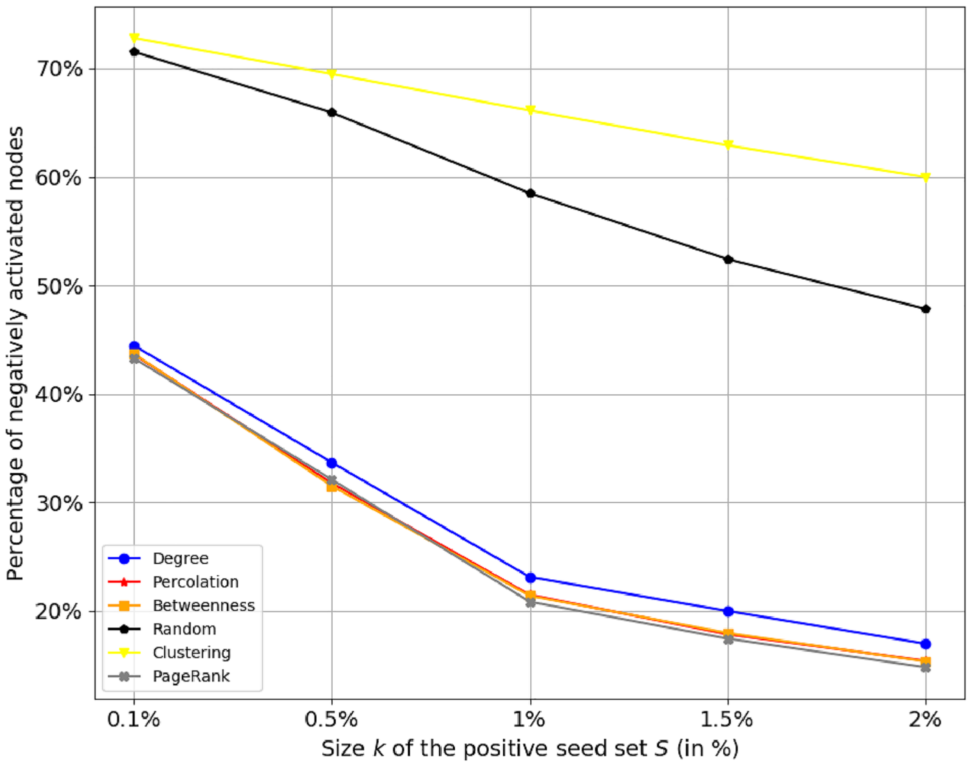
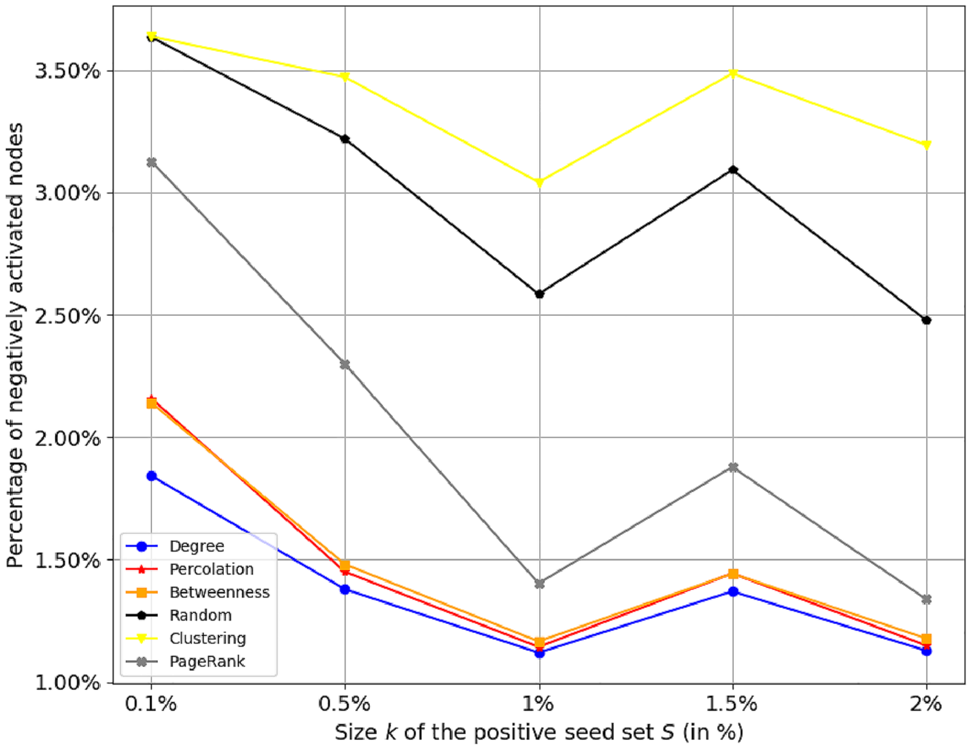


Fig. 9 Uniform cost function in directed graphs with low spread



We hypothesize that the similar behaviors between degree, betweenness, PageRank, and percolation come from the fact that the set of positive seeds chosen by these strategies are similar. In order to test this hypothesis, we take the set of positive seeds of the degree centrality as a basis for the comparison and measure the similarity between the sets returned by the other strategies. More formally, let S_1 and S_2 be the sets returned by using, respectively, the degree

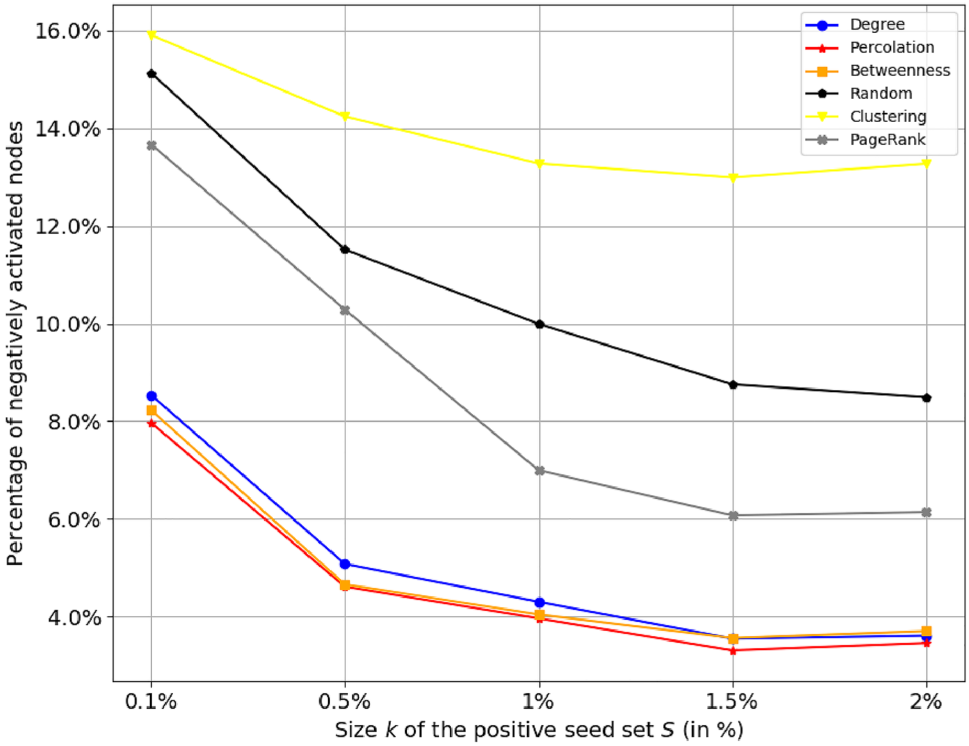


Fig. 10 Uniform cost function in directed graph with normal spread

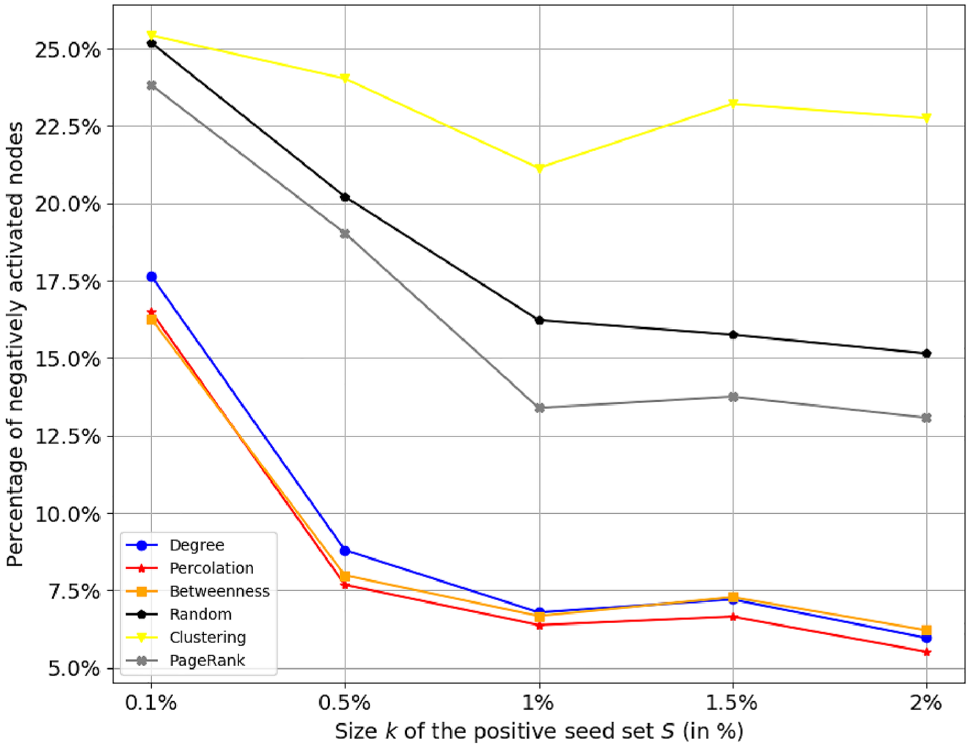


Fig. 11 Uniform cost function in directed graph with high spread

centrality and some other strategy. To measure the similarity between the sets, we use the *overlap coefficient*, defined as $\frac{|S_1 \cap S_2|}{\min\{|S_1|, |S_2|\}}$.

Figure 12 shows the results of the similarities between the solutions on the DBLP dataset. On the vertical axis, we have the overlap coefficient, taking the degree centrality as the base comparison. The horizontal axis represents the size of the set, and we show solutions up to 250 nodes since this is roughly the size of the largest sets for the solutions in the experiments and, additionally, with the solution size approaching the entire node set obviously they have a large overlap. We note that solutions using betweenness, percolation, and PageRank as strategies have a high overlap coefficient. This means that the solution sets returned by these strategies are similar to the degree centrality strategy. On the other hand, solutions obtained using clustering coefficient and random sampling as strategy have a very small overlap, so they are very different from the set nodes with highest degree.

7.2 Degree penalty cost function

In this section, we analyze the results for the *degree penalty* cost function. In this case, the costs are directly proportional to the degree, so we define the sizes of N_0 and S as a fraction of the sum of the degrees (i.e., twice the number of edges). More specifically, we set the size of N_0 to be equal to 1% of the sum of the degrees and choose k to be 0.1%, 0.5%, 1%, 1.5%, and 2% of that same sum. In Table 7, we show the parameter k for each percentage scenario.

Initially, we analyze the results for undirected graphs. Differently from the case with uniform cost function where the node degree is the central attribute that characterize the success of a given strategy, in the degree penalty cost function the node weight “amortizes” the advantage that the degree exerts in these strategies where high-degree nodes are prioritized, i.e., betweenness, percolation, and PageRank (recall Fig. 12 where we show the overlap of such strategies with the set of highest degree nodes). Therefore, these strategies are not as successful in the scenario using the degree penalty cost function as shown in Figs. 13, 14, and 15 (for low, normal, and high propagation, respectively). Generally speaking, compared to the *uniform* cost function, the degree penalty cost function had more negatively influenced nodes in the three settings of the spreading probability. Also, in the degree penalty cost function problem, the clustering and random strategies present the best performances among the metrics we choose.

In particular, we believe that the good performance of the clustering coefficient can be explained by the dissimilarity

Table 7 Size of k (as a function of the % of $2|E|$)

Network	0.1%	0.5%	1%	1.5%	2%
CORA	178	891	1783	2674	3566
DBLP	99	495	991	1487	1983
Wiki	201	1007	2014	3021	4028

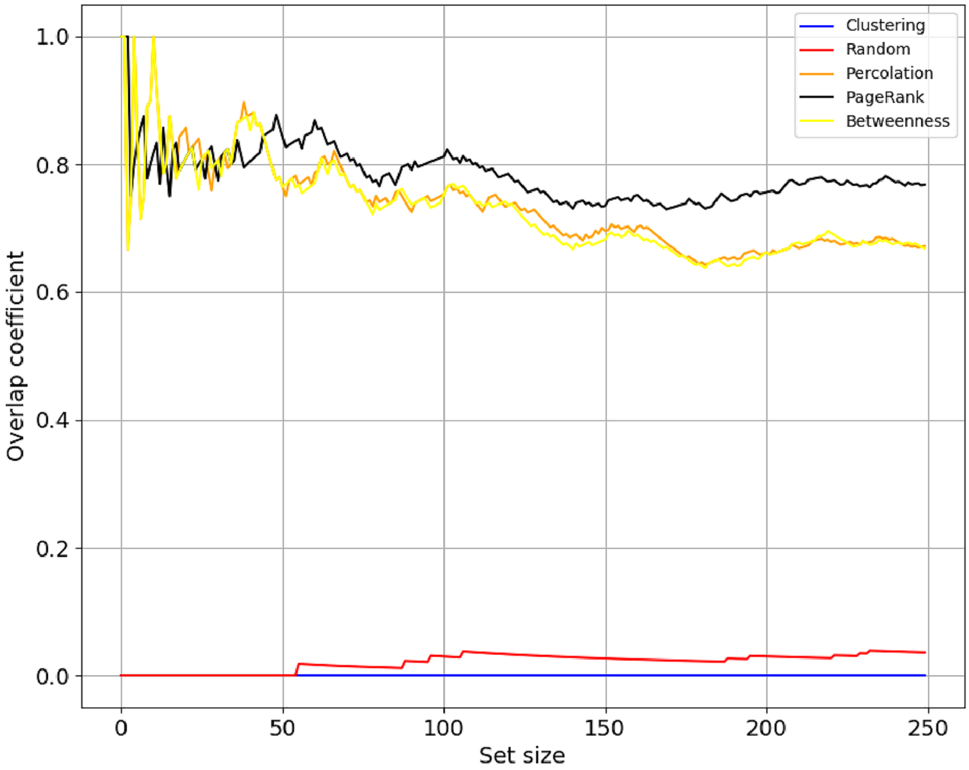


Fig. 12 Overlap coefficient in DBLP undirected graph: value 0 (resp. value 1) is the case where the elements of the solution are completely different (resp. exactly the same) from nodes of k highest degrees

Fig. 13 Degree penalty cost function in undirected graph with low spread

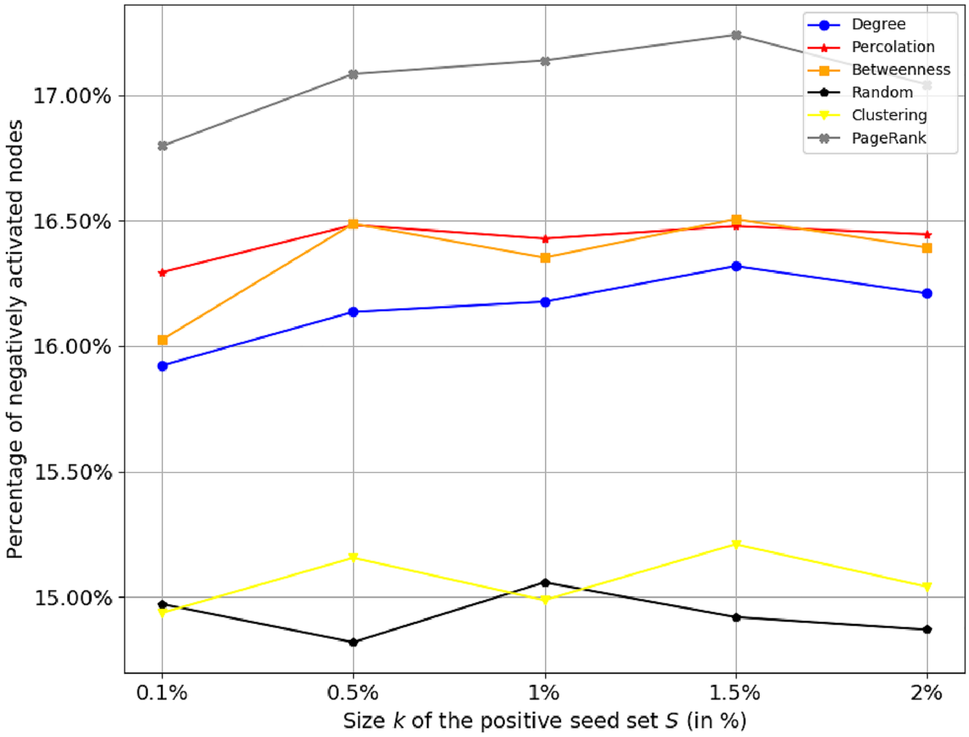
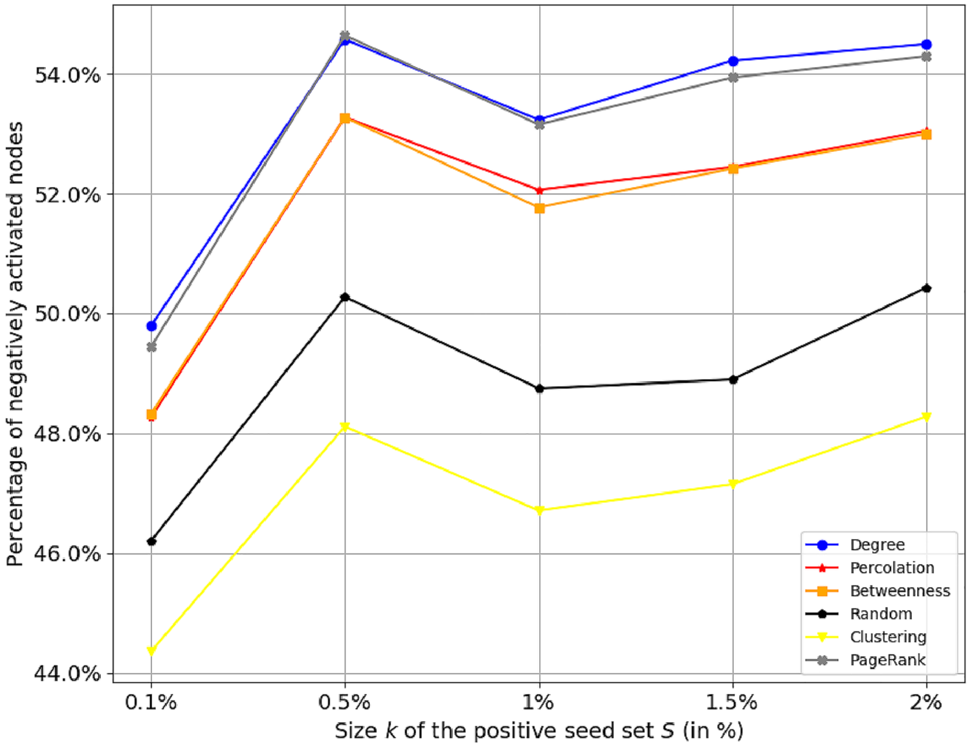


Fig. 14 Degree penalty cost function in undirected graph with normal spread



between this measure and the degree centrality (also shown in Fig. 12). In the analyzed datasets, the nodes with the highest clustering coefficient are those with the lowest degree.

Since the strategy that uses the clustering coefficient selects nodes with low degree, this means that it chooses a large number of nodes for the solution, since the cost of the nodes in this case is low. Therefore, the clustering coefficient strategy may succeed by being able to choose a high fraction of the nodes of a graph.

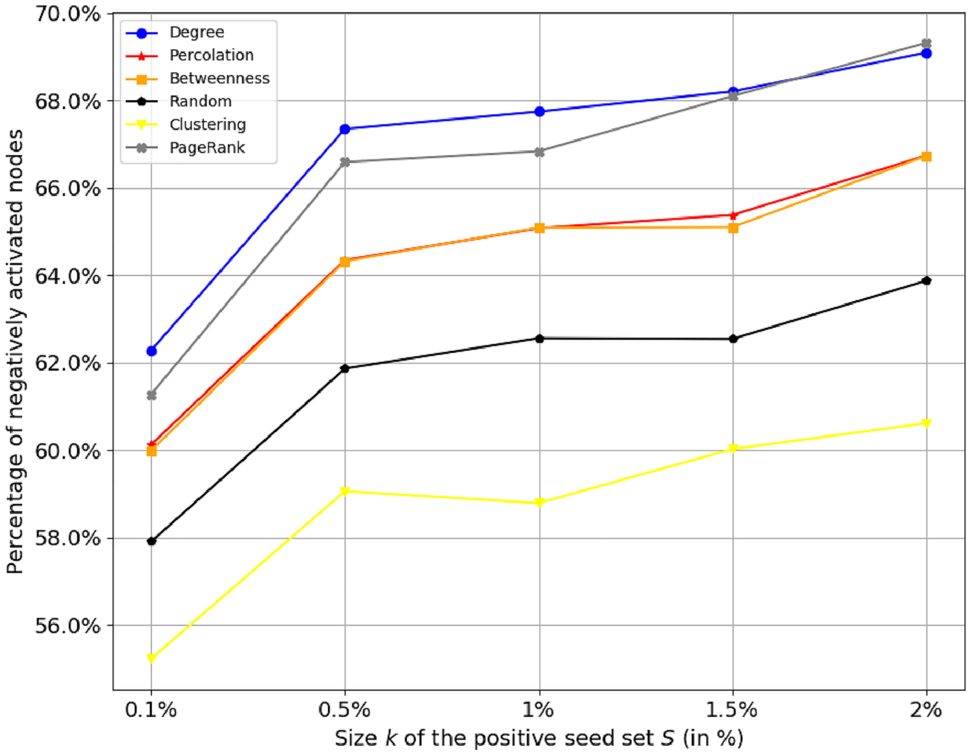


Fig. 15 Degree penalty cost function in undirected graph with high spread

The random strategy also had good results, and we have some supposition for its success. Since, in real-world graphs, typically, the degree distribution is approximated by a power law distribution, roughly speaking theses graphs contain a large number of low-degree nodes. This may explain, in part, the good performance of the random metric. The idea is that by randomly selecting the graph nodes, the vast majority are low-degree nodes and therefore more nodes are selected until reaching the maximum budget limit.

The behavior of degree penalty cost function in directed graphs is different from the other cases analyzed so far. Figure 16 shows the case where the network has a low spreading probability. In this figure, we see that the metrics perform practically the same, with the exception of the PageRank, which presents a slightly inferior performance. Figure 17 considers the normal spreading probability. In this case, betweenness and percolation show a better result than the other metrics. Finally, Fig. 18 shows the results in a highly influential environment. The percolation and betweenness metrics continue to show good results; however, the random strategy also ends up having a result close to them.

The clustering coefficient strategy had opposite performances in the directed and undirected cases. A possible explanation is that in the directed case, the nodes chosen by this strategy have low degree. Thus, due to the edge directions, many may have out-degree equal to zero, making the spreading impossible.

8 Conclusion

In this work, we present the generalized influence blocking maximization (GIBM) problem and analyze the behavior of strategies for the problem based on well-known network metrics for two particular cost functions: uniform and degree penalty. The uniform cost function case has appeared in the literature as the influence blocking maximization problem. For this case, the betweenness, percolation, and PageRank metrics obtain similar results to the simple degree centrality. We show that this similarity is related to the overlapping of the solution sets. On the other hand, in the degree penalty case, the results show that the same metrics have opposite performances. In addition, our results suggest at least two conclusions in the case of algorithms that have a high level of similarity with the node degree. First, however sophisticated an algorithm for the uniform cost function may be, if there is a high similarity with the degree, then one should not expect substantial improvements in their performance. So this might be the case that recent results in the literature obtained only slight improvements (about 1% better) when compared with the node degree strategy in the uniform cost function case. Second, algorithms with solutions correlated with the set of high-degree nodes do not perform well in degree penalty scenario (in any case for the GIBM problem where high-degree nodes are expensive). This naturally leads us to consider futures research where goal is to design solutions that take into consideration other cost functions

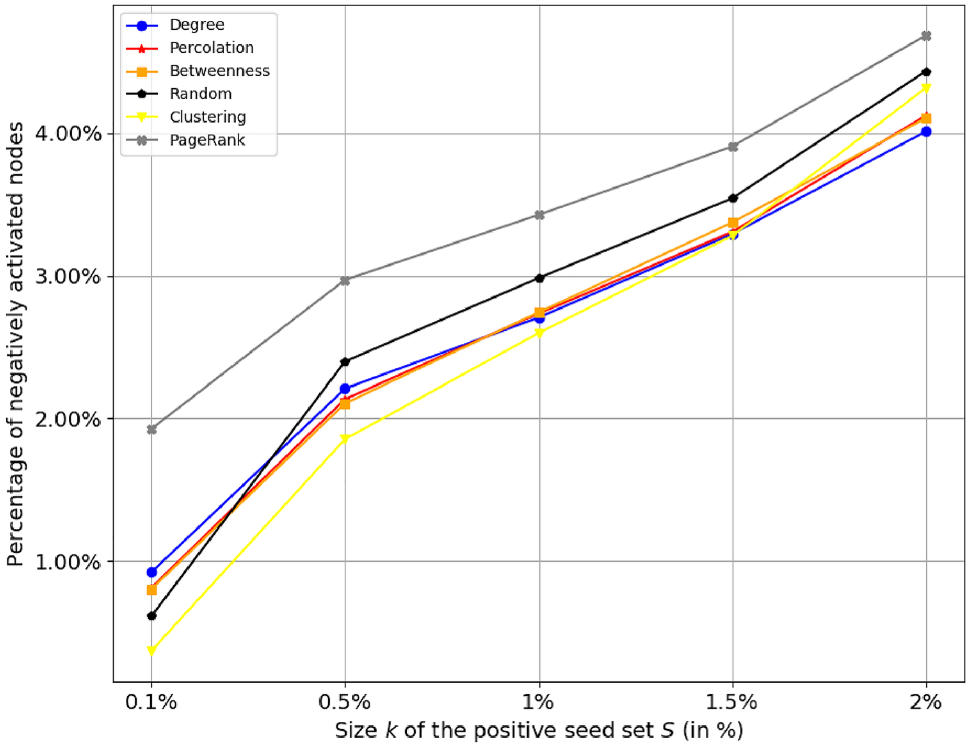


Fig. 16 Degree penalty cost function in directed graphs with low spread

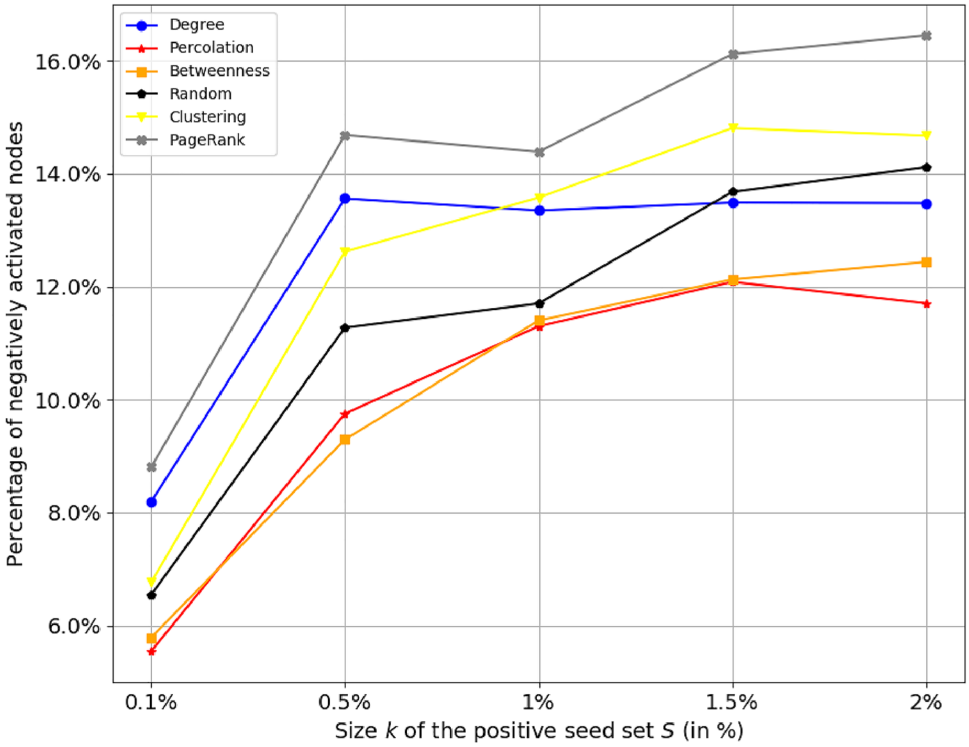
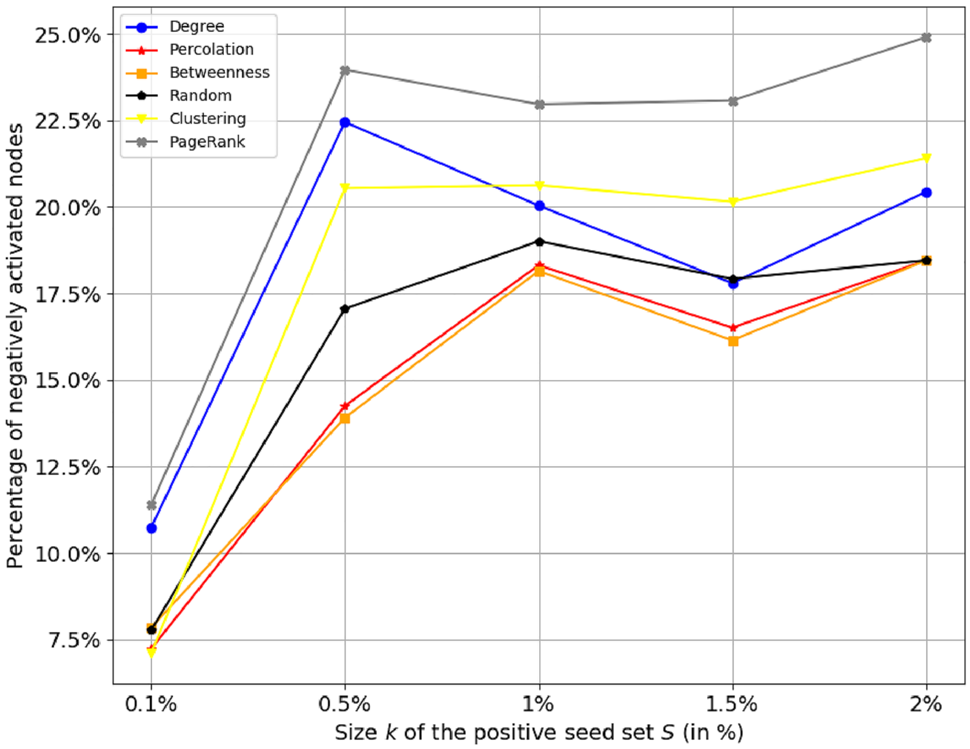


Fig. 17 Degree penalty cost function in directed graphs with normal spread

for the generalized version of the problem. We also analytically address submodularity and approximation properties for the COICM, MCICM, and CLT models. We show that the problem admits constant approximation for the COICM and CLT models, assuming weak suppositions in the latter. In the MCICM, we show that the submodularity property does not hold, which prevents the use of a greedy method to obtain an approximation.

Fig. 18 Degree penalty cost function in directed graphs with high spread



References

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31:211–236. https://doi.org/10.1257/jep.31.2.211

Arazkhani N, Meybodi MR, Rezvanian A (2019) An efficient algorithm for influence blocking maximization based on community detection. In: 5th International Conference on Web Research (ICWR), pp. 258–263

Arazkhani N, Meybodi MR, Rezvanian A (2019) Influence blocking maximization in social network using centrality measures. In: 5th Conference on Knowledge Based Engineering and Innovation (KBEI), pp. 492–497

Brandes U (2004) A faster algorithm for betweenness centrality. J Mathem Sociol. https://doi.org/10.1080/0022250X.2001.9990249

Budak C, Agrawal D, Abbadi A (2011) Limiting the spread of misinformation in social networks. Proceedings of the 20th International Conference on World Wide Web, WWW 2011 pp. 665–674 . https://doi.org/10.1145/1963405.1963499

Chen W, Yuan Y, Zhang L (2010) Scalable influence maximization in social networks under the linear threshold model. Proceedings - IEEE International Conference on Data Mining, ICDM pp. 88–97 . https://doi.org/10.1109/ICDM.2010.118

Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. ArXiv **abs/2003.05004** (2020)

Cornuejols G, Fisher ML, Nemhauser GL (1977) Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms. Manag Sci 23(8):789–810

Erd FC, Vignatti AL, da Silva MVG (2020) Blocking the spread of misinformation in a network under distinct cost models. IEEE/ACM International Conference on. Advances in Social Networks Analysis and Mining (2020)

Fagiolo G (2007) Clustering in complex directed networks E, Statistical, nonlinear, and soft matter physics. Phys Rev 76(2):026107

Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conference, pp. 11 – 15

He X, Song G, Chen W, Jiang Q (2012) Influence blocking maximization in social networks under the competitive linear threshold model technical report. Proceedings of the 12th SIAM International Conference on Data Mining, SDM . https://doi.org/10.1137/1.9781611972825.40

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 137–146. https://doi.org/10.1145/956750.956769

Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018) The science of fake news. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998

Leskovec J, Huttenlocher D, Kleinberg J (2010) Governance in social media: A case study of the Wikipedia promotion process. In: Proc. Int. Conf. on Weblogs and Social Media

Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, p. 420–429. Association for Computing Machinery, New York, NY, USA . https://doi.org/10.1145/1281192.1281239

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 13(3):106–131. https://doi.org/10.1177/1529100612451018 ((**PMID: 26173286**))

Ley M (2002) The DBLP computer science bibliography: Evolution, research issues, perspectives. In: Proc. Int. Symposium on String Processing and Information Retrieval, pp. 1–10

Nemhauser GL, Fisher LAW (1978) An analysis of approximations for maximizing submodular set functions. Math Program 14:365

Nguyen H, Zheng R (2013) On budgeted influence maximization in social networks. IEEE J Selected Areas Commun 31(6):1084–1094. https://doi.org/10.1109/JSAC.2013.130610

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab . Previous number = SIDL-WP-1999-0120

Pham CV, Duong HV, Hoang HX, Thai MT (2019) Competitive influence maximization within time and budget constraints in online social networks: an algorithmic approach. Appl Sci. https://doi.org/10.3390/app9112274

Piraveenan M, Prokopenko M, Hossain L (2013) Percolation centrality: quantifying graph-theoretic impact of nodes during percolation in networks. PLOS ONE 8(1):1–14. https://doi.org/10.1371/journal.pone.0053095

Šubelj L, Bajec M (2013) Model of complex networks based on citation dynamics. In: Proceedings of the WWW Workshop on Large Scale Network Analysis, pp. 527–530

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559

Wu P, Pan L (2017) Scalable influence blocking maximization in social networks under competitive independent cascade models. Computer Netw. https://doi.org/10.1016/j.comnet.2017.05.004

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.