# Graphical Approach for Influence Maximization in Social Networks Under Generic Threshold-based Non-submodular Model

Liang Ma
IBM T. J. Watson Research
Yorktown, NY, USA
Email: maliang@us.ibm.com

Guohong Cao
Pennsylvania State University
University Park, PA, USA
Email: gcao@cse.psu.edu

Lance Kaplan
Army Research Laboratory
Adelphi, MD, USA
Email: lance.m.kaplan.civ@mail.mil

***Abstract*—As a widely observable social effect, influence diffusion refers to a process where innovations, trends, awareness, etc. spread across the network via the social impact among individuals. Motivated by such social effect, the concept of influence maximization is coined, where the goal is to select a bounded number of the most influential nodes (seed nodes) from a social network so that they can jointly trigger the maximal influence diffusion. A rich body of research in this area is performed under statistical diffusion models with provable submodularity, which essentially simplifies the problem as the optimal result can be approximated by the simple greedy search. When the diffusion models are non-submodular, however, the research community mostly focuses on how to bound/approximate them by tractable submodular functions. In other words, there is still a lack of efficient methods that can directly resolve non-submodular influence maximization problems. In this regard, we fill the gap by proposing seed selection strategies using network graphical properties in a generalized non-submodular threshold-based model, called influence barricade model. Under this model, we first establish theories to reveal graphical conditions that ensure the network generated by node removals has the same optimal seed set as that in the original network. We then exploit these theoretical conditions to develop efficient algorithms by strategically removing less-important nodes and selecting seeds only in the remaining network. To the best of our knowledge, this is the first graph-based approach that directly tackles non-submodular influence maximization. Evaluations on both synthetic and real-world Facebook/Twitter datasets confirm the superior efficiency of the proposed algorithms, which are orders of magnitude faster than benchmarks for large networks.**

***Keywords*-Influence Maximization; Viral Marketing; Influence Barricade Model; Theory; Algorithm; Non-submodular**

## I. Introduction

The development of high-end portable devices and utmost prevalence of social networks around the globe have drastically increased the speed and frequency of interactions among individuals. Unlike communication networks, social networks provide a unique substrate through which social behaviors, e.g., the adoption of innovations, trends, social awareness, etc., can also propagate. In particular, by leveraging the social impact, individuals with certain social behaviors can affect their friends, and the influenced ones can further affect their friends, etc. The spread of such social influence is referred to as *influence diffusion* in social networks. As one canonical application of influence diffusion, viral marketing takes advantage of the "word-of-mouth" effect [1] to promote the vast spread of product awareness and adoption in a viral replication manner. Since the influence among friends is generally more reliable than third-party media (e.g., TV, radio, and billboard), viral marketing demonstrates substantial advantages than traditional commercial campaigns. Motivated by such influence-diffusion applications, *influence maximization* emerges as a fundamental research issue. Specifically, influence maximization explores a computationally efficient way to select a small subset of the most influential nodes from a social network so that the selected nodes can trigger the maximal influence diffusion. These selected nodes are called *seed nodes*, which act as initial influential sources; moreover, seed selection is subject to practical constraints, e.g., budget for free samples, and thus only a small subset is allowed to serve as seed nodes. In addition, when the social networks are extremely large, i.e., in the context of big data [2]–[4], selecting the most influential nodes efficiently is even more challenging.

For influence maximization, it is first investigated as an algorithmic problem by [1], [5] using statistical models and parameter estimation. Following their work, influence maximization is formulated as a discrete optimization problem in [6], where two diffusion models, Independent Cascade model (IC) and Linear Threshold model (LT), are explored. Under these two models, influence maximization is reduced to an easier combinatorial problem where the objective function is submodular. For such submodular influence maximization problems, simple hill-climbing greedy algorithm performs well with guaranteed $(1 - 1/e)$ approximation ratio [6], [7]. As such, a large amount of research works based on IC/LT models are stimulated thereafter, e.g., algorithms with improved complexity in [8]–[12] and optimization under more constraints in [13]–[15]. Due to the limitation of IC/LT in capturing various social effects in influence propagation, other complicated diffusion models are proposed, e.g., [16], [17], which are, however, no longer submodular. For influence maximization under these non-submodular models, one common approach is to find other submodular models that can approximate and/or bound the objective function in the original problem; the derived results associated with these relaxed models are then used as the estimated solution to the original non-submodular problem. In particular, [16] establishes a Sandwich Approximation (SA) strategy that bounds the original problem from both sides (lower/upper bounds) by two submodular functions, and proves the approxiamation ratio of the solution derived from these submodular problems. However, depending on how the bounding submodular functions are chosen, the corresponding result can be arbitrarily worse than the optimal solution. Therefore, there is still a lack of efficient seed selection strategies that can directly tackle non-submodular influence maximization problems. In this regard, we target to investigate efficient algorithms specifically designed to solve non-submodular problems by leveraging network intrinsic attributes, i.e., graphical properties.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

In this paper, we consider one generic threshold-based non-submodular model, called *influence barricade model*, where the influence requirement is a hard constraint (not a random variable) that must be satisfied on all cases for the spread of influence. More importantly, no parameter constraints (e.g., social impacts and influence requirements) in [6], [17] are imposed to this model. Such generic model can capture any barricade-like social effect that is abundant [18], [19] in real life. Moreover, this barricade model incorporates the case where there may exist people who are too conservative to be influenced by their friends, which is generally ignored in other threshold-based models. Under such generic model, we first establish theoretical conditions in terms of network graphical properties to reveal the relationships between the optimal seed sets in two networks that differ by only a few nodes/edges; the established theories indicate under what conditions these two graphs maintain the same (or similar) optimal seed set. Based on such fundamental understanding, we then develop efficient seed selection algorithms for non-submodular influence maximization problems. Evaluations in both synthetic and real Facebook/Twitter networks confirm the orders of magnitude improvement of our graph-based algorithms, especially in large networks.

The remainder of the paper is organized as follows. Section II formulates the problem. Theoretical results for non-submodular problems are presented in Section III, based on which seed selection algorithms are developed in Section IV. Experiments under both synthetic and real networks are conducted in Section V. Section VI concludes the paper.

## II. Problem Formulation

In this section, we first introduce the generic threshold-based diffusion model. We then formally formulate our influence maximization problem and state our research objective.

### A. Influence Diffusion in Social Networks

Let directed graph $\mathcal{G} = (V, L)$ represent a social network, where $V$ ($L$) denotes the set of nodes (directed edges). $\mathcal{G}$ can be either a connected or disconnected graph; our proposed algorithms are independent of the network connectivity. In $\mathcal{G}$, $\overrightarrow{uv} \in L$ is a directed edge starting from node $u$ and pointing to node $v$; in addition, edge $\overrightarrow{uv}$ associates with a positive weight $W_{\overrightarrow{uv}}$ (i.e., $W_{\overrightarrow{uv}} > 0$, which can be *different* for different edges), which represents the quantified social influence of node $u$ on node $v$. In $\mathcal{G} = (V, L)$, regarding $u \in V$, $v$ is called an *out-neighbor* (or *in-neighbor*) of $u$ if there exists $\overrightarrow{uv} \in L$ (or there exists $\overrightarrow{vu} \in L$). For each node in $\mathcal{G}$, its status is either *active* or *inactive*. A node $u \in V$ is active if $u$ adopts the innovations, trends, awareness, etc., that are being propagated in the network; otherwise, $u$ is inactive. In this paper, we consider the progressive diffusion process, where node status can switch from inactive to active, but status cannot change in the opposite direction. In social network $\mathcal{G} = (V, L)$, let $S$ ($S \subseteq V$) be the set of initial active nodes, referred to as *seed nodes*, which act as the original influence sources (e.g., nodes selected by a company for a product campaign). Based on such seed set, inactive nodes may become active in subsequent time steps via influence diffusion.

### B. Influence Barricade Model

In this paper, we consider the following generic threshold-based diffusion model, called *influence barricade model*. Suppose $\forall u \in V$, $u$ associates with a fixed non-negative influence threshold $b_u$ (i.e., $b_u \geq 0$, which can be *different* for different nodes), called the *barricade factor*, which determines the difficulty level of $u$ to be influenced by its neighbors. In particular, given the network state at time $t$, for any inactive node $u$, $u$ becomes active at time $t+1$ if and only if the sum weight from all its active in-neighbors is at least $b_u$ (i.e., $\sum_{v \text{ is an active in-neighbor of } u} W_{\overrightarrow{vu}} \geq b_u$); such diffusion cascade proceeds as a *discrete*-time process. In this model, no constraints are imposed on the values of barricade factors or the edge weights in the social network. Here, the barricade factor with respect to a node represents the hard and uncompromisable requirement that must be satisfied by its active neighbors so as to influence this node.

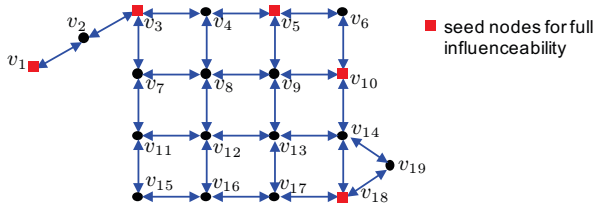



Fig. 1. Seed selection in a sample network (weight is 1 for all edges; barricade factor is 2 for all nodes; $v_1$ must be selected as a seed node for full influenceability).

### C. Influence Maximization Problem

With the influence barricade model, we now formally define the influence maximization problem. Let $A(S)$ denote the set of all active nodes ($S \subseteq A(S)$) that are influenced by seed set $S$ at the moment when no other nodes can become active even if more time is given; this particular moment is called *the end of the influence process* in the sequel. Let $k$ be the maximum number of seed nodes that are allowed to be selected from a network, i.e., $|S| \leq k$, and $\sigma(S) := |A(S)|$ the total number of active nodes at the end of the influence process. Our objective is to maximize the influence by selecting a constrained number of seed nodes, as stated below.

**Objective:** Given graph $\mathcal{G} = (V, L)$, select a set of seed nodes $S$ from $V$ with $|S| \leq k$ such that $\sigma(S)$ is maximized under the influence barricade model.

Let seed set $S_k^*$ denote the optimal solution to the above problem when up to $k$ seed nodes are allowed to be selected from $V$. We aim to determine $S_k^*$ under any $k$. Regarding the global network status, we call the network under one seed set achieves *full influenceability* if every node is active at the end of the influence process; otherwise, we call it *partial influenceability*. With this concept, let $S_{\mathcal{G}}^*$ be a minimum seed set required for full influenceability in $\mathcal{G} = (V, L)$, i.e., $\sigma(S_{\mathcal{G}}^*) = |V|$ and $\forall S \in \{Q \subseteq V : |Q| < |S_{\mathcal{G}}^*|\}$ $\sigma(S) < |V|$. Then we can further define $S_k^*$ as $S_k^* = S_{\mathcal{G}}^*$ if $k \geq |S_{\mathcal{G}}^*|$ since $\sigma(\cdot)$ already achieves the maximum when $k = |S_{\mathcal{G}}^*|$. Moreover, $S_k^*$ generally is not unique; see Section II-D for examples.

### D. Illustrative Example

Fig. 1 shows a sample bidirectional network with 19 nodes. Suppose the weight is 1 for each edge and the barricade factor is 2 for each node. Then the optimal seed set for full influenceability is $S_{\mathcal{G}}^* = \{v_1, v_3, v_5, v_{10}, v_{18}\}$. Let $A_t(S)$ be the set of active nodes at time step $t$ under seed set $S$, where $A_0(S) = S$. Then $A_0(S_{\mathcal{G}}^*) = S_{\mathcal{G}}^*$, $A_1(S_{\mathcal{G}}^*) = A_0(S_{\mathcal{G}}^*) \cup \{v_2, v_4, v_6, v_{14}\}$, $A_2(S_{\mathcal{G}}^*) = A_1(S_{\mathcal{G}}^*) \cup \{v_9, v_{19}\}$, $A_3(S_{\mathcal{G}}^*) = A_2(S_{\mathcal{G}}^*) \cup \{v_8, v_{13}\}$, $A_4(S_{\mathcal{G}}^*) = A_3(S_{\mathcal{G}}^*) \cup \{v_7, v_{12}, v_{17}\}$, $A_5(S_{\mathcal{G}}^*) = A_4(S_{\mathcal{G}}^*) \cup \{v_{11}, v_{16}\}$, $A_6(S_{\mathcal{G}}^*) = A_5(S_{\mathcal{G}}^*) \cup \{v_{15}\}$; therefore, $\sigma(S_{\mathcal{G}}^*) = 19$. In

$S^*_{\mathcal{G}}$, node $v_1$ must be selected as a seed node for full influenceability, as $v_1$ cannot be influenced even if all other nodes in the network are active. Given $S^*_{\mathcal{G}}$, we know that $S^*_k = \{v_1, v_3, v_5, v_{10}, v_{18}\}$ for $k \geq 5$. For other values of $k$, we have $S^*_1 = \{v_{18}\}$ with $A(S^*_1) = \{v_{18}\}$, $S^*_2 = \{v_{13}, v_{18}\}$ with $A(S^*_2) = \{v_{13}, v_{14}, v_{17}, v_{18}, v_{19}\}$, $S^*_3 = \{v_8, v_{13}, v_{18}\}$ with $A(S^*_3) = \{v_8, v_9, v_{10}, v_{12}, v_{13}, v_{14}, v_{16}, v_{17}, v_{18}, v_{19}\}$, and $S^*_4 = \{v_3, v_8, v_{13}, v_{18}\}$ with $A(S^*_4) = \{v_i : i = 3, 4, \ldots, 19\}$. Note that the optimal solution is not unique, e.g., seed set $\{v_4, v_9, v_{14}, v_{15}\}$ is also optimal when $k = 4$, and seed set $\{v_1, v_3, v_8, v_{13}, v_{18}\}$ also achieves full influenceability using 5 seed nodes.

## III. Theoretical Foundations for Seed Set Selection

Under the influence barricade model, we prove that the influence maximization problem is NP-hard. Note that all theoretical proofs in this paper are presented in [20] due to page limitations.

*Theorem 1:* Maximizing $\sigma(S)$ subject to $|S| \leq k$ is NP-hard under the influence barricade model.

Therefore, we need to develop heuristic solutions. In addition, the influence barricade model causes the influence maximization problem to be non-submodular, as stated below.

*Claim 2:* The influence maximization problem under the influence barricade model is *not* submodular.

Therefore, influence maximization under the influence barricade model is a challenging problem, which requires fundamental understanding of the linkage between network properties and node importance levels of being seeds. In this regard, we investigate how minor changes to the network (e.g., add/remove a node) may affect set $S^*_{\mathcal{G}}$ (recall that $S^*_{\mathcal{G}}$ denotes the minimum seed set for full influenceability in $\mathcal{G}$). The idea for our following-up algorithm design is that if a certain sequence of minor changes is proved to have little or no impact on $S^*_{\mathcal{G}}$, then we are able to narrow $S^*_{\mathcal{G}}$ down to the nodes in the remaining network. Since node changes generally also incur edge changes, we therefore first study the impact of minor edge changes on the optimal seed set selection.

*Theorem 3:* Let $\mathcal{G} = (V, L)$ and $\mathcal{G}' = (V, L \cup \{\overrightarrow{v_1v_2}, \overrightarrow{v_2v_1}\})$ with the corresponding minimum seed set for full influenceability being $S^*_{\mathcal{G}}$ and $S^*_{\mathcal{G}'}$ respectively, where $v_1$, $v_2 \in V$ and $\overrightarrow{v_1v_2}, \overrightarrow{v_2v_1} \notin L$. Then $0 \leq |S^*_{\mathcal{G}}| - |S^*_{\mathcal{G}'}| \leq 1$.

Following similar arguments in the proof [20] of Theorem 3, the following corollary also holds.

*Corollary 4:* Let $\mathcal{G} = (V, L)$ and $\mathcal{G}' = (V, L \cup \{\overrightarrow{v_1v_2}\})$ with the corresponding minimum seed set for full influenceability being $S^*_{\mathcal{G}}$ and $S^*_{\mathcal{G}'}$ respectively, where $v_1, v_2 \in V$ and $\overrightarrow{v_1v_2} \notin L$. Then $0 \leq |S^*_{\mathcal{G}}| - |S^*_{\mathcal{G}'}| \leq 1$.

Theorem 3 and Corollary 4 consider a simple case where $\mathcal{G}$ and $\mathcal{G}'$ differ by only one or two edges between a pair of nodes, for which Theorem 3 and Corollary 4 quantify the impact of edge changes by providing lower and upper bound with the gap being only 1. More importantly, these results are independent of the values of $\{b_u\}_{u \in V}$ and $\{W_{\overrightarrow{uv}}\}_{\overrightarrow{uv} \in L}$, thus applicable to networks under any parameter settings.

*Definition 5:* *1)* Node $u$ with $b_u > \sum_{w \in N_{\text{in}}} W_{\overrightarrow{wu}}$, where $N_{\text{in}}$ is the set of all in-neighbors of $u$ in network $\mathcal{G}$, is an *influence deficient node*, for which $u \in S^*_{\mathcal{G}}$ always holds.

*2)* An edge is *redundant* if the removal of which does not affect the size of the minimum seed set for full influenceability, i.e., edges $L'$ in $\mathcal{G} = (V, L)$ are redundant if and only if $|S^*_{\mathcal{G}}| = |S^*_{\mathcal{G}'}|$, where $\mathcal{G}' = \{V, L \setminus L'\}$.

With these definitions, we derive lower/upper bounds to quantify the maximum impact of single node changes on the optimal seed set for full influenceability as follows, where $\mathcal{V}(\mathcal{G}')$ denotes the set of nodes in graph $\mathcal{G}'$.

*Corollary 6:* Given graph $\mathcal{G} = (V, L)$, suppose an extra node $v$ ($v \notin V$) is added and connected to a set of nodes $N$ in $\mathcal{G}$, thus forming a new graph $\mathcal{G}'$. Let $S^*_{\mathcal{G}}$ and $S^*_{\mathcal{G}'}$ denote the minimum seed set for full influenceability in $\mathcal{G}$ and $\mathcal{G}'$ respectively. Then $\max(|M|, |S^*_{\mathcal{G}}|+1-|N|) \leq |S^*_{\mathcal{G}'}| \leq |S^*_{\mathcal{G}}|+1$, where $M = \{w \in \mathcal{V}(\mathcal{G}') : b_w > \sum_{z \in \mathcal{V}(\mathcal{G}')} W_{\overrightarrow{zw}}\}$ (i.e., influence deficient nodes in $\mathcal{G}'$).

We next develop more concrete conditions to characterize the relationships between $S^*_{\mathcal{G}}$ and $S^*_{\mathcal{G}'}$; see [20] for the proof.

*Theorem 7:* Given graph $\mathcal{G} = (V, L)$, suppose node $v$ ($v \notin V$) is added and connected to a set of nodes $N$ in $\mathcal{G}$ ($N \subseteq V$, $N = N_{\text{in}} \cup N_{\text{out}}$, $N_{\text{in}}$: all in-neighbors of $v$, $N_{\text{out}}$: all out-neighbors of $v$), thus forming a new graph $\mathcal{G}' = (V \cup \{v\}, L \cup L')$, where $L' = \{\overrightarrow{wv} : w \in N_{\text{in}}\} \cup \{\overrightarrow{vw} : w \in N_{\text{out}}\}$ and each edge in $L'$ associates with a positive weight. Let $Q$ be the largest subset of $N$ ($Q \subseteq N$) that satisfies the following three conditions.

(a) $\exists$ a minimum seed set $S^*_{\mathcal{G}}$ for full influenceability in $\mathcal{G}$ with $Q \subset S^*_{\mathcal{G}} \subseteq V$;
(b) let $Z := \{z : z \notin Q, z$ is a neighbor of at least one node in $Q\}$. All edges between $Q$ and $Z$ (in either direction) are redundant in $\mathcal{G}$;
(c) $\forall q \in Q$, $\sum_{z \in Z} W_{\overrightarrow{zq}} < b_q \leq W_{\overrightarrow{vq}} + \sum_{z \in Z} W_{\overrightarrow{zq}}$.

Then regarding the minimum seed set $S^*_{\mathcal{G}'}$ for full influenceability in $\mathcal{G}'$, the following results hold.

(1) If $\sum_{w \in N'_{\text{in}}} W_{\overrightarrow{wv}} < b_v \leq \sum_{w \in N_{\text{in}}} W_{\overrightarrow{wv}}$ (for any $N'_{\text{in}} \subset N_{\text{in}}$), $N_{\text{out}} \subseteq N_{\text{in}}$, and $W_{\overrightarrow{vw}} < b_w$ (for any $w \in N_{\text{out}}$), then the necessary and sufficient condition for $|S^*_{\mathcal{G}'}| < |S^*_{\mathcal{G}}|$ is that $|Q| > 1$. Specifically,
$$\begin{cases} S^*_{\mathcal{G}'} = (S^*_{\mathcal{G}} \setminus Q) \cup \{v\}, & |Q| > 1 \\ S^*_{\mathcal{G}'} = S^*_{\mathcal{G}}, & |Q| \leq 1 \end{cases}.$$
(2) If $\sum_{w \in N_{\text{in}}} W_{\overrightarrow{wv}} < b_v$ and $W_{\overrightarrow{vw}} < b_w$ (for any $w \in N_{\text{out}}$), then the necessary and sufficient condition for $|S^*_{\mathcal{G}'}| < |S^*_{\mathcal{G}}|$ is that $|Q| > 1$. Specifically, $S^*_{\mathcal{G}'} = (S^*_{\mathcal{G}} \setminus Q) \cup \{v\}$ for any $|Q| \geq 0$.

For graphs $\mathcal{G} = (V, L)$ and $\mathcal{G}' = (V \cup \{v\}, L \cup L')$, Theorem 7 states under what conditions and to what extent the size of $S^*_{\mathcal{G}'}$ varies. Besides being of theoretical value, these theorems and corollaries can also guide our algorithm design. Specifically, using these results, we are able to identify which nodes are critical in affecting the optimal seed set for full influenceability. Based on such understanding, we develop efficient seed selection algorithms in the next section.

## IV. Seed Set Selection Algorithms

In this section, we build efficient algorithms for seed set selection. Theoretical results in Section III motivate us to remove nodes iteratively with the goal of maintaining the optimal seed set in the remaining graph; this removal process continues until the objective is accomplished. Such removal-based method therefore becomes the fundamental principal to our algorithm design for full and partial influenceability.

### A. Seed Selection for Full Influenceability

We note that, depending on the network structure, the conditions in Theorem 7 are not always easily verifiable, and thus alternative approaches are required. In particular,

**Algorithm 1:** Minimum Seed Selection (MSS)

**input :** Network $\mathcal{G} = (V, L)$, barricade factors $\{b_v\}_{v \in V}$
**output:** Set of seed nodes $S$

1 $U \leftarrow \{u \in \mathcal{V}(\mathcal{G}) : b_u \leq \sum_{z \in \mathcal{V}(\mathcal{G})} W_{\overrightarrow{zu}}\}$; // $\mathcal{V}(\mathcal{G})$: set of nodes in $\mathcal{G}$, $W_{\overrightarrow{zu}} = 0$ if $\overrightarrow{zu}$ does not exist
2 **while** $|U| > 0$ **do**
3  **if** $|U| > 1$ **then**
4  $m_1 \leftarrow \min_{u \in U} \left( \sum_{z \in \mathcal{V}(\mathcal{G})} (W_{\overrightarrow{zu}} + W_{\overrightarrow{uz}}) \right)$;
5  $U \leftarrow \{u \in U : \sum_{z \in \mathcal{V}(\mathcal{G})} (W_{\overrightarrow{zu}} + W_{\overrightarrow{uz}}) = m_1\}$;
6  **end**
7  **if** $|U| > 1$ **then**
8  $m_2 \leftarrow \min_{u \in U}$(number of non-trivial connected components in $\mathcal{G} - u$); // a connected component containing at least two nodes is *non-trivial*
9  $U \leftarrow \{u \in U : \mathcal{G} - u$ has $m_2$ non-trivial connected components$\}$; // $\mathcal{G} - u$: remove $u$ and all edges (in either direction) incident to $u$ from $\mathcal{G}$
10  **end**
11  **if** $|U| > 1$ **then**
12  $m_3 \leftarrow \min_{u \in U}$(number of influence deficient nodes in $\mathcal{G} - u$); // see Definition 5
13  $U \leftarrow \{u \in U : \mathcal{G} - u$ has $m_3$ influence deficient nodes$\}$;
14  **end**
15  $\mathcal{G} \leftarrow \mathcal{G} - q$; // $q$: a randomly picked node from $U$
16  $U \leftarrow \{u \in \mathcal{V}(\mathcal{G}) : b_u \leq \sum_{z \in \mathcal{V}(\mathcal{G})} W_{\overrightarrow{zu}}\}$;
17 **end**
18 $S \leftarrow \mathcal{V}(\mathcal{G})$; // $\mathcal{V}(\mathcal{G})$: set of nodes in $\mathcal{G}$

to enable high efficiency of the proposed algorithm, at each iteration in the removal process, we delete the node that least violates the optimal-set-maintaining conditions in Theorem 7; this process continues until all remaining nodes are influence deficient nodes (see Definition 5). Based on this rule, we develop Algorithm 1, Minimum Seed Selection (MSS), for full influenceability in a given network.

In MSS, lines 2–17 iteratively remove nodes in the given network. For each removed node, MSS guarantees that it is influenceable if all nodes in the remaining graph are active, while also aiming to minimize its impact on the optimal seed set in the remaining graph. Specifically, at each iteration, MSS first identifies set $U$ containing all non-influence-deficient nodes in line 1 (or line 16). For any $u \in U$, $u$ does not need to be a seed node as $u$ can be influenced by the rest of the network (if active); therefore, $u$ can be excluded from the seed set for full influenceability. However, generally $|U| > 1$, i.e., there exist ties. According to Theorem 7, nodes that do not associate with large edge weights from/to their neighbors may maintain the optimal seed set in the rest of the network. Therefore, lines 3–6 refine $U$ by only keeping the nodes with the minimum impact from/to their neighbors. After this step, if the tie still exists, then lines 7–10 further select nodes that incur the minimum number of non-trivial components. Here *non-trivial connected component* after the removal operation refers to a connected component containing at least two nodes, and $\mathcal{G} - u$ denotes removing node $u$ and all edges incident to $u$ from $\mathcal{G}$. Intuitively, when removing a node generates many non-trivial components, a large portion of the original social connections are damaged, thus requiring more seed nodes for full influenceability. Next, if the tie persists, then lines 11–14 only keep nodes resulting in the minimum number of influence deficient nodes. Finally, a randomly picked node from $U$ (if $|U| > 1$) is removed from $\mathcal{G}$ in line 15. This removal process is repeated until all nodes in the remaining graph are influence deficient, i.e., $U = \emptyset$; then all remaining nodes form the selected seed set for full influenceability (line 18). Note that the output of MSS may not be unique as the node removal sequence is not; the performance of MSS is evaluated in Section V.

*Correctness of MSS:* Let $s$ be the time sequence of how nodes are removed. Then the reverse order of $s$ corresponds to one influence diffusion process starting from seeds. Thus, the seed set selected by MSS achieves full influenceability.

*Example:* For the sample network in Fig. 1, one possible node removal sequence (recall this sequence is not unique) by MSS is: (i) nodes $v_{19}$, $v_{15}$, $v_{11}$, $v_{16}$, $v_7$, $v_{12}$, $v_{17}$, $v_8$, $v_{13}$, and $v_9$ by the tie breaking rule in lines 3–6, and (ii) $v_2$, $v_4$, $v_6$, and $v_{14}$ by the tie breaking rule in lines 7–14. Then all remaining nodes $S_{\mathcal{G}} = \{v_1, v_3, v_5, v_{10}, v_{18}\}$ form a seed set for full influenceability. It can be verified that $S_{\mathcal{G}}$ is optimal as no sets with less number of nodes achieve full influenceability.

*Complexity:* In MSS, each iteration for removing a non-influence-deficient node takes $O(|U|)$ time as the properties (the impact from/to neighbors in lines 3–6, the resulting number of non-trivial connected components in lines 7–10, and the resulting number of influence deficient nodes in lines 11–14) associated with each $u \in U$ can be computed in a constant time. In addition, when updating $U$ after the removal of node $q$ (line 16), only nodes that are neighbors of $q$ (before $q$ is removed) are candidates to be included in $U$, because other nodes have the unchanged sets of neighbors. Therefore, the total time complexity of MSS is $O(cn) = O(n)$ ($n = |V|$ and $c$ is a constant), i.e., linear time complexity.

**Algorithm 2:** Seed selection for Influence Maximization (SIM)

**input :** Network $\mathcal{G} = (V, L)$, barricade factors $\{b_v\}_{v \in V}$, budget $k$
**output:** Set of seed nodes $S$

1 $S \leftarrow$ seed set selected by MSS (Algorithm 1);
2 **while** $k < |S|$ **do**
3  $u = \arg\max_{s \in S} \sigma(S \setminus \{s\})$; // $\sigma(\cdot)$ is determined by the discrete cascade process as illustrated in Fig. 1
4  $S \leftarrow S \setminus \{u\}$;
5 **end**

### B. Seed Selection for Partial Influenceability

For partial influenceability, we propose Seed selection for Influence Maximization (SIM) to further refine the seed set $S$ generated by MSS by selecting the top-$k$ critical nodes from $S$, as presented in Algorithm 2. In Algorithm 2, SIM removes nodes from set $S$ constructed by Algorithm 1 iteratively until the seed budget $k$ is reached.

*Complexity:* Since the discrete cascade process can be determined in $O(n)$ time under a given seed set, line 3 in Algorithm 2 takes $O(n|S|)$ time. Let $m$ denote the size of seed set generated by Algorithm 1. Then the total time complexity for Algorithm 2 is $O(nm^2)$.

## V. Experiments

To evaluate the performance of MSS and SIM, we conduct a set of experiments on both randomly-generated and real Facebook and Twitter networks [21]. For comparison, we use the classic greedy algorithm as a benchmark.

### A. Synthetic Networks

We first consider synthetic topologies generated according to the widely used Random Geometric (RG), Erdös-Rényi, and Random Power Law graph models [22]. In this experiment, we obtain similar observations under different graph models; we therefore only report results under RG due to page limitations. Under the RG model [22], nodes are first randomly distributed in a unit square, and then each pair of nodes $u$ and $v$ are connected by edges $\overrightarrow{uv}$ and $\overrightarrow{vu}$ if their distance is no larger than a threshold $\theta$. To focus on evaluating MSS and SIM under different network properties, we fix the weight of each edge to 1 under synthetic networks; the impact under various edge weights will be examined in real Facebook/Twitter networks.

In addition to the greedy algorithm, we also compare our algorithms to the optimal results obtained by enumerating all possible seed sets. Since it is expensive to compute the optimal results, we randomly generate small RG graph realizations, with each realization containing only 30 nodes. The evaluation results averaged over 10 graph realizations are reported in Fig. 2. In Fig. 2, we first study the selected seed sets under various budgets ($k$). We tune parameter $\theta$ in the RG model to make sure the expected number of (directed) edges is $E[|L|] = 320$; furthermore, the barricade factor $b_u$ for each node $u \in V$ is selected from $[E[|L]/(2|V|), E[|L]/|V|]$ (i.e., $[5.33, 10.66]$) uniformly at random. The corresponding results are shown in Fig. 2(a). Fig. 2(a) illustrates that when the budget $k$ is small, both SIM and the greedy algorithm achieve the optimal value; moreover, $\sigma(S) = k$ when $k \leq 5$. This is because when the selection budget is extremely small, the social influence can hardly propagate, i.e., barricade factors are large. Nevertheless, when $k$ increases, SIM significantly outperforms the greedy benchmark, especially in the case of $k = 11$, where SIM almost exhibits a 2-fold improvement. In addition, the optimal result further confirms that the objective function under the barricade model is not submodular. In Fig. 2(a), we also note that for a range of $k$ (e.g., $6 \leq k \leq 10$), SIM is unable to accurately approximate the optimal value. Regarding this observation, we argue that SIM is an algorithm that is capable of achieving superior performance over the best-known heuristic (pure greedy solution) while experiencing much smaller complexity. Next, we evaluate the algorithm performance under different network densities (i.e., number of edges to nodes ratio), focusing on seed selection for full influenceability. We adopt the same network parameter settings as those in Fig. 2(a), except that $E[|L|]$ is changed to a set of values, i.e., $E[|L|] = \{320, 400, 480, \ldots, 720\}$; the average results are presented in Fig. 2(b). In Fig. 2(b), it shows that the performance of both MSS and the greedy benchmark improves in networks with high densities, and MSS can even achieves optimality. Furthermore, MSS and the greedy solution converge to the optimal value when the network density is sufficiently high. Intuitively, this is because there exists a large number of optimal seed sets when the network exhibits high density, and thus it is easier for both MSS and the greedy benchmark to find one of them. Finally, we examine how barricade factors may affect the seed selection algorithms. For this goal, let $B_i$ denote $[5.33, 5.33(0.5i + 1.5)]$ ($i = 1, 2, \ldots, 6$), and barricade factor $b_u$ is selected from $B_i$ uniformly at random for each $i$. We compare the cardinality of the generated seed sets for full influenceability under various $B_i$ ($i = 1, 2, \ldots, 6$), as reported in Fig. 2(c). Fig. 2(c) shows that when the average barricade factors are large, MSS accurately approximates the optimal value, owing to the fact that MSS strategically removes nodes unnecessary to serve as seeds. However, the greedy benchmark generates seed sets that are almost 25% more than necessary, thus resulting in a waste of resources.

TABLE I
SEED SELECTION IN THE FACEBOOK/TWITTER NETWORK FOR FULL INFLUENCEABILITY UNDER VARIOUS BARRICADE FACTORS ($\forall u \in V\ b_u \in [5, 5(i+1)]$ IN $B_i$)

(a) Facebook Network

| | Algorithm | $B_1$ | $B_2$ | $B_3$ | $B_4$ |
|---|---|---|---|---|---|
| #Seed Nodes | MSS | 140.0 | 203.0 | 242.8 | 270.8 |
| | Greedy | 157.0 | 206.8 | 243.5 | 278.5 |
| Running Time (s) | MSS | 0.9 | 0.6 | 0.5 | 0.4 |
| | Greedy | 4798 | 4597 | 4543 | 4700 |

(b) Twitter Network

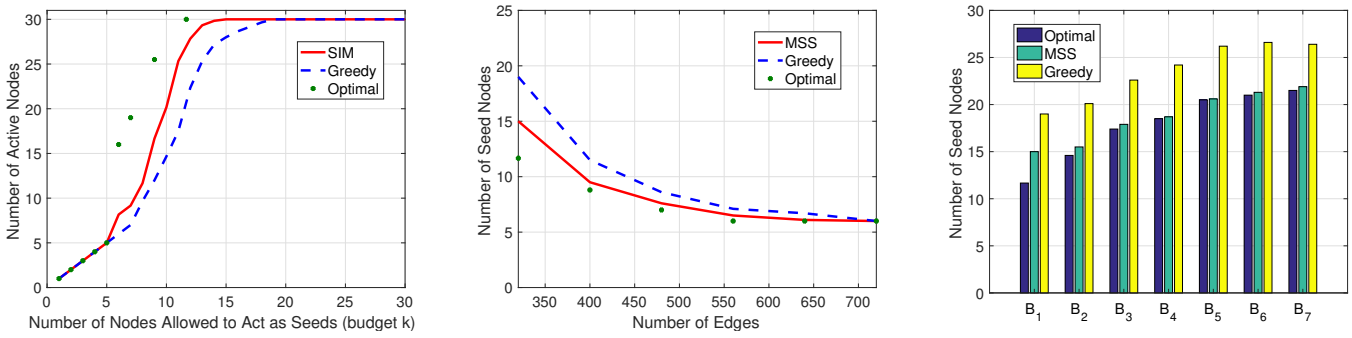
| | Algorithm | $B_1$ | $B_2$ | $B_3$ | $B_4$ |
|---|---|---|---|---|---|
| #Seed Nodes | MSS | 139.0 | 211.0 | 257.0 | 285.3 |
| | Greedy | 164.8 | 230.0 | 280.0 | 314.0 |
| Running Time (s) | MSS | 1.3 | 1.0 | 0.9 | 0.8 |
| | Greedy | 5058 | 5180 | 5871 | 5327 |

### B. Facebook and Twitter Networks

We next use publicly available social network datasets collected by the Stanford SNAP project [21], from which we select two representative networks, Facebook and Twitter, for algorithm evaluations. For these two selected networks, (i) Facebook is a bidirected graph as the existence of edge $\overrightarrow{uv}$ directly implies the existence of edge $\overrightarrow{vu}$, whereas (ii) Twitter is not a bidirected graph as edge $\overrightarrow{uv}$ does not necessarily imply the existence of edge $\overrightarrow{vu}$. In this dataset, there are $4,039$ ($107,614$) nodes and $176,468$ ($13,673,453$) directed edges in the Facebook (Twitter) network; the huge sizes of these networks cause extremely long running time in the algorithm evaluation. As such, we randomly sample a smaller subgraph, i.e., subgraph with 500 (600) nodes and $7,280$ ($9,899$) directed edges, from the original Facebook (Twitter) dataset for evaluations.

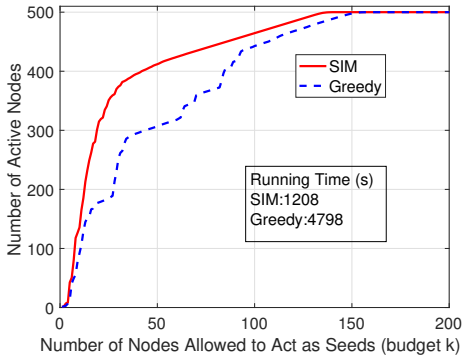
Using real network data, again we first examine the number of active nodes achieved under various budgets ($k$). In particular, we set edge weights (barricade factors) to be numbers generated between 1 and 2 (5 and 10) uniformly at random. Ten such parameter realizations are generated as the algorithm input; the results averaged over these tests are reported in Fig. 3. Similar to random graphs, again we observe that SIM outperforms the greedy benchmark by up to 2-fold (e.g., when $k \approx 25$ in Fig. 3(a) and $k \approx 40$ in Fig. 3(b)) in both Facebook and Twitter networks. Besides, we also report the algorithm average running time for the whole spectrum of $k$, i.e., $k = 1, 2, \ldots, |V|$, in Fig. 3. These results show that SIM achieves a roughly 4-fold speedup, thus confirming the efficiency of SIM.

Next, we study the algorithm performance under different barricade factors. Similar to Fig. 2(c), we select a set of barricade factors, i.e., $B_i = [5, 5(i + 1)]$ ($i \in \{1, \ldots, 4\}$). Under each $B_i$, $b_u$ of node $u \in V$ has the value selected from $B_i$ uniformly at random, and edge weights are still randomly and uniformly generated between 1 and 2. The experiment is repeated 10 times for each $B_i$; the average results for full influenceability are reported in Table I. In Table I, we first observe that MSS outperforms the greedy benchmark in term of the number of selected seed nodes; however, the improvement is not always significant. This is because, as suggested by Fig 2(b), when the network density ($E[|L|]/E[|V|]$ is 14.6 for Facebook and 16.5 for Twitter) is high, the greedy benchmark is likely to experience improved performance. Besides, we also report the algorithm running time in Table I, which justifies the momentous advantages of MSS in comparison

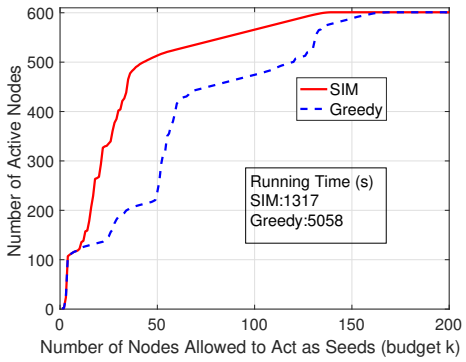



(a) (b) (c)

Fig. 2. Evaluations on Radom Geometric graphs ($|V| = 30$, $\forall \overrightarrow{uv} \in L\ W_{\overrightarrow{uv}} = 1$, 10 graph realizations under each parameter setting). (a) Number of achieved active nodes under different budgets: $E[|L|] = 320$, $\forall u \in V\ b_u \in [5.33, 10.66]$; (b) Size of the seed set for full influenceability under different $E[|L|]$: $E[|L|] = \{320, 400, 480, \ldots, 720\}$, $\forall u \in V\ b_u \in [5.33, 10.66]$; (c) Size of the seed set for full influenceability under different $\{b_u\}_{u \in V}$: $E[|L|] = 320$, $\forall u \in V\ b_u \in [5.33, 5.33(0.5i + 1.5)]$ in $B_i$.




(a) Facebook ($|V| = 500, |L| = 7280$)




(b) Twitter ($|V| = 600, |L| = 9899$)

Fig. 3. Number of achieved active nodes under different budgets in real networks.

to the benchmark. In particular, Table I shows that MSS achieves roughly three orders of magnitude improvement over all ranges of barricade factors for both the Facebook and Twitter networks, which thus confirms the superior efficiency of MSS in finding the seed set for full influenceability.

## VI. Conclusion

We investigated efficient solutions specifically designed for non-submodular influence maximization problems. For this issue, we first established theoretical results to relate the optimal seed set to network graphical properties. Employing these results, we then developed seed selection algorithms to handle non-submodularity by iteratively removing less-critical nodes. Evaluations in both synthetic and real networks confirm the efficacy of the proposed graph-based algorithms in identifying a superior set of the influential nodes as well as the significance in reducing the algorithm execution time.

## References


[1] P. Domingos and M. Richardson, "Mining the network value of customers," in *ACM KDD*, 2001.

[2] L. Wu, I. E. Yen, J. Chen, and R. Yan, "Revisiting random binning features: Fast convergence and strong parallelizability," in *ACM KDD*, 2016.

[3] J. Chen, L. Wu, K. Audhkhasi, B. Kingsbury, and B. Ramabhadrari, "Efficient one-vs-one kernel ridge regression for speech recognition," in *IEEE ICASSP*, 2016.

[4] L. Wu, K. J. Wu, A. Sim, M. Churchill, J. Y. Choi, A. Stathopoulos, C.-S. Chang, and S. Klasky, "Towards real-time detection and tracking of spatio-temporal features: Blob-filaments in fusion plasma," *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 262–275, 2016.

[5] M. Richardson and P. Domingos, "Mining knowledge- sharing sites for viral marketing," in *ACM KDD*, 2002.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *ACM KDD*, 2003.

[7] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *ACM KDD*, 2007.

[9] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *ACM KDD*, 2009.

[10] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *ACM KDD*, 2010.

[11] Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, and K. Xie, "Simulated annealing based influence maximization in social networks," in *AAAI*, 2011.

[12] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *SIAM SODA*, 2014.

[13] H. Ma, H. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," in *ACM CIKM*, 2008.

[14] H. Zhang, T. N. Dinh, and M. T. Thai, "Maximizing the spread of positive influence in online social networks," in *IEEE ICDCS*, 2013.

[15] Y. Zhu, D. Li, and Z. Zhang, "Minimum cost seed set for competitive social influence," in *IEEE INFOCOM*, 2016.

[16] W. Lu, W. Chen, and L. V. S. Lakshmanan, "From competition to complementarity: Comparative influence diffusion and maximization," *Proc. VLDB Endow.*, vol. 9, no. 2, pp. 60–71, 2015.

[17] D.-N. Yang, W.-C. Lee, N.-H. Chia, M. Ye, and H.-J. Hung, "On bundle configuration for viral marketing in social networks," in *ACM CIKM*, 2012.

[18] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[19] S. Aral and D. Walker, "Creating social contagion through viral product design: A randomized trial of peer influence in networks," *Management Science*, vol. 57, no. 9, pp. 1623–1639, 2011.

[20] L. Ma, "Influence maximization under generic threshold-based non-submodular model," IBM T. J. Watson Research Center, Yorktown, NY, USA, October 2017. [Online]. Available: https://resedit.watson.ibm.com/researcher/files/us-maliang/Ma17IMTR.pdf

[21] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," June 2014. [Online]. Available: http://snap.stanford.edu/data

[22] F. Chung and L. Lu, *Complex Graphs and Networks*. American Mathematical Society, 2006.