Analyzing competitive influence maximization problems with partial information: An approximation algorithmic framework

Yishi Lin *, John C.S. Lui

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

ABSTRACT

Given the popularity of the viral marketing campaign in online social networks, finding a computationally efficient method to identify a set of most influential nodes so as to compete well with others is of the utmost importance. In this paper, we propose a general model to describe the influence propagation of multiple competing sources in the same network. We formulate the *Competitive Influence Maximization with Partial information* (CIMP) problem: given an influence propagation model and the probability of a node being in the competitor's seed set, how to find a set of k seeds so to trigger the largest expected influence cascade under the presence of other competitors? We propose a general algorithmic framework, *Two-phase Competitive Influence Maximization* (TCIM), to address the CIMP problem. TCIM returns a $(1 - 1/e - \epsilon)$-approximate solution with probability of at least $1 - n^{-\ell}$, where $\ell \geq 1/2$ is a parameter controlling the trade-off between the success probability and the computational efficiency. TCIM has an efficient expected time complexity of $O(c(k+\ell)(m+n) \log n/\epsilon^2)$, where n and m are the number of nodes and edges in the network, and c is a function of the given propagation model (which may depend on k and the underlying network). To the best of our knowledge, this is the first work which provides a general framework for the competitive influence maximization problem where the seeds of the competitor could be given as an *explicit* set of seed nodes or a *probability distribution* of seed nodes. Moreover, our algorithmic framework provides both quality guarantee of solution and practical computational efficiency. We conduct extensive experiments on real-world datasets under three specific influence propagation models, and show the efficiency and accuracy of our framework. In particular, for the case where the seed set of the competitor is given explicitly, we achieve up to four orders of magnitude speedup as compared to previous algorithms with the same quality guarantee. When the competitor's seed set is not given explicitly, running TCIM using the *probability distribution* of the competitor's seeds returns nodes with higher expected influence than those nodes returned by TCIM using an *explicit guess* of the competitor's seeds.

1. Introduction

With the popularity of online social networks (OSNs), viral marketing has become a powerful method for companies to promote sales. In 2003, Kempe et al. [1] formulated the influence maximization problem: given a network G and an integer

---

* Corresponding author.
 E-mail addresses: yslin@cse.cuhk.edu.hk (Y. Lin), cslui@cse.cuhk.edu.hk (J.C.S. Lui).

k, how to select a set of *k* nodes in *G* so that they can trigger the largest influence cascade under a predefined influence propagation model. The selected nodes are often referred to as *seed nodes*. Kempe et al. proposed the *Independent Cascade* (IC) model and the *Linear Threshold* (LT) model to describe the influence propagation process. They proved that the influence maximization problem under these two models is NP-hard and a natural greedy algorithm could return $(1 − 1/e − \epsilon)$-approximate solutions for any $\epsilon > 0$. Recently, Tang et al. [2] presented an algorithm with a high probability of finding an approximate solution, and at the same time, with a low computational overhead.

Recognizing that companies are competing in a viral marketing, a thread of work studied the competitive influence maximization problem under a series of competitive influence propagation models, where multiple sources spread the information in a network simultaneously (e.g., [3–5]). Many of these work assumed that there are two companies competing with each other and studied the problem from the "follower's perspective". For example, in the viral marketing, a company introducing a new product into an existing market can be regarded as the follower, and consumers of the existing products can be treated as nodes influenced by this company's competitor. Formally, the problem of *Competitive Influence Maximization* (CIM) is defined as follows: suppose we are given a network *G* and the set of *seed nodes* selected by our competitor, how to select *seed nodes* for our product so as to trigger the largest influence cascade? In this work, we propose a more general problem, the *Competitive Influence Maximization with Partial information* (CIMP) problem, where we are given the probability of each node being in the competitor's seed set. We refer to this given probability as the *seed distribution* of the competitor. The assumption of knowing the *seed distribution* in the CIMP problem is much milder than the assumption in the CIM problem where one needs to know the explicit set of seed nodes of the competitor. Note that the CIM problem is essentially a special case of our CIMP problem.

Contributions: We believe that there can be many influence propagation models representing different viral marketing scenarios. However, existing works are usually restricted to specific propagation models (e.g., [4,6]). In this work, we propose a *general framework* that can solve the competitive influence maximization problem under a variety of propagation models. Our contributions are:

- We present a *General Competitive Independent Cascade* (GCIC) model which can accommodate different influence propagation models, and apply it to the *Competitive Influence Maximization* (CIM) problem. We proceed to formulate a more general *Competitive Influence Maximization with Partial information* (CIMP) problem, which generalizes the CIM problem as a special case. Both the CIM and CIMP problems are NP-hard.
- For the CIMP problem under the GCIC model, we propose a *Two-phase Competitive Influence Maximization* (TCIM) algorithmic framework. With probability of at least $1 − n^{−\ell}$, TCIM guarantees a $(1 − 1/e − \epsilon)$-approximate solution. It runs in $O(c(\ell + k)(m + n) \log n / \epsilon^2)$ expected time, where $n = |V|$, $m = |E|$, $\ell \geq 1/2$ is a quality control knob, and *c* depends on the specific propagation model, seed-set size *k* and the network $G = (V, E)$. To the best of our knowledge, this is the first general algorithm with *both* $(1 − 1/e − \epsilon)$ approximation guarantee and practical efficiency. Moreover, as we will explain later, the $(1 − 1/e − \epsilon)$ approximation guarantee is in fact the best guarantee one could obtain in polynomial time.
- To demonstrate the generality, accuracy and efficiency of our framework, we analyze the performance of TCIM using three popular influence propagation models: *Campaign-Oblivious Independent Cascade model* [6], *Distance-based model* [4] and *Wave propagation model* [4].
- We conduct extensive experiments using real datasets to demonstrate the efficiency and effectiveness of TCIM. For the CIM problem, when $k = 50$, $\epsilon = 0.5$ and $\ell = 1$, TCIM returns solutions comparable with those returned by baseline algorithms with the same quality guarantee, but runs *up to four orders of magnitude faster*. We also illustrate via experiments the benefits of taking the "seed distribution" of the competitor as input to our TCIM framework.

The outline of our paper is as follows. Background and related work are given in Section 2. We define the *General Competitive Independent Cascade* model, the *Competitive Influence Maximization* problem and the *Competitive Influence Maximization problem with Partial information* in Section 3. We present the TCIM framework in Section 4 and analyze its performance under various influence propagation models in Section 5. We compare TCIM with the greedy algorithm with performance guarantee in Section 6, and show experimental results in Section 7. Section 8 concludes.

2. Background and related work

Single source influence maximization. In the seminal work [1], Kempe et al. proposed the *Independent Cascade* (IC) and the *Linear-Threshold* (LT) propagation model and formally defined the *influence maximization problem*. In the IC model, a network $G = (V, E)$ is given and each edge $e_{uv} \in E$ is associated with a probability p_{uv}. Initially, a set of nodes *S* is *active* and *S* is referred to as the *seed nodes*. Each active node *u* has a single chance to influence its inactive neighbor *v* and succeeds with probability p_{uv}. Let $\sigma(S)$ be the expected number of nodes in *G* that *S* could activate, the influence maximization problem is defined as how to select a set of *k* nodes such that $\sigma(S)$ is maximized. Kempe et al. showed that this problem is NP-hard under both proposed models. Moreover, they showed that $\sigma(S)$ is a monotone and submodular function of *S* under both models. Therefore, for any $\epsilon > 0$, a greedy algorithm returns a solution whose expected influence is at least $(1 − 1/e − \epsilon)$ times the expected influence of the optimal solution. The research on this problem went on for around ten years (e.g., [7–12]). Recently, Borgs et al. [13] made a breakthrough and presented an algorithm that simultaneously maintains the performance guarantee and significantly reduces the time complexity. Tang et al. [2] further improved the method in [13] and presented