Online Influence Maximization in Non-Stationary Social Networks

Yixin Bao*, Xiaoke Wang*, Zhi Wang†, Chuan Wu*, Francis C.M. Lau*
*Department of Computer Science, The University of Hong Kong, Email: {yxbao,xkwang,cwu,fcmlau}@cs.hku.hk
†Graduate School at Shenzhen, Tsinghua University, Email: wangzhi@sz.tsinghua.edu.cn

***Abstract*—Social networks have been popular platforms for information propagation. An important use case is viral marketing: given a promotion budget, an advertiser can choose some influential users as the seed set and provide them free or discounted sample products; in this way, the advertiser hopes to increase the popularity of the product in the users' friend circles by the world-of-mouth effect, and thus maximizes the number of users that information of the production can reach. There has been a body of literature studying the influence maximization problem. Nevertheless, the existing studies mostly investigate the problem on a one-off basis, assuming fixed known influence probabilities among users, or the knowledge of the exact social network topology. In practice, the social network topology and the influence probabilities are typically unknown to the advertiser, which can be varying over time, i.e., in cases of newly established, strengthened or weakened social ties. In this paper, we focus on a dynamic non-stationary social network and design a randomized algorithm, RSB, based on multi-armed bandit optimization, to maximize influence propagation over time. The algorithm produces a sequence of online decisions and calibrates its explore-exploit strategy utilizing outcomes of previous decisions. It is rigorously proven to achieve an upper-bounded regret in reward and applicable to large-scale social networks. Practical effectiveness of the algorithm is evaluated using real-world datasets, which demonstrates that our algorithm outperforms previous stationary methods under non-stationary conditions.**

I. Introduction

Influence maximization in social networks is an important problem that seeks the best seed users to maximize the spread of information [1]. Prominent use cases include advertising and viral marketing [1]. When a company is promoting a new product, it can engage some influential users as seeds in a social network, providing them samples for free or at discounted prices. These seed users may inform their friends of this product, and their friends will further influence other users, and so on. Through world-of-mouth distribution, the product will get to be known by more and more users in the social network. As it is common for a company to have a promotion budget, it is most beneficial to identify the best set of seeds so as to maximize the number of users that information can eventually reach.

Existing studies mostly tackle the influence maximization problem on a one-off basis, assuming that both the social network topology and influence probabilities are fixed and available as input [1]. In real-world social networks, exact network topology and influence probabilities are typically unknown to a third party advertiser, and are time-varying. For example, new social ties are set up when people make new friends, and the ties can be strengthened over time when they become more familiar; two people become connected when collaborating on a short-term project and the tie may weaken after the project has ended; a couple may break up and be no longer connected in the social network. It is therefore more realistic to describe the influence probabilities and social network topology as non-stationary. In addition, it is often hard to determine an accurate stochastic distribution assumption for the variance of influence probabilities, since no assumption may exist for human behavior.

To handle unknown underlying distributions in online optimization, multi-armed bandit optimization has been applied in related scenarios. The multi-armed bandit problem [2] is a problem in which an agent has multiple arms to choose from, and needs to decide a policy to select an arm at each time. When chosen, an arm provides a random reward from an unknown distribution specific to the arm, and the agent utilizes the outcome to update his strategy. The objective is to maximize the overall reward in the whole time span through selecting a sequence of arms, thus minimizing regret, which is the gap between offline optimal overall reward and the actual overall reward the agent has obtained. The design of multi-armed bandit algorithms mainly focuses on how to handle the trade-off between exploration and exploitation [2], *i.e.*, to try the arm that has not been attempted before (*exploration*) or the arm that has brought high reward so far (*exploitation*).

This paper designs an online randomized algorithm, referred to as *RSB*, based on multi-armed bandit optimization, to maximize influence propagation in a dynamic non-stationary social network with unknown and non-stationary influence probabilities between pairs of users. Our algorithm design does not assume knowledge of the social graph and the influence probability distributions, nor requires any initialization stage. Regardless of the concrete influence probabilities or the topology of the social network, an $O(\sqrt{TN \ln N})$ regret bound is rigorously proven where T is the number of time stages in the entire system span and N is the number of nodes. To the best of our knowledge, this is the first influence maximization algorithm dealing with both unknown and non-stationary influence probabilities.

The rest of the paper is organized as follows. We discuss

The project was supported in part by grants from Hong Kong RGC under the contracts HKU 717812E, C7036-15G (CRF), and the National Natural Science Foundation of China (NSFC) under Grant No. 61402247

related work in Sec. II and present the problem model in Sec. III. In Sec. IV and Sec. V, we present the detailed online algorithm and provide theoretical analysis of its regret bound. Simulation results are presented in Sec. VI. We conclude the paper in Sec. VII.

II. Related Work

Recently, multi-armed bandit optimization has been applied to solve the influence maximization problem with incomplete information of the social network. The existing algorithms have been relying on assumptions of the rewards to guarantee nice theoretical bounds on regret. In particular, combinatorial bandits are highly related to the influence maximization problem, where the decision-making agent needs to select multiple arms in each time stage. Chen *et al.* [3] define a general bandit framework to deal with both linear and non-linear reward functions. One application of their framework is social influence maximization with unknown influence probabilities. However, the reward of each arm must be an i.i.d random process, *i.e.*, the reward distribution is stationary over time. Lei *et al.* [4] present an online influence maximization framework utilizing exploration-exploitation for seed selection and strategy updating. They assume that all connections in the social network are known and only evaluate the performance of their algorithm experimentally without theoretical analysis. Lin *et al.* [5] develop a more general bandit optimization algorithm for a class of problems using greedy methods, guaranteeing a good approximation ratio. However, they only consider stochastic rewards with stationary distributions.

The simplest idea to tackle non-stationary rewards is to decrease the weights of earlier feedback in next-step decision making [6]. The problem it may lead to is that without sufficient feedback information, it is hard to achieve a good accuracy of reward estimation. Some algorithm designs assume abrupt changes of the distributions occurring at arbitrary intervals [7], and allow the agent to query a set of arms not picked before and obtain outcomes as if these arms were played. This assumption is reasonable in a stock market, *i.e.*, people can acquire information of stocks they have never purchased by following bearish or bullish trends, but not for influence propagation, where there is no channel to obtain outcomes of untried arms. Besbes *et al.* [8] assume that the total variation of the rewards is given and design a randomized algorithm based on Exp3 [9]. Only one arm is selected in each time stage, while we focus on the case of combinatorial bandits. Gai *et al.* [10] investigate restless bandits with Markov rewards, where the states of each arm evolve as an irreducible finite-state Markov process over time. The algorithm relies on an initialization stage, in which each arm is tried for at least once. This is impractical for influence propagation (*e.g.*, market campaign) in a large-scale network, as the cost of trying all nodes is unaffordable.

III. Problem Model

We model the social network as an influence graph $G = (\mathcal{N}, \mathcal{E})$. $\mathcal{N} = \{1, 2, \ldots, N\}$ is the set of users (nodes), where N is the total number of nodes. $\mathcal{E}$ is the set of social connections among the nodes. An unknown influence probability $p^t_{n,m}$ is associated with each edge $(n, m) \in \mathcal{E}$, which is time varying following an unknown, non-stationary distribution: after user n is activated (*e.g.*, obtained information of a product), he may activate his neighbor m (*e.g.*, share information of the product) with different probabilities at different time stages t. In this way, each edge (n, m) is associated with a non-stationary Bernoulli distribution: in t, user n may activate his neighbor m with probability $p^t_{n,m}$, or not with probability $1 - p^t_{n,m}$. We do not assume any cascade model of the information propagation system (*e.g.*, independent cascade model [1] or linear threshold model [1]), and our algorithm works with various cascade models as long as the information spread brought by an activate node can be modeled as a random variable.

Let T be the total number of time stages that the system spans. In each time stage, a set of K seeds are selected as information sources (*e.g.*, the seed users that an advertiser directly promotes the product to, whose number is decided by the promotion budget), from which the information spreads to other nodes in the network. The seed set is repeatedly selected over different time stages. For example, a company may carry out a promotion campaign for a series of time stages, *e.g.*, a number of consecutive days. After the promotion in each time stage via a potentially different set of seeds, the company collects statistics on the number of purchases of their promoted product(s) and utilizes this feedback to update its seed selection strategies in later time stages. The goal is to maximize the expected overall influence spread in the whole time span $1, 2, \ldots, T$, *i.e.*, the expected total number of activated nodes. Let $\mathcal{M}$ be the collection of all subsets of $\mathcal{N}$. In our bandit optimization framework, we define $a|S$, meaning node a under a given set $S \in \mathcal{M}$, as an *arm*. The expected *reward* of selecting an arm $a|S$ is the expected marginal gain by adding a into the existing seed set S, *i.e.*, the expected additional number of activated users after we add a into S. Let $f_t(S)$ be an influence spread function in time stage t, indicating the total number of activated nodes in t based on seed set S. The value of $f_t(S)$ is a random variable. The expectation $\mathbb{E}[f_t(S)]$ is non-negative, monotone and submodular, as proven in [11]. The submodularity of the spread function is useful such that we can utilize the benchmark based on greedy optimal value. The expected reward of selecting an arm $a|S$ in t is hence $\mathbb{E}[f_t(S \cup \{a\})] - \mathbb{E}[f_t(S)]$. Note that the expectation $\mathbb{E}[\cdot]$ is taken over both randomized rewards and randomized policies, where a *policy* refers to the agent's strategy for seed selection, which is random given the random nature of our algorithm.[1]

In each time stage t, starting from an empty set $S_t = \emptyset$, we obtain a seed set of size K by adding nodes to S_t one by one in some order. Let $S_t = (a^1_t, \ldots, a^K_t)$ be the completed seed set, in which the k^{th} element is the k^{th} seed selected in this time stage. Let $S^{(1:k-1)}_t$ represent the selected seed set with

[1] Although [11] does not consider randomized policies, the submodularity of $\mathbb{E}[f_t(S)]$ still holds following results in [11], as expectation over policies is a linear combination of submodular functions.

elements $1,\dots,k-1$, and $a_t^k|S_t^{(1:k-1)}$ mean that node a_t^k is selected as the k^{th} seed in t given previous choices in $S_t^{(1:k-1)}$. Let $\bar{r}_t^k(a_t^k|S_t^{(1:k-1)}) = \mathbb{E}[f_t(S_t^{(1:k-1)} \cup \{a\})] - \mathbb{E}[f_t(S_t^{(1:k-1)})]$ denote the expected marginal gain of choosing a_t^k as the k^{th} seed in t. The expected total reward in time stage t is $\bar{r}_t(S_t) = \sum_{k=1}^{K} \bar{r}_t^k(a_t^k|S_t^{(1:k-1)}) = \mathbb{E}[f_t(S_t)]$.

In this model, maximizing the expected total number of activated nodes in $1,\dots,T$ is equivalent to maximizing the expected overall reward in the entire span, $\sum_{t=1}^{T} \bar{r}_t(S_t) = \sum_{t=1}^{T} \mathbb{E}[f_t(S_t)]$. It is further equivalent to minimizing the regret, the gap between the expected overall reward that the agent can obtain by running our online algorithm and the offline optimal expected overall reward computed using full knowledge of the system. In our algorithm design, we aim to minimize the weak regret, *i.e.*, the gap between the expected overall reward achieved by our algorithm and the offline expected overall reward achieved by using the same best seed set S^* in all time stages, namely $S^* \in \arg\max_{S\in\mathcal{M}} \sum_{t=1}^{T} \mathbb{E}[f_t(S)]$, computed based on full knowledge of the entire system. Such a weak regret is the difference between the expected overall reward obtained by our algorithm and that achieved by the best single action, *i.e.*, sticking with one seed set in all time stages. Weak regret is commonly used in the literature on analysing non-stationary bandit algorithms [10], and the key ingredient is to form accurate estimates on the average condition for each arm [12], so as to find the arm performing best in a long term. In particular, we analyze a greedy weak regret, with detailed definition given in Definition 2 in Sec. V, that compares the expected overall reward produced by our algorithm with the lower bound of an approximate offline overall reward achieved by a single best seed set derived by a greedy approach. Greedy weak regret is a concept narrowed down from weak regret, when the best single action is decided by a greedy algorithm. We apply this notion so as to compare with the lower bound of the greedy optimal value.

IV. RSB: Randomized Multi-armed Bandit Algorithm for Non-Stationary Social Networks

Main Idea. We next design an online multi-armed bandit algorithm to minimize the greedy weak regret. In each time stage, we select the best seed set by sequentially selecting the next best node given previous seed decisions. Given the set of already selected seeds, we associate weights with candidate arms, and deal with the varying environment (time-varying underlying distributions of influence probabilities) by adjusting the weights of arms based on rewards received due to previous seed selection (the *exploitation* component of our algorithm). Besides, we also include a constant $\frac{\gamma}{N}$ in the weight of each arm, where $\gamma \in (0,1]$ is a gaugeable value, in order to enable *exploration* of arms never tried before. Different from deterministic stationary bandit algorithms, our algorithm is randomized in arm selection according to the weights, and hence even if the environment changes abruptly, the algorithm still has a chance to switch to the new best arm.

Algorithm Steps. Our multi-armed bandit algorithm for selecting the best seed set in each time stage t is given in Alg. 1. Here $w_t^{n|S_t^{(1:k-1)}}$ is the weight for selecting node n as the k^{th} seed in time stage t, while the set of already selected seeds in t is $S_t^{(1:k-1)}$. $v_t^{n|S_t^{(1:k-1)}}$ is an auxiliary quantity to compute the weights, updated based on the past reward information of arm $n|S_t^{(1:k-1)}$, as an exploration measure. $q_t^{n|S_t^{(1:k-1)}}$ is the probability of playing arm $n|S_t^{(1:k-1)}$ in t, derived from the weights of the arms. $r_t^k(a|S_t^{(1:k-1)})$ denotes the realization of the reward (actual marginal influence spread) by choosing node a as the k^{th} seed in t. C is an input parameter to the algorithm, which satisfies $C \geq \frac{\gamma r_t^k(n|S)}{N q_t^k(n|S)}, \forall n \in \mathcal{N},\ S \in \mathcal{M}$.

In Alg. 1, the K seeds are selected sequentially (line 3). The weights $\mathbf{w}$ associated with the nodes should be equal at the beginning of each time stage, and adjusted based on updated $\mathbf{v}$, each time after the seed set has been updated (lines 4-6). The computation of $w_t^{n|S_t^{(1:k-1)}}$ aims to balance exploitation and exploration: the first term is calculated based on past reward information (*exploitation*) and the second constant term is assigned for each arm no matter how many times it has been tried (*exploration*). Next, the probability for adding an additional node into the already selected set of seeds is decided by normalizing its weight over the weights of all the remaining nodes not in the existing seed set (lines 7-9). An arm is randomly selected according to the probability distribution and a reward $a|S_t^{(1:k-1)}$ is observed (lines 10-12), *e.g.*, the additional number of product purchases received by promoting to node a is collected. We then update $v_t^{n|S_t^{(1:k-1)}}$ by multiplying an exponential factor (line 16), decided by $\hat{r}_t^k(a|S_t^{(1:k-1)})$, which can be understood as an unbiased estimation of the reward of the arm. Computing $\hat{r}_t^k(a|S_t^{(1:k-1)})$ by dividing the actual reward by the probability of selecting the arm (line 13) compensates the reward of actions with less probability to be chosen and guarantees that the expectation of the estimated reward and the actual reward are equal, when the expectation is taken over both randomized policies and randomized rewards. This equality helps us to derive the expected reward of RSB in the proof. The updated weights will be used in selecting future seeds in this time stage.

We will evaluate the impact of the input parameter γ under practical settings in simulations. The input parameter C is related to the largest spread brought by a seed, which is unknown before running the algorithm. In fact, requiring $C \geq \frac{\gamma r_t^k(n|S)}{N q_t^k(n|S)}$ is only needed for regret analysis. We can set the value of C empirically when running the algorithm in practice, and will evaluate the performance of the algorithm under an empirical value of C in simulations, which does not necessarily satisfy the above condition.

We note that our algorithm does not rely on any knowledge of the underlying social network topology and the influence probabilities, but only utilizes the outcomes that are decided by them. In addition, although the entire space of arms, $a|S, \forall a \in \mathcal{N},\ S \in \mathcal{M}$, is exponential, the number of arms that need to be dealt with in each time stage in Alg. 1 (weights

Algorithm 1 RSB: Randomized Sequential Multi-armed Bandit Algorithm for Non-Stationary Networks

Input: $\mathcal{N}$, K, C, γ

Output: the seed set $S_t^{(1:K)}$ for each time stage t

set $v_1^{n|S_1^{(1:k-1)}} = 1$, $\forall n \in \mathcal{N}$, $k = 1, \ldots, K$
for $t = 1, 2, \ldots, T$ **do**
  for $k = 1, 2, \ldots, K$ **do**
    for each node $n \in \mathcal{N}$ **do**
      set $w_t^{n|S_t^{(1:k-1)}} = (1-\gamma)\frac{v_t^{n|S_t^{(1:k-1)}}}{\sum_{n' \in \mathcal{N}} v_t^{n'|S_t^{(1:k-1)}}} + \frac{\gamma}{N}$
    end for
    for each node $n \in \mathcal{N} \backslash S_t^{(1:k-1)}$ **do**
      $q_t^{n|S_t^{(1:k-1)}} = \frac{w_t^{n|S_t^{(1:k-1)}}}{\sum_{n' \in \mathcal{N} \backslash S_t^{(1:k-1)}} w_t^{n'|S_t^{(1:k-1)}}}$
    end for
    draw an arm $a|S_t^{(1:k-1)}$ according to the distribution $\{q_t^{n|S_t^{(1:k-1)}}\}_{n \in \mathcal{N} \backslash S_t^{(1:k-1)}}$
    receive a reward $r_t^k(a|S_t^{(1:k-1)})$
    set $S_t^{(1:k)} = S_t^{(1:k-1)} \cup \{a\}$
    set $\hat{r}_t^k(a|S_t^{(1:k-1)}) = \frac{r_t^k(a|S_t^{(1:k-1)})}{q_t^{a|S_t^{(1:k-1)}}}$
    for all $n \in \mathcal{N} \backslash \{a\}$, set $\hat{r}_t^k(n|S_t^{(1:k-1)}) = 0$
    for each arm $n|S_t^{(1:k-1)}$, $\forall n \in \mathcal{N}$ **do**
      update $v_{t+1}^{n|S_{t+1}^{(1:k-1)}} = v_t^{n|S_t^{(1:k-1)}} \exp\{\frac{\gamma \hat{r}_t^k(n|S_t^{(1:k-1)})}{NC}\}$
    end for
  end for
end for

and probabilities computed and used in seed selection) is still polynomial, as given in the following theorem. The proof can be found in our technical report [13].

Theorem 1. *The time complexity of Alg. 1, executed in each time stage t, is $O(KN)$.*

V. Regret Analysis

We next analyze an upper bound of the greedy weak regret achieved by Alg. 1. Let OPT denote the offline maximal value of the expected overall reward $\sum_{t=1}^{T} \bar{r}_t(S) = \sum_{t=1}^{T} \mathbb{E}[f_t(S)]$ over all $S \in \mathcal{M}$, computed based on complete knowledge of the influence probability distributions and the social graph topologies in $1, \ldots, T$. Let S^* be the offline optimal seed set, *i.e.*, the single best seed set that maximizes $\sum_{t=1}^{T} \bar{r}_t(S)$.

We define a *position* optimal reward OPT^k as the sum of the expected marginal gains achieved by using the best k^{th} seed in all time stages. The best k^{th} seed maximizes $\sum_{t=1}^{T} \bar{r}_t^k(a|S_t^{(1:k-1)})$ based on full knowledge of the system, given the first $k-1$ seeds in $S_t^{(1:k-1)}$ in each t derived using RSB. Based on [14], the idea is to reduce the original problem of finding the best solution of the full set to a parallel bandit setting, finding the best k^{th} element under the condition determined by our algorithm. Let $\tilde{a}^k$ denote this optimal k^{th} seed, *i.e.*, $\tilde{a}^k \in \arg\max_{a \in \mathcal{N}} \sum_{t=1}^{T} \bar{r}_t^k(a|S_t^{(1:k-1)})$. Such a best k^{th} seed may form different arms, $\tilde{a}^k|S_t^{(1:k-1)}$, under different seed sets $S_t^{(1:k-1)}$ in different time stages. We have $OPT^k = \max_{a \in \mathcal{N}} \sum_{t=1}^{T} \bar{r}_t^k(a|S_t^{(1:k-1)}) = \sum_{t=1}^{T} \bar{r}_t^k(\tilde{a}^k|S_t^{(1:k-1)})$.

Definition 1. *The position weak regret for the k^{th} seed is*

$$Reg^k(T) = \sum_{t=1}^{T} \bar{r}_t^k(\tilde{a}^k|S_t^{(1:k-1)}) - \sum_{t=1}^{T} \bar{r}_t^k(a_t^k|S_t^{(1:k-1)})$$

where $\tilde{a}^k \in \arg\max_{a \in \mathcal{N}} \sum_{t=1}^{T} \bar{r}_t^k(a|S_t^{(1:k-1)})$ and $a_t^k|S_t^{(1:k-1)}$ is the arm selected by Alg. 1 in time stage t. The conditional set $S_t^{(1:k-1)}$ is also decided by Alg. 1.

The following theorem states the relationship between position weak regret and OPT, which will be used to bound the greedy weak regret in Theorem 3. All missing proofs in this section can be found in [13].

Theorem 2. *For any position $k = 1, 2, \ldots, K$, we have*

$$\sum_{t=1}^{T} \Big(\bar{r}_t(S_t^{(1:k)}) - \bar{r}_t(S_t^{(1:k-1)})\Big) \geq \frac{1}{K}\Big(OPT - \sum_{t=1}^{T} \bar{r}_t(S_t^{(1:k-1)})\Big) - Reg^k(T).$$

Let $F(S) = \sum_{t=1}^{T} \mathbb{E}[f_t(S)]$, $\forall S \in \mathcal{M}$, which denotes the expected overall influence spread over the whole system span. $F(S)$ is a submodular function since it is the summation of submodular functions $\mathbb{E}[f_t(S)], \forall t = 1, \ldots, T$. Then we can design a greedy approach to compute a S that approximately maximizes the expected overall reward $\sum_{t=1}^{T} \bar{r}_t(S) = \sum_{t=1}^{T} \mathbb{E}[f_t(S)]$ based on full knowledge of the system: after deciding $S^{(1:k-1)}$, we select a local optimal node as the k^{th} seed, that maximizes the expected marginal influence spread, *i.e.*, node u such that $u \in \arg\max_{v \in \mathcal{N} \backslash S^{(1:k-1)}} \{F(S^{(1:k-1)} \cup \{v\}) - F(S^{(1:k-1)})\}$. We can easily prove that the approximate offline solution computed this way achieves an approximation ratio of $1 - \frac{1}{e}$, *i.e.*, the expected overall reward it achieves is at least $(1 - \frac{1}{e})OPT$, following Theorem 3.5 in [11], based on submodularity of the spread function and local optimality when selecting each seed. The reason that we compute this approximate offline solution using the greedy approach (which runs in polynomial time) is that computing S^* has been shown to be an NP hard problem [1].

Using the approximate offline overall reward computed as above, we define a greedy weak regret as follows, which we use to evaluate the performance of our algorithm RSB.

Definition 2. *The greedy weak regret is defined as the gap between the lower bound of the approximate offline overall* *reward derived by the greedy approach and the expected overall reward produced by RSB in Alg. 1,* i.e.,

$$Reg_G(T) = (1 - \frac{1}{e})OPT - \sum_{t=1}^{T} \bar{r}_t(S_t^{(1:K)}).$$

The following theorem shows that the overall position weak regret provides an upper bound of the greedy weak regret.

Theorem 3. *The greedy weak regret is upper bounded by the sum of position weak regrets over all positions* $k = 1, 2, \ldots, K$*,* i.e.*,*

$$Reg_G(T) \le \sum_{k=1}^{K} Reg^k(T).$$

Based on Theorem 3, we seek to bound the position weak regret for each k, in order to derive an upper bound of $Reg_G(T)$. According to Definition 1, the position weak regret for the k^{th} seed is

$$\begin{aligned} Reg^k(T) &= \sum_{t=1}^{T} r_t^k(\tilde{a}^k | S_t^{(1:k-1)}) - \sum_{t=1}^{T} r_t^k(a_t^k | S_t^{(1:k-1)}) \\ &= \max_{n \in \mathcal{N}} \sum_{t=1}^{T} r_t^k(n | S_t^{(1:k-1)}) - \sum_{t=1}^{T} r_t^k(a_t^k | S_t^{(1:k-1)}). \end{aligned}$$

Let D be the upper bound of the realization of reward, *i.e.*, $r_t^k(n|S) \le D,\ \forall n \in \mathcal{N},\ S \in \mathcal{M}$. The following theorem states an upper bound of the position weak regret for each k. In particular, if γ is set to a special value, it can minimize the regret bound.

Theorem 4. *Let* $R_{\max}^k = \max_{n \in \mathcal{N}} \sum_{t=1}^{T} \bar{r}_t^k(n|S_t^{(1:k-1)})$ *be the expected overall reward achieved by selecting the best* k^{th} *arm given* $S_t^{(1:k-1)}, \forall t = 1, \ldots, T$*, derived by Alg. 1. Let* $R_{RSB}^k = \sum_{t=1}^{T} \mathbb{E}[r_t^k(a_t^k|S_t^{(1:k-1)})]$ *denote the expected overall marginal gain obtained by adding the* k^{th} *seeds into the given* $S_t^{(1:k-1)}, \forall t = 1, \ldots, T$*. For any parameter* $\gamma \in (0, 1]$*, we have*

$$\begin{aligned} Reg^k(T) &= R_{\max}^k - R_{RSB}^k \\ &\le (1 + (e-2)\frac{D}{C})\gamma R_{\max}^k + \frac{NC \ln N}{\gamma}. \end{aligned}$$

If we set $\gamma = \min\{1, \sqrt{\frac{NC \ln N}{(1+(e-2)\frac{D}{C})g}}\}$ *where constant* $g \ge R_{\max}^k,\ \forall k = 1, \ldots, K$*, we have the following minimum upper bound*

$$\sum_{k=1}^{K} Reg^k(T) \le 2K \sqrt{1 + (e-2)\frac{D}{C}} \sqrt{gCN \ln N}.$$

Corollary 1. *The greedy weak regret achieved by Alg. 1 is upper bounded as follows:*

$$Reg_G(T) \le 2K \sqrt{1 + (e-2)\frac{D}{C}} \sqrt{DCTN \ln N},$$

i.e.*, the upper bound of the greedy weak regret of Alg. 1 is* $O(\sqrt{TN \ln N})$*.*

It shows that our greedy weak regret is sublinear with both N and T.

VI. Performance Evaluation

A. Data Sets and Time-varying Influence Probabilities

We produce a dynamic social network based on Tencent Weibo[2] traces containing the *following* links among 4257 users for 7 consecutive days during November 2011. Each directed *following* link (n, m) indicates that user n follows user m [15]. The links among the users vary from one day to the next, giving a dynamic social graph. To prolong the trace duration, we further repeat the variation of the social graph on 7-day cycles to form a 100-day duration (T), which we believe reasonable since human behavior may well follow a weakly periodicity.

We employ the following three models to generate nonuniform and time-varying influence probabilities in a social graph.

- The Trivalency (TR) model [16]: in each time stage, the influence probability of an edge in the social graph is assigned a value among $\{0.1,\ 0.01,\ 0.001\}$ uniformly randomly, corresponding to three types of social ties - strong, medium and weak. The assigned probability on an edge may change from one time stage to the next.
- A Fluctuating Reward (FR) model. We design this model such that influence probabilities evolve over time in a similar fashion as a sinusoidal wave (also similar to that used in [17]): the influence probability of each edge starts from a random value drawn uniformly within $[0, 0.1]$; then in each time stage, it increases or decreases at a constant rate $\frac{0.3}{T}$ until reaching the largest value 0.1 or the smallest value 0.

B. Schemes for Comparison and Evaluation Results

We compare RSB with a random algorithm and OG-UCB proposed in [5]. With the random algorithm, the agent always selects a seed uniformly randomly among all candidate nodes. OG-UCB is designed for stationary scenarios, which associates a confidence bound with each arm and chooses the arm with the highest upper confidence bound greedily.

We note that although a number of bandit algorithms have been proposed for influence maximization (as discussed in Sec. II), most are not directly comparable since they run with the complete knowledge of a social network. We compare with OG-UCB since it is the only existing bandit algorithm without requiring knowledge of the social graph topology. In addition, the bandit algorithms designed for non-stationary systems in Sec. II either deal with 1 arm or assume Markov rewards, and hence cannot be readily extended for comparison.

In computing greedy weak regret, we also compute the approximate offline optimal overall reward by the greedy offline algorithm discussed before Definition 2 in Sec. V.

To show greedy weak regret values in a unified range in our figures, we plot the ratio between greedy weak regret and the approximate offline optimal overall reward, *i.e.*, $\frac{\text{approx. offline opt. overall reward} - \text{overall reward by RSB}}{\text{approx. offline opt. overall reward}}$. Especially, a data point at a specific T represents the above

[2] http://t.qq.com/

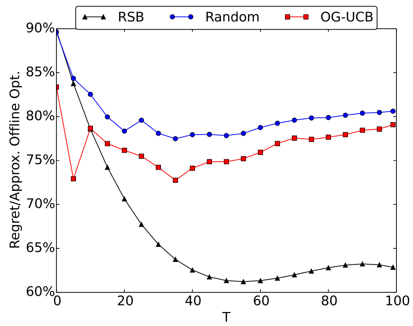

Fig. 1. Tencent Weibo trace and FR model: $\gamma = 0.2$.

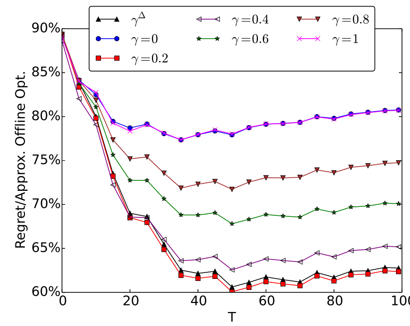

Fig. 2. Tencent Weibo trace and FR model: different values of γ.

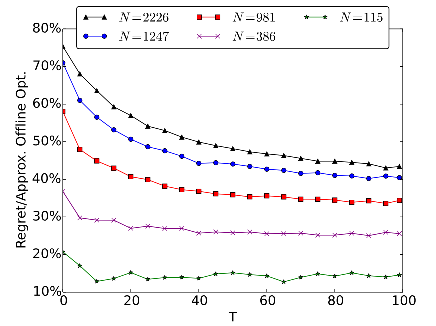

Fig. 3. Tencent Weibo trace and TR model: different graph sizes N.

ratio computed using overall rewards in $[1, T]$. We set $K = 5$, $\gamma = 0.2$, $D = 120$ and $C = 1$ in our experiments.

Fig. 1 shows the results obtained using Tencent Weibo traces under FR model. We observe that RSB gets better than the other algorithms (lower regret and hence better spread) after more time stages, validating that RSB can improve with more feedback received from the real system. Besides, OG-UCB performs worse than RSB especially with the ongoing of time, showing that it is only suitable for fixed influence probability distributions and does not work well in cases of time-varying influence probabilities. The increase of cumulative regret by RSB with the increase of time stages, if any, is always slower than that of the other algorithms. In Fig. 2, we compare the regret ratios of RSB achieved under different values of input parameter γ, using Tencent Weibo traces under the FR model. From line 5 of Alg. 1, we can see $\gamma = 0$ represents pure exploitation and $\gamma = 1$ indicates pure exploration. RSB performs worst in these extreme cases. $\gamma^{\Delta} = 0.18$ is computed following the formula in Theorem 4 which minimizes the theoretical upper bound. We observe that γ^{Δ} achieves near-lowest regrets in actual execution of our algorithm under practical settings as well. In Fig. 3, we evaluate the impact of different graph sizes N, by extracting subgraphs of different sizes using Tencent Weibo traces. We observe that the regret is larger in larger networks, but it always improves when the system runs for a longer period of time.

VII. Conclusion

This paper investigates online influence maximization in dynamic social networks with non-stationary influence probability distributions among participants. We design a randomized algorithm based on multi-armed bandit optimization to guide source selection for information dissemination over multiple time stages, aiming to maximize the overall spread over the system span. The algorithm is simple and neat, relying on carefully designed, continuously updating preferences on seed selection, which exploit real-world feedback from previous decisions, as well as explore new choices. As the first in the literature, the algorithm does not require knowledge of the dynamic social graph topology, nor time-varying influence probabilities, but is able to achieve an upper-bounded weak regret, as compared to an approximate offline optimal reward. Simulations based on real-world datasets further validate that our algorithm is more adaptive to a changing environment than heuristic and stationary bandit algorithms.

References

[1] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the Spread of Influence through a Social Network," in *Proc. of the 9th ACM SIGKDD*, 2003, pp. 137–146.

[2] S. Bubeck, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[3] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial Multi-Armed Bandit: General Framework, Results and Applications," in *Proc. of the 30th ICML*, 2013, pp. 151–159.

[4] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart, "Online Influence Maximization," in *Proc. of the 21th ACM SIGKDD*, 2015, pp. 645–654.

[5] T. Lin, J. Li, and W. Chen, "Stochastic Online Greedy Learning with Semi-bandit Feedbacks," in *NIPS*, 2015, pp. 352–360.

[6] A. Garivier and E. Moulines, "On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems," *arXiv preprint arXiv:0805.3415*, 2008.

[7] J. Y. Yu and S. Mannor, "Piecewise-stationary Bandit Problems with Side Observations," in *Proc. of the 26th ICML*, 2009, pp. 1177–1184.

[8] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards," in *NIPS*, 2014, pp. 199–207.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Non-stochastic Multi-armed Bandit Problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2003.

[10] Y. Gai, B. Krishnamachari, and M. Liu, "Online Learning for Combinatorial Network Optimization with Restless Markovian Rewards," in *Proc. of the 9th IEEE SECON*, 2012, pp. 28–36.

[11] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and Influence Propagation in Social Networks," *Synthesis Lectures on Data Management*, vol. 5, no. 4, pp. 1–177, 2013.

[12] Y. Liu and M. Liu, "An Online Approach to Dynamic Channel Access and Transmission Scheduling," in *Proc. of the 16th ACM MobiHoc*, 2015, pp. 187–196.

[13] Y. Bao, X. Wang, Z. Wang, C. Wu, and F. C. Lau, "Online Influence Maximization in Non-Stationary Social Networks," *arXiv preprint arXiv:1604.07638*, 2016.

[14] Y. Yue and C. Guestrin, "Linear Submodular Bandits and their Application to Diversified Retrieval," in *NIPS*, 2011, pp. 2483–2491.

[15] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a Network Highway for Big Data: Architecture and Challenges," *Network, IEEE*, vol. 28, no. 4, pp. 5–13, 2014.

[16] W. Chen, C. Wang, and Y. Wang, "Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks," in *Proc. of the 16th ACM SIGKDD*, 2010, pp. 1029–1038.

[17] O. Besbes, Y. Gur, and A. Zeevi, "Optimal Exploration-Exploitation in a Multi-Armed-Bandit Problem with Non-stationary Rewards," *arXiv preprint arXiv:1405.3316*, 2014.