Submodular Optimization beyond Nonnegativity: Adaptive Seed Selection in Incentivized Social Advertising

Shaojie Tang[1], **Jing Yuan** [2]

[1]Naveen Jindal School of Management, University of Texas at Dallas
[2] Department of Computer Science and Engineering, University of North Texas
shaojie.tang@utdallas.edu, jing.yuan@unt.edu

Abstract

Social advertising, also known as social promotion, is a method of promoting products or ideas through the use of influential individuals, known as "seeds," on online social networks. Advertisers and platforms are the main players in this ecosystem, with platforms selling viral engagements, such as "likes," to advertisers by inserting ads into the feeds of seeds. Seeds are given monetary incentives by the platform in exchange for their participation in the campaign, and when a follower of a seed engages with an ad, the platform receives payment from the advertiser. Specifically, at the beginning of a campaign, the advertiser submits a budget to the platform and this budget can be used for two purposes: recruiting seeds and paying for the viral engagements generated by the seeds. Note that the first part of payment goes to the seeds and the latter one is the actual revenue collected by the platform. The challenge for the platform is to select a group of seeds that will generate the most revenue within the budget constraints set by the advertiser. This problem is challenging as the objective function can be non-monotone and may take on negative values. This makes traditional methods of submodular optimization and influence maximization inapplicable. We study this problem under both non-adaptive and adaptive settings, and propose effective solutions for each scenario.

Introduction

Social advertising (or social promotion) has been proved to be an effective approach that can produce a significant cascade of adoptions through word-of-mouth effect. It has been shown that social advertising is more effective than conventional advertising channels, including both demographically targeted and untargeted ads (Bakshy et al. 2012; Tucker 2012). Social advertising is often implemented as *promoted posts* that are displayed in the news feeds of their online users. Under the price per engagement (PPE) pricing model, the advertiser pays the platform for all users engaged with their ad. Examples of engagement include "like", "share", or "comment". One unique feature of promoted posts, as compared with traditional online ads, is that they can be propagated from user to user. In particular, once a user v engages with an ad, such an engagement will appear in the feed of v's followers, who could be influenced to engage with the same ad. This potentially could trigger viral contagion.

In this paper, we study the revenue maximization problem in the context of incentivized social advertising (Aslay et al. 2016; Han et al. 2021). Our model involves two major players: advertiser and platform. At the beginning of a campaign, the advertiser submits a budget limit B to the platform. The platform is in charge of running the campaign and planning budget on behalf of the advertiser. In particular, the platform can spend the budget on behalf of the advertiser in two ways: recruiting seeds and paying for the viral engagements generated by the seeds. Note that the first part of the payment goes to the seeds and the latter one is the actual revenue collected by the platform. This motivates us to define the objective function of the platform as the minimum of the remaining budget after paying seeds and the value of viral engagements generated by those seeds. Formally, if the platform selects S as seeds at cost $c(S)$ and it generates $g(S, \Phi)$ engagements at the end of the day, where Φ is a random variable capturing the uncertainty about the propagation of the engagements, we can represent the expected revenue $f_{exp}(S)$ of the platform as

$$f_{exp}(S) = \mathbb{E}_{\Phi}[\min\{\mathsf{PPE} \times g(S, \Phi), B - c(S)\}]. \tag{1}$$

Example. Suppose the budget of the advertiser is $B = \$20$ and the PPE is \$1. The platform hires a group of seeds S at $c(S) = \$7$ such that it generates $g(S, \Phi) = 15$ engagements at the end of the campaign. In this case, the remaining budget after hiring S is $B - c(S) = \$13$. Thus, the actual revenue collected from viral engagements is $\min\{15 \times \$1, \$13\} = \$13$.

We study the revenue maximization seed selection problem from the platform's perspective. Our goal is to select the best S (resp. a policy) that maximise $f_{exp}(S)$ under the non-adaptive setting (resp. the adaptive setting). Intuitively, we aim to collect as many engagements as possible, while reducing the cost of hiring seed users.

Our Results: We next summarize the main contributions made in this paper.

1. Under the non-adaptive setting, one must pick a group of seed users S all at once in advance to maximize $f_{exp}(S)$. Notice that $f_{exp}(S)$ is non-monotone and it might take on negative values, making most of the existing results on submodular optimization and influence maximization not appli-

cable to our setting. This problem becomes even more challenging by having to take into account the uncertainty about the spread of engagements through a social network. For the general non-adaptive seed selection problem, we develop an algorithm that achieves a $\frac{1-e^{-1/2}}{4}$ approximation ratio.

2. Under the more complicated adaptive setting, one is allowed to pick seed users sequentially and adaptively, where each selection is based on the partial realizations of selected seeds (e.g., we choose the next seed given who we have selected as seeds so far, and how many engagements they have generated). Formally, we can encode any policy using a function from a set of partial realizations to the set of users, specifying which seed to select next under a particular partial realization of selected seeds. In this setting, our goal is to design a policy, rather than finding a fixed set of seeds, to maximize the expected revenue. We develop an adaptive strategy that achieves a $\frac{\kappa(1-e^{-C/B})}{2}$ approximation ratio against the optimal adaptive policy where B is the budget constraint and C, as well as κ, is some term that is dependent on the cost of the most expensive seed. We show that if the cost of the most expensive seed is no larger than $B/2$, then the above approximation ratio is lower bounded by $\frac{1-e^{-1/2}}{4}$.

3. We conduct extensive experiments on four large-scale benchmark social networks (*Wikivote*, *NetHEPT*, *NetPHY* and *Epinions*) to evaluate the performance of our solutions. And the experiment results validate the effectiveness and the efficacy of our algorithms.

Related Work

Influence Maximization and Social Advertising. The problem of influence maximization has been studied in (Kempe, Kleinberg, and Tardos 2003; Golovin and Krause 2011) where their objective is to select a group of seed nodes to maximize the expected size of influence. They assumed that advertisers could observe the structure of the entire social network and select seed users from this network on their own. However, this assumption does not hold true in the context of social advertising, where the platform selects seed users on behalf of the advertisers. As a result, their objective functions are monotone and (adaptive) submodular, while ours is non-monotone. Our research aligns closely with previous investigations into social advertising. For instance, (Chalermsook et al. 2015) study the social advertising problem involving multiple advertisers. However, their approach treats the cost of selecting seeds as sunk, with each advertiser i able to select up to a predetermined k_i seeds. Consequently, their focus lies in maximizing a monotone function within cardinality constraints. The channel allocation problem is explored in (Alon, Gamzu, and Tennenholtz 2012), while (Abbassi, Bhaskara, and Misra 2015) tackles the user ordering problem. However, neither of these studies addresses engagement propagation. In (Aslay et al. 2014), they introduce the concept of regret to capture the tradeoff between maximizing the social advertising revenue and minimizing the impact of free-riding. Their goal is to select a group of seeds to minimize the regret. This problem is revisited in (Tang and Yuan 2016), they convert it to a new optimization problem which admits constant approximation algorithms under some conditions. In (Aslay et al. 2016), they initiate the study of revenue maximization incentivized social advertising problem. Although the basic business model adopted in their paper is similar to ours, we propose to use a new utility function to capture the revenue of the platform.Specifically, in (Aslay et al. 2016)'s setting, it is not possible to provide any "free engagements" to the advertiser due to the strict budget constraint imposed on the total cost of seed hiring and viral engagements. However, our setting offers greater flexibility to the platform, enabling it to offer free-riding services to advertisers. Our problem can be viewed as a relaxation of (Aslay et al. 2016)'s problem, as any feasible solution for their problem is also feasible for ours, but not the other way around. Therefore, the optimal solution under our setting always yields a higher expected revenue from the platform's perspective. This observation is also backed by the results of our experiment. From a technical point of view, the objective function defined in (Aslay et al. 2016) is monotone and submodular, while our objective function is non-monotone and it might take on negative values. This makes the existing results on submodular optimization (Nemhauser, Wolsey, and Fisher 1978) and influence maximization (Kempe, Kleinberg, and Tardos 2003) not applicable to our setting.

Incentive Mechanism Design in Social Media Numerous studies have been conducted on developing incentive mechanisms aimed at promoting participation in social media (Brady, Morris, and Bigham 2015; Alperin et al. 2017). These studies explore different types of incentives, such as personalized message (Grau, Naderi, and Kim 2018) or using online bots (Savage, Monroy-Hernandez, and Höllerer 2016), and evaluating their effectiveness in motivating people to participate in a particular activity or behavior. In contrast, our research is focused on selecting a best group of seed users given that the incentive mechanism is well designed and pre-given (e.g., in our study, we assume an incentive mechanism where a fixed amount of monetary reward is offered to each seed user). Our work distinguishes itself from and complements the research cited in the aforementioned studies. On the one hand, the research on incentive mechanism design can assist us in formulating better incentives, such as combining monetary rewards with other types of incentives, to more effectively recruit seed users in the context of social advertising. On the other hand, given that distributing incentives to a large number of users can often be expensive, our study may provide valuable guidance on how to selectively target a group of users to receive those incentives.

Submodular Optimization. Our study is also related to non-adaptive (Nemhauser, Wolsey, and Fisher 1978) and adaptive submodular maximization (Golovin and Krause 2011). While most of the existing studies in this field assume non-negative functions, (Harshaw et al. 2019; Sviridenko, Vondrák, and Ward 2017) study the problem of maximizing a regularized submodular function, which may take on negative values, in the form of a sum of a non-negative monotone increasing submodular function and a linear function. (Feldman 2020) develop a faster algorithm using a surrogate objective that varies with time. For the case of a cardinality constraint and a non-positive linear part, (Harshaw et al. 2019; Kazemi et al. 2020) develop the first practical algorithms. Our objective function is different from theirs, i.e., it is the expectation of the minimum of an increasing stochastic submodular function and a decreasing linear function. Moreover, we are the first to extend this study to the adaptive setting where the seed users are selected in a sequential and adaptive manner.

Preliminaries and Problem Statement

Engagement Propagation Model and Submodularity

The platform owns a social network which is represented as a directed graph $G = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V}$ represents a online user and an edge $(u, v) \in \mathcal{E}$ means that user v is a follower of user u, and thus v is exposed to u's posts and may be influenced by u. We adopt the Independent Cascade Model (IC) (Kempe, Kleinberg, and Tardos 2003) to govern the way in which engagements (e.g., impressions, clicks and shares) propagate in G. Under the IC model, each edge $(u, v) \in \mathcal{E}$ has a binary random variable X_{uv} that denotes whether u has influenced v (to engage with the ad). That is $X_{uv} = 1$ indicates that u successfully influences v and otherwise $X_{uv} = 0$. The random variables X_{uv} are independent of each other and have a known mean of $\mathbb{E}[X_{uv}] = \rho_{uv}$, where ρ_{uv} represents the probability of influence that u has on v. An edge (u, v) will be referred to as a *live* edge if $X_{uv} = 1$, whereas an edge with $X_{uv} = 0$ will be referred to as a *blocked* edge. Once a node u becomes activated (e.g., u has engaged with the ad), the model samples the edges X_{uv} for each of its neighbors v. If the edge (u, v) is a live edge ($X_{uv} = 1$), then the node v becomes activated. This process can then continue, as influence spreads from u's neighbors to their neighbors and so on, following the same mechanism. We illustrate this engagement propagation model in Figure 1.

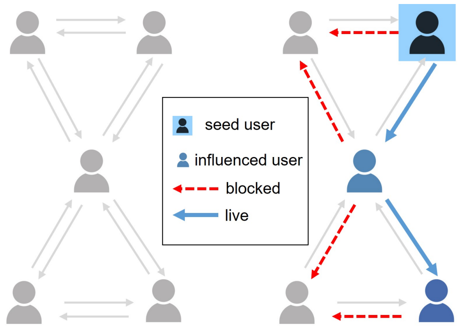

Figure 1: An illustration of the engagement propagation model. The underlying social network is shown on the left-hand side; and the individuals influenced and observed following the selection of a single seed user are shown on the right-hand side.

Since the diffusion process can lead to the activation of multiple nodes through the network, triggering the activation of a single node u can potentially have far-reaching effects. We assume that selecting a seed u can reveal the status (`Blocked` or `Live`) of every out-going edge of every user that can be reached by u though a path composed of *live* edges. We model the state $\phi(u)$ of a user u as a function $\phi(u) : \mathcal{E} \rightarrow \{\texttt{Blocked}, \texttt{Live}, ?\}$. Specifically, $\phi(u)((v, w)) = \texttt{Blocked}$ means that activating u has revealed that (v, w) is *blocked*, $\phi(u)((v, w)) = \texttt{Live}$ means that activating u has revealed that (v, w) is *live*, and $\phi(u)((v, w)) =?$ means that activating u can not reveal the status of (v, w). Let $\phi = \{\phi(v) \mid v \in \mathcal{V}\}$ denote a *full realization* of the diffusion process and Φ denote the random variable of ϕ. Given a graph $G = (\mathcal{V}, \mathcal{E})$ and influence probabilities $\{\rho_{uv} \mid (u, v) \in \mathcal{E}\}$, there is a known prior probability distribution $p(\phi) = \{\Pr[\Phi = \phi] : \phi \in \Omega\}$ over all possible realizations Ω. Given a realization ϕ and a set of seed users S, we define $g(S, \phi)$ as the number of engagements generated by S conditional on ϕ. In other words, $g(S, \phi)$ is the number of users who can be reached by at least one user in S (including S itself) through live edges conditional on ϕ. The expected number of engagements of S is $g_{exp}(S) = \mathbb{E}[g(S, \Phi)]$ where the expectation is taken over Φ according to $p(\phi)$. We next introduce the concept of submodularity and monotonicity.

Definition 1 (Submodularity and Monotonicity) *Consider any two subsets $A \subseteq \mathcal{V}$ and $B \subseteq \mathcal{V}$ such that $A \subseteq B$ and any element $v \in \mathcal{V} \setminus B$, a function $q : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ is submodular if $q(A \cup \{v\}) - q(A) \geq q(B \cup \{v\}) - q(B)$. A function $q : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ is monotone if $q(A \cup \{v\}) - q(A) \geq 0$.*

In (Kempe, Kleinberg, and Tardos 2003), they show that $g(\cdot, \phi)$ is both monotone and submodular for any realization ϕ with $\Pr[\Phi = \phi] > 0$. Moreover, $g_{exp}(\cdot)$ is also monotone and submodular.

Business Model

The platform is hosting the social advertising campaign. After receiving a campaign budget B, as well as a PPE amount ppe, from the advertiser, the platform selects a group of seed users $S \subseteq \mathcal{V}$ on behalf of the advertiser to endorse her ad, and each seed user $v \in S$ receives an incentive $c(v)$. Hence, the cost of hiring S as seed users is $c(S) = \sum_{v \in S} c(v)$. The total payment made by the advertiser is composed of two parts: The payment for the engagements and the cost for hiring the seed users. Notice that the revenue of the platform is just the payment for the engagements, as $c(S)$ is paid to the seed users. If there is no budget constraint, i.e, $B = \infty$, the revenue of the platform conditional on a realization ϕ is simply $ppe \cdot g(S, \phi)$ where $g(S, \phi)$ is the number of engagements generated by S under ϕ. For a general budget constraint B, we must ensure that the total payment from the advertiser does not exceed B. Thus, as the cost of hiring seed users is $c(S)$, the highest possible payment for engagements is upper bound by $B - c(S)$. As a result, we model the actual revenue $f(S, \phi)$ of the platform subject to a budget constraint B as $f(S, \phi) = \min\{ppe \cdot g(S, \phi), B - c(S)\}$. Note that our model allows for the possibility of providing "free" engagements to the advertiser if $ppe \cdot g(S, \phi)$ exceeds $B - c(S)$. Without

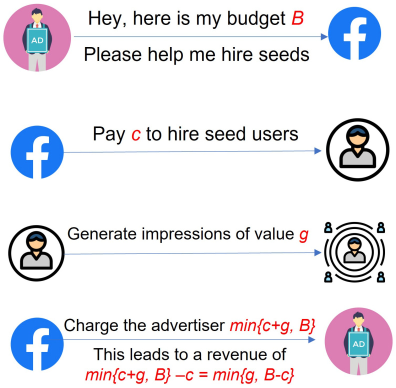
Hey, here is my budget B
Please help me hire seeds
Pay c to hire seed users
Generate impressions of value g
Charge the advertiser min{c+g, B}
This leads to a revenue of
min{c+g, B} –c = min{g, B-c}
AD

Figure 2: Business model of incentivized social advertising (under a fixed realization ϕ). In step 1, the advertiser submitted a budget B to the platform (e.g., facebook). In step 2, the platform spent c on hiring seed users. In step 3, assume those seed users hired by the platform successfully generate viral engagements (e.g., viral impressions) of value g. In the final step, the platform charge the advertiser $\min\{c+g, B\}$ (this is because the total payment from the advertiser can not exceed his budget B). Note that although the total payment by the advertiser is $\min\{c+g, B\}$, but c was given to the seed users. In this case, the actual revenue collected by the platform is $\min\{c+g, B\} - c = \min\{g, B-c\}$.

loss of generality, we normalize the value of *ppe* to one to obtain a simplified form of $f(S, \phi)$ as

$$f(S, \phi) = \min\{g(S, \phi), B - c(S)\}. \tag{2}$$

With the above notation, we can represent the expected revenue $f_{exp}(S)$ of the platform when S is selected as

$$f_{exp}(S) = \mathbb{E}[f(S, \Phi)] = \mathbb{E}[\min\{g(S, \Phi), B - c(S)\}] \tag{3}$$

where the expectation is taken over Φ according to $p(\phi)$. To gain a better understanding of the business model, please refer to a walkthrough example provided in Figure 2.

In this paper, we study the revenue maximization seed selection problem from the platform's perspective under both non-adaptive and adaptive settings. Under the non-adaptive setting, one must pick a group of seed users all at once in advance. Under the more complicated adaptive setting, one is allowed to pick seed users sequentially and adaptively, where each selection is based on the feedback from the past observations. Under both settings, the platform faces the tradeoff of generating as many engagements as possible and reducing the cost of hiring influential seed users.

Non-adaptive Seed Selection Problem

We first describe the seed selection problem under the non-adaptive setting. Our objective is to select a fixed set of seed users to maximize the expected revenue $f_{exp}(S)$. Formally,

$$\max_{S \subseteq \mathcal{V}} \{f_{exp}(S) \mid c(S) \leq B\}.$$

One can drop the constraint from the above formulation without affecting its optimal solution. This is because if S is an optimal solution to the above problem, then it must satisfy $f_{exp}(S) \geq 0$, which implies $c(S) \leq B$, otherwise, we can pick $\emptyset$ as a better solution which contradicts to the assumption that S is an optimal solution. Notice that the utility function $f_{exp}(\cdot)$ might take on negative values, which renders the existing results on submodular optimization and influence maximization ineffective.

Adaptive Seed Selection Problem

Unlike the non-adaptive setting where we select a fixed set of seed users, an adaptive solution selects seed users sequentially and adaptively, where the selection taken in each step depends on information obtained in the previous steps. In particular, it monitors the realized influence of the current set of seed users, and uses this information to adjust the selection criteria when selecting subsequent seed users. This can improve the outcomes of social advertising campaigns. For instance, if an online user v has already been successfully influenced in an intermediate stage, selecting v itself or v's neighbors as seed users in subsequent rounds may be less effective, as there may be diminishing returns to targeting users who have already been influenced. Instead, it may be more beneficial to select other users who have not yet been exposed to the campaign, or who have different characteristics that could help expand the reach of the campaign to new audiences.

While an adaptive solution can have significant advantages, designing an effective adaptive solution can be extremely challenging. In adaptive settings, feasible solutions are no longer simple subsets, but rather complex policies like decision trees, which specify actions based on the current system state. These policies can have an exponentially large search space, making it difficult to find a good solution.

We follow the framework of (Golovin and Krause 2011) to introduce some notations. Formally, we can represent any policy using a mapping function π that maps a set of partial realizations to $\mathcal{V}$: $\pi : 2^{\mathcal{V}} \times O^{\mathcal{V}} \rightarrow \mathcal{V}$ where $O = \{\texttt{Blocked}, \texttt{Live}, ?\}^{\mathcal{E}}$ denotes the set of all possible states of a user. Intuitively, a policy specifies which seed user to select next after observing a partial realization. Consider the following example for an illustration. Assume the current observation is $(u, \phi(u))$, i.e., u is the sole seed that is chosen, and $\phi(u)$ denotes the resulting influence that u generates. Assume π is designed such that $\pi(\{(u, \phi(u))\}) = w$, then π selects w as the next seed user. One is also allowed to design a randomized policy that maps a partial realization to a distribution of users. Since every randomized policy can be represented as a distribution of a set of deterministic policies, we focus on deterministic policies in this paper.

Let $\mathcal{V}(\pi, \phi)$ denote the set of seed users selected by π conditional on a realization ϕ. Then the expected utility $f_{avg}(\pi)$ of a policy π can be written as

$$f_{avg}(\pi) = \mathbb{E}[f(\mathcal{V}(\pi, \Phi), \Phi)]$$

where the expectation is taken over Φ according to $p(\phi)$. With the above notations, the adaptive seed selection problem can be formulated as follows:

$$\max_{\pi} \{f_{avg}(\pi) \mid c\,(\mathcal{V}(\pi, \phi)) \leq B, \forall \phi : \Pr[\Phi = \phi] > 0\}.$$

Again, one can drop the constraint from the above formulation without affecting its optimal solution. We next introduce some additional notations from (Golovin and Krause 2011). Given any $S \subseteq \mathcal{V}$, let $\psi = \{\phi(v) \mid v \in S\}$ denote a *partial realization* (e.g., ψ encodes who we have selected as seeds and who have been influenced by them) and $\mathrm{dom}(\psi) = S$ is the *domain* of ψ. Given a partial realization ψ and a realization ϕ, we say ϕ is consistent with ψ, denoted $\phi \sim \psi$, if they are consistent everywhere in $\mathrm{dom}(\psi)$.

All missing proofs are moved to our technical report (Tang and Yuan 2021).

Non-Adaptive Seed Selection Problem

In this section, we focus on the non-adaptive seed selection problem and propose an algorithm that provides an approximate solution with an approximation ratio of $\frac{1-e^{-\frac{1}{2}}}{4}$. Here, e refers to Euler's number, a mathematical constant with an approximate value of 2.71828. That is, our algorithm identifies a group of seed users that can produce a revenue that is at least a constant (e.g., $\frac{1-e^{-\frac{1}{2}}}{4}$) fraction of the revenue achieved from an optimal solution.

The rest of this section is organized as follows. We first study a relaxed version of the non-adaptive seed selection problem by assuming that the cost $c(S^*) = \sum_{e \in S^*} c(e)$ of the optimal solution S^* is known. Although the value of $c(S^*)$ is rarely known in practise, we start with this simplified case to make it easier to explain the idea of our approach for solving the general problem without knowing $c(S^*)$.

Warming Up: Seed Selection with Known $c(S^*)$

To facilitate our algorithm design, we first introduce a new utility function. For any $z \in [0, B]$, define $l(\cdot, z) : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ as follows:

$$l(S, z) = \mathbb{E}[\min\{g(S, \Phi), B - z\}].$$

We first show that $l(\cdot, z) : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ is monotone and submodular for any $z \in [0, B]$.

Lemma 1 *For any $z \in [0, B]$, $l(\cdot, z) : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ is monotone and submodular.*

Proof: It has been proved in (Kempe, Kleinberg, and Tardos 2003) that $g(\cdot, \phi)$ is monotone and submodular for any fixed realization ϕ, thus $\min\{g(S, \phi), B - z\}$ is also monotone and submodular for any $z \in [0, B]$. This is because monotone submodular functions remain so under *truncation* (Krause and Golovin 2014) , i.e., the minimum of any monotone and submodular function and any constant is still monotone and submodular. Moreover, because submodularity is preserved under taking nonnegative linear combinations, we have $\mathbb{E}[\min\{g(S, \Phi), B - z\}]$ is also monotone and submodular for any $z \in [0, B]$. This finishes the proof of this lemma. □

In the rest of this paper, for any $x \in \mathbb{R}_{\geq 0}$, let $\mathcal{V}(x) = \{e \in \mathcal{V} \mid c(e) \leq x\}$ denote the set of users whose cost is no larger than x. Assume $c(S^*)$ is given, we introduce a new optimization problem **P.1** as follows.

P.1: *Maximize* $l\left(S, c(S^*)\right)$ **subject to:** $\begin{cases} S \subseteq \mathcal{V}\left(c(S^*)\right) \\ c(S) \leq c(S^*) \end{cases}$

Because $l\left(\cdot, c(S^*)\right) : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ is submodular and monotone (Lemma 1), **P.1** is a classical monotone submodular maximization problem subject to a knapsack constraint. To solve this problem efficiently, we consider two candidate solutions. One is the singleton e^* maximizing $l(\cdot, c(S^*)) : 2^{\mathcal{V}} \rightarrow \mathbb{R}_{\geq 0}$ among all users in $\mathcal{V}\left(c(S^*)\right)$, i.e., $e^* = \arg\max_{e \in \mathcal{V}(c(S^*))} l\left(\{e\}, c(S^*)\right)$, and the other one is the output from a benefit-cost greedy algorithm $\mathsf{Greedy}\left(c(S^*), c(S^*)\right)$ listed in Algorithm 1. Note that $\mathsf{Greedy}(x, y)$ is presented as a general template that takes two parameters, i.e., x and y, and this template makes it easier to describe our solution for solving the general problem later. Intuitively, $\mathsf{Greedy}(x, y)$ refers to a benefit-cost greedy algorithm that runs on a ground set $\mathcal{V}(x)$ using the utility function $l(\cdot, y)$ subject to a knapsack constraint x. We next describe $\mathsf{Greedy}\left(c(S^*), c(S^*)\right)$ in details. It starts with iteration $t = 0$ and an initial solution $S_0 = \emptyset$, and in each subsequent iteration $t + 1$, it adds s_{t+1} to the existing solution S_t, i.e., $S_{t+1} \leftarrow S_t \cup \{s_{t+1}\}$, where

$$s_{t+1} = \underset{e \in \mathcal{V}(c(S^*)) \setminus S_t}{\arg\max} \frac{l\left(S_t \cup \{e\}, c(S^*)\right) - l\left(S_t, c(S^*)\right)}{c(e)}$$

denotes the user that maximizes the benefit-cost ratio with respect to S_t. This process iterates until it reaches some iteration t such that $c(S_{t+1}) + c(s_{t+1}) > c(S^*)$.

Algorithm 1: $\mathsf{Greedy}(x, y)$

```
$S_0 = \emptyset, t = 0, U = \mathcal{V}(x)$
while $U \setminus S_t \neq \emptyset$ do
    let $s_{t+1}$ denote the user $e$ maximizing $\frac{l(S_t \cup \{e\}, y) - l(S_t, y)}{c(e)}$ among all users in $U \setminus S_t$
    if $c(S_{t+1}) + c(s_{t+1}) \leq x$ then
        let $S_{t+1} = S_t \cup \{s_{t+1}\}$
        $t \leftarrow t + 1$
    else
        return $S_t$
return $S_t$
```

Let $S^{\mathsf{Greedy}(c(S^*), c(S^*))}$ denote the solution returned from $\mathsf{Greedy}(c(S^*), c(S^*))$. We next show that the better one between $S^{\mathsf{Greedy}(c(S^*), c(S^*))}$ and $\{e^*\}$ achieves a $\frac{1-1/e}{2}$ approximation ratio for our original problem.

Theorem 1 *Let $S^{\mathsf{Greedy}(c(S^*), c(S^*))}$ denote the solution returned from $\mathsf{Greedy}(c(S^*), c(S^*))$ and $e^* = \arg\max_{e \in \mathcal{V}(c(S^*))} l\left(\{e\}, c(S^*)\right)$,*

$$\max\left\{f_{exp}\left(S^{\mathsf{Greedy}(c(S^*), c(S^*))}\right), f_{exp}\left(\{e^*\}\right)\right\}$$

$$\geq \frac{1 - 1/e}{2} f_{exp}(S^*).$$

Algorithm 2: Non-adaptive Seed Selection Algorithm

```
run $\mathsf{Greedy}(\frac{B}{2}, 0)$ and obtain an output $S^{\mathsf{Greedy}(\frac{B}{2},0)}$
let $S^{phase1} = \arg\max_{S\in\{S^{\mathsf{Greedy}(\frac{B}{2},0)}, v(\frac{B}{2},0)\}} f_{exp}(S)$
let $\mathcal{V}_{large} = \mathcal{V} \setminus \mathcal{V}(\frac{B}{2})$
for $e \in \mathcal{V}_{large}$ do
    run $\mathsf{Greedy}(c(e), c(e))$ and obtain an output $S^{\mathsf{Greedy}(c(e),c(e))}$
let $S^{phase2}$ be the best solution in $\bigcup_{e\in\mathcal{V}_{large}}\{S^{\mathsf{Greedy}(c(e),c(e))}, v\left(c(e), c(e)\right)\}$
return the better solution between $S^{phase1}$ and $S^{phase2}$
```

Solving the General Seed Selection Problem

Now we are in position to drop the assumption about knowing the value of $c(S^*)$. One naive approach of dealing with unknown $c(S^*)$ is to try all possibilities of $c(S^*)$. Then we feed each "guess" to Algorithm 1 to obtain an output. At last, we choose the output maximizing the expected utility among all returned solutions as the final solution. This approach can secure an approximation ratio no worse than $\frac{1-1/e}{2}$. However, the number of possible values of $c(S^*)$ is exponentially large in terms of n, it is clearly not affordable to enumerate all those possibilities. Perhaps surprisingly, we next show that it is not necessary to find out the value of $c(S^*)$, rather, we only need to find out the cost of the "most expensive" user in the optimal solution S^*, i.e., $\max_{e\in S^*} c(e)$. This can be done in $O(n)$ time since there are only n possibilities of this value. Our algorithm (Algorithm 2) is composed of two phases. The first phase is dealing with the case when $\max_{e\in S^*} c(e) \leq \frac{B}{2}$, i.e., the cost of the most expensive seed from the optimal solution is no larger than $B/2$, and the second phase is used to handle the rest of the cases. At last, we return the solution maximizing the expected utility among all candidate outputs.

Before describing the design of our algorithm in details, we introduce some notations. For any $x, y \in \mathbb{R}_{\geq 0}$, let $v(x, y) \in \arg\max_{e\in\mathcal{V}(x)} l(\{e\}, y)$ denote the singleton maximizing $l(\cdot, y)$ among users in $\mathcal{V}(x)$. Recall that for any $x \in \mathbb{R}_{\geq 0}$, we use $\mathcal{V}(x) = \{e \in \mathcal{V} \mid c(e) \leq x\}$ to denote the set of users whose cost is no larger than x, and $\mathsf{Greedy}(x, y)$ (Algorithm 1) refers to a benefit-cost greedy algorithm that runs on a ground set $\mathcal{V}(x)$ using the utility function $l(\cdot, y)$ subject to a knapsack constraint x. Now we are ready to describe our algorithm. Our solution is composed of two phases:

Phase 1 Run $\mathsf{Greedy}(\frac{B}{2}, 0)$ to obtain $S^{\mathsf{Greedy}(\frac{B}{2},0)}$. Let S^{phase1} be the better solution in $\{S^{\mathsf{Greedy}(\frac{B}{2},0)}, v(\frac{B}{2}, 0)\}$.

Phase 2 Let $\mathcal{V}_{large} = \mathcal{V} \setminus \mathcal{V}(\frac{B}{2})$ denote the set of all users whose cost is larger than $\frac{B}{2}$. For each $e \in \mathcal{V}_{large}$, run $\mathsf{Greedy}(c(e), c(e))$ to obtain $S^{\mathsf{Greedy}(c(e),c(e))}$. Let S^{phase2} be the best solution in

$$\bigcup_{e\in\mathcal{V}_{large}} \left\{S^{\mathsf{Greedy}(c(e),c(e))}, v\left(c(e), c(e)\right)\right\}.$$

Output Return the better solution between S^{phase1} and S^{phase2} as the final output.

We next analyze the approximation ratio of our algorithm. We first present two technical lemmas. In the first lemma, we show that if $\max_{e\in S^*} c(e) > \frac{B}{2}$, i.e., the cost of the most expensive seed from the optimal solution is larger than $B/2$, then the solution returned from the second phase of our algorithm S^{phase2} is near optimal. In the second lemma, we show that if $\max_{e\in S^*} c(e) \leq \frac{B}{2}$, , i.e., the cost of the most expensive seed from the optimal solution is no larger than $B/2$, then the solution returned from the first phase of our algorithm S^{phase1} is near optimal. Combining these two lemmas, we are able to derive an approximation bound of our algorithm for the original problem.

Lemma 2 *If* $\max_{e\in S^*} c(e) > \frac{B}{2}$*, then* $f_{exp}(S^{phase2}) \geq \frac{1-e^{-\frac{1}{2}}}{2} f_{exp}(S^*)$.

Proof: Let $e' \in \arg\max_{e\in S^*} c(e)$ denote the most expensive user in the optimal solution S^*. To prove this lemma, we first introduce a new optimization problem **P.3**.

P.3: *Maximize* $l\left(S, c(e')\right)$ **subject to:** $\begin{cases} S \subseteq \mathcal{V}(c(e')) \\ c(S) \leq B \end{cases}$

Because $\max_{e\in S^*} c(e) \leq c(e')$, we have $S^* \subseteq \mathcal{V}\left(c(e')\right)$. Thus, S^* is a feasible solution to **P.3**. Moreover, we have $l\left(S^*, c(e')\right) = \mathbb{E}[\min\{g(S^*, \Phi), B - c(e')\}] \geq \mathbb{E}[\min\{g(S^*, \Phi), B - c(S^*)\}] = f_{exp}(S^*)$ where the inequality is due to $e' \in S^*$. Let S^{p3} denote the optimal solution to **P.3**, then we have $l\left(S^{p3}, c(e')\right) \geq l\left(S^*, c(e')\right) \geq f_{exp}(S^*)$. To prove this lemma, it suffices to show that $f_{exp}(S^{phase2}) \geq \frac{1-e^{-\frac{1}{2}}}{2} l\left(S^{p3}, c(e')\right)$.

Note that because $l(\cdot, c(e'))$ is monotone and submodular (due to the same proof as for Lemma 1), **P.3** is a classical monotone submodular maximization problem subject to a knapsack constraint. By abuse of notation, let $R = \arg\max_{S\in\left\{S^{\mathsf{Greedy}\left(c(e'),c(e')\right)}, \{v(c(e'),c(e'))\}\right\}} l\left(S, c(e')\right)$ denote the solution maximizing $l(\cdot, c(e'))$ between $S^{\mathsf{Greedy}\left(c(e'),c(e')\right)}$ and $\{v\left(c(e'), c(e')\right)\}$, we next show that $l\left(R, c(e')\right) \geq \frac{1-e^{-\frac{c(e')}{B}}}{2} l\left(S^{p2}, c(e')\right)$. Consider a "one-step-further" version $\mathsf{Greedy}^{+}\left(c(e'), c(e')\right)$ of $\mathsf{Greedy}\left(c(e'), c(e')\right)$ which is obtained by first running $\mathsf{Greedy}\left(c(e'), c(e')\right)$ then selecting one more user according to the same greedy manner. One can verify that $\mathsf{Greedy}^{+}\left(c(e'), c(e')\right)$ always violates the budget constraint $c(e')$ (assuming $\mathcal{V}(c(e'))$ contains more than one users to avoid trivial cases). Let $S^{\mathsf{Greedy}^{+}\left(c(e'),c(e')\right)}$ denote the solution returned from $\mathsf{Greedy}^{+}\left(c(e'), c(e')\right)$. According to Theorem 2[1] in (Tang and Yuan 2020), if $l(\cdot, c(e'))$ is monotone and submodular, then

$$l\left(S^{\mathsf{Greedy}^{+}\left(c(e'),c(e')\right)}, c(e')\right) \geq (1 - e^{-\frac{c(e')}{B}}) l\left(S^{p2}, c(e')\right). \quad (4)$$

[1] The original theorem provides a stronger result than what we need here. I.e., they show that this result holds even under the adaptive setting.

Because $l(\cdot, c(e'))$ is monotone and submodular, the marginal utility brought by the last added user in $S^{\textsf{Greedy}^{+}\left(c(e'),c(e')\right)}$ is no larger than expected utility of the best singleton $v\left(c(e'), c(e')\right)$. Thus,

$$\begin{aligned} & l\left(S^{\textsf{Greedy}^{+}\left(c(e'),c(e')\right)}, c(e')\right) \\ & \leq l\left(S^{\textsf{Greedy}\left(c(e'),c(e')\right)}, c(e')\right) + l\left(\{v(c(e'), c(e'))\}, c(e')\right). \end{aligned} \tag{5}$$

This implies that

$$\begin{aligned} & l\left(R, c(e')\right) \\ & \geq \frac{l\left(S^{\textsf{Greedy}\left(c(e'),c(e')\right)}, c(e')\right) + l\left(\{v(c(e'), c(e'))\}, c(e')\right)}{2} \\ & \geq \frac{l\left(S^{\textsf{Greedy}^{+}\left(c(e'),c(e')\right)}, c(e')\right)}{2} \\ & \geq \frac{1-e^{-\frac{c(e')}{B}}}{2} l\left(S^{p2}, c(e')\right). \end{aligned} \tag{6}$$

The first inequality is due to the definition of R. The second inequality is due to (5). The third inequality is due to (4). Then we have

$$\begin{aligned} l\left(R, c(e')\right) & \geq \frac{1-e^{-\frac{c(e')}{B}}}{2} l\left(S^{p2}, c(e')\right) \\ & \geq \frac{1-e^{-\frac{1}{2}}}{2} l\left(S^{p2}, c(e')\right). \end{aligned} \tag{7}$$

The second inequality is due to $c(e') > \frac{B}{2}$. Second, due to $c\left(v(c(e'), c(e'))\right) \leq c(e')$, which is because $v(c(e'), c(e')) \in \mathcal{V}(c(e'))$, and $c(S^{\textsf{Greedy}\left(c(e'),c(e')\right)}) \leq c(e')$, which is because of the design of $\textsf{Greedy}\left(c(e'), c(e')\right)$, we have $B - c\left(v(c(e'), c(e'))\right) \geq B - c(S^*)$ and $B - c(S^{\textsf{Greedy}\left(c(e'),c(e')\right)}) \geq B - c(S^*)$. Thus,

$$B - c(R) \geq B - c(S^*). \tag{8}$$

Inequalities (7) and (8) imply that $f_{exp}(R) = \mathbb{E}[\min\{g(R, \Phi), B - c(R)\}] \geq \mathbb{E}[\min\{g(R, \Phi), B - c(S^*)\}] = l\left(R, c(e')\right) \geq \frac{1-e^{-\frac{1}{2}}}{2} l\left(S^{p3}, c(e')\right)$. Note that $S^{\textsf{Greedy}\left(c(e'),c(e')\right)}$ and $\{v(c(e'), c(e'))\}$ are two of the solutions considered in Phase 2 for its output S^{phase2}. Thus, we have $f_{exp}(S^{phase2}) \geq f_{exp}(R) \geq \frac{1-e^{-\frac{1}{2}}}{2} l\left(S^{p3}, c(e')\right)$. □

Lemma 3 *If* $\max_{e \in S^*} c(e) \leq \frac{B}{2}$, *then* $f_{exp}(S^{phase1}) \geq \frac{1-e^{-\frac{1}{2}}}{4} f_{exp}(S^*)$.

Proof: To prove this lemma, we first introduce a new optimization problem **P.2**.

P.2: *Maximize* $l(S, 0)$ **subject to:** $\begin{cases} S \subseteq \mathcal{V}(\frac{B}{2}) \\ c(S) \leq B \end{cases}$

If $\max_{e \in S^*} c(e) \leq \frac{B}{2}$, i.e., the cost of every seed in S^* is no larger than $B/2$, then we have $S^* \in \mathcal{V}(\frac{B}{2})$. It follows that S^* is a feasible solution to **P.2**. Moreover, we have $l(S^*, 0) = \mathbb{E}[\min\{g(S^*, \Phi), B\}] \geq \mathbb{E}[\min\{g(S^*, \Phi), B - c(S^*)\}] = f_{exp}(S^*)$. Let S^{p2} denote the optimal solution to **P.2**, then we have $l(S^{p2}, 0) \geq l(S^*, 0) \geq f_{exp}(S^*)$ where the first inequality is due to S^{p2} is the optimal solution to **P.2** and S^* is a feasible solution to **P.2**. To prove this lemma, it suffices to show that $f_{exp}(S^{phase1}) \geq \frac{1-e^{-\frac{1}{2}}}{4} l(S^{p2}, 0)$ which is equivalent to showing that $\max\left\{f_{exp}(S^{\textsf{Greedy}(\frac{B}{2},0)}), f_{exp}\left(v(\frac{B}{2}, 0)\right)\right\} \geq \frac{1-e^{-\frac{1}{2}}}{4} l(S^{p2}, 0)$.

By abuse of notation, let $R = \arg\max_{S \in \left\{S^{\textsf{Greedy}(\frac{B}{2},0)}, \{v(\frac{B}{2},0)\}\right\}} l(S, 0)$ denote the solution maximizing $l(\cdot, 0)$ between $S^{\textsf{Greedy}(\frac{B}{2},0)}$ and $\{v(\frac{B}{2}, 0)\}$. Note that because $l(\cdot, 0)$ is monotone and submodular (due to the same proof as for Lemma 1), **P.2** is a classical monotone submodular maximization problem subject to a knapsack constraint. With the same proof as for (7), we can show that

$$l(R, 0) \geq \frac{1-e^{-\frac{1}{2}}}{2} l(S^{p2}, 0). \tag{9}$$

Second, due to $c\left(v(\frac{B}{2}, 0)\right) \leq \frac{B}{2}$, which is because $v(\frac{B}{2}, 0) \in \mathcal{V}(\frac{B}{2})$, and $c(S^{\textsf{Greedy}(\frac{B}{2},0)}) \leq \frac{B}{2}$, which is because of the design of $\textsf{Greedy}(\frac{B}{2}, 0)$, we have $B - c\left(v(\frac{B}{2}, 0)\right) \geq B - \frac{B}{2}$ and $B - c(S^{\textsf{Greedy}(\frac{B}{2},0)}) \geq B - \frac{B}{2}$. Thus,

$$B - c(R) \geq B - \frac{B}{2} = \frac{B}{2}. \tag{10}$$

Then we have $f_{exp}(R) = \mathbb{E}[\min\{g(R, \Phi), B - c(R)\}] \geq \mathbb{E}[\min\{g(R, \Phi), \frac{B}{2}\}] \geq \frac{1}{2}\mathbb{E}[\min\{g(R, \Phi), B\}] = \frac{1}{2} l(R, 0) \geq \frac{1-e^{-\frac{1}{2}}}{4} l(S^{p2}, 0)$ where the first inequality is due to (10) and the second inequality is due to (9). □

Lemma 3 and Lemma 2 together imply that $\max\{f_{exp}(S^{phase1}), f_{exp}(S^{phase2})\} \geq \frac{1-e^{-\frac{1}{2}}}{4} f_{exp}(S^*)$. Thus, we have the following main theorem.

Theorem 2 *Algorithm 2 achieves a* $\frac{1-e^{-\frac{1}{2}}}{4}$ *approximation ratio for the non-adaptive seed selection problem.*

Adaptive Seed Selection Problem

In this section, we study the seed selection problem under the adaptive setting. It was worth noting that our objective function defined in (2) is non-monotone and it might take on negative values. This makes most of existing studies (Golovin and Krause 2011) on monotone adaptive submodular maximization not applicable to our setting. Our key finding is an adaptive policy that attains an approximation ratio of $\frac{1-e^{-\frac{1}{2}}}{4}$ against the optimal adaptive policy, assuming the cost of the costliest user is no larger than half of the total budget (e.g., $B/2$).

Algorithm 3: Greedy Policy π^1

```
$t = 0; S_0 = \emptyset; \psi_0 = \emptyset; U = \mathcal{V}$.
while $U \setminus S_t \neq \emptyset$ do
    let $s_{t+1} = \arg\max_{e \in U \setminus S_t} \frac{\Delta_{h(\cdot,\cdot,0)}(e|\psi_t)}{c(e)}$;
    if $c(s_{t+1}) + c(S_t) \leq C$ then
        $S_{t+1} \leftarrow S_t \cup \{e_{t+1}\}$
        $\psi_{t+1} \leftarrow \psi_t \cup \{(s_{t+1}, \Phi(s_{t+1}))\}$;
        $t \leftarrow t + 1$;
    else
        return $S_t$
return $S_t$
```

Before presenting our adaptive policy, we first introduce some additional notations. Given any number $z \in [0, B]$, define $h(\cdot, \cdot, z) : 2^{\mathcal{V}} \times O^{\mathcal{V}} \to \mathbb{R}_{\geq 0}$ as follows:

$$h(S, \phi, z) = \min\{g(S, \phi), B - z\}$$

And the expected utility $h_{avg}(\pi, z)$ of a policy π under $h(\cdot, \cdot, z)$ is defined as

$$\begin{aligned} h_{avg}(\pi, z) &= \mathbb{E}[h(\mathcal{V}(\pi, \Phi), \Phi, z)] \\ &= \mathbb{E}[\min\{g(\mathcal{V}(\pi, \Phi), \Phi), B - z\}]. \end{aligned}$$

For any $z \in [0, B]$ and partial realization ψ, let $\Delta_{h(\cdot,\cdot,z)}(e \mid \psi)$ denote the marginal utility of e on top of ψ under $h(\cdot, \cdot, z)$, i.e.,

$$\begin{aligned} &\Delta_{h(\cdot,\cdot,z)}(e \mid \psi) \\ &= \mathbb{E}_{\Phi \sim \psi}[h(\mathrm{dom}(\psi) \cup \{e\}, \Phi, z)] \\ &\qquad - \mathbb{E}_{\Phi \sim \psi}[h(\mathrm{dom}(\psi), \Phi, z)]. \end{aligned}$$

Let $\overline{e}$ denote the most expensive user in $\mathcal{V}$, i.e., $\overline{e} \in \arg\max_{e \in \mathcal{V}} c(e)$. Let $C = \max\{c(\overline{e}), \frac{B}{2}\}$. Now we are ready to introduce our *Adaptive Seed Selection Policy* π^s.

Design of Adaptive Seed Selection Policy π^s**.** The design of π^s involves two candidate policies: π^1 and π^2. And π^s samples a policy uniformly at random from $\{\pi^1, \pi^2\}$. Thus, the expected utility $f_{avg}(\pi^s)$ of π^s can be represented as $f_{avg}(\pi^s) = \frac{f_{avg}(\pi^1) + f_{avg}(\pi^2)}{2}$. We next describe the design of π^1 and π^2 in details.

- **Design of** π^1**.** The first candidate π^1 (Algorithm 3) is an adaptive version of benefit-cost greedy algorithm presented in the earlier section. π^1 starts with $t = 0$, an initial solution $S_0 = \emptyset$ and an initial partial realization $\psi_0 = \emptyset$. In each iteration $t + 1$, it adds s_{t+1} to the current solution S_t. i.e., $S_{t+1} \leftarrow S_t \cup \{s_{t+1}\}$, where

  $$s_{t+1} = \arg\max_{e \in U \setminus S_t} \frac{\Delta_{h(\cdot,\cdot,0)}(e \mid \psi_t)}{c(e)}$$

  is the user maximizing the benefit-cost ratio with respect to the current observation ψ_t under $h(\cdot, \cdot, 0)$. Then we update the observation using $\psi_{t+1} \leftarrow \psi_t \cup \{(s_{t+1}, \Phi(s_{t+1}))\}$ and enter the next iteration. This process iterates until it reaches some iteration t such that $c(S_{t+1}) + c(s_{t+1}) > C$. Recall that $C = \max\{c(\overline{e}), \frac{B}{2}\}$.
- **Design of** π^2**.** The second candidate π^2 simply returns the singleton e that maximizes $f_{avg}(\{e\})$.

Performance Analysis Now we are ready to give a performance bound of π^s.

Theorem 3 *Let $C = \max\{c(\overline{e}), \frac{B}{2}\}$ and $\kappa = \min\{\frac{1}{2}, 1 - \frac{C}{B}\}$, and let π^* denote the optimal policy for the adaptive seed selection problem, we have $f_{avg}(\pi^s) \geq \kappa \frac{1 - e^{-\frac{C}{B}}}{2} f_{avg}(\pi^*)$.*

Notice that when $c(\overline{e}) \leq \frac{B}{2}$, i.e., the cost of the most expensive user is no larger than $B/2$, we have $C = B/2$ and $\kappa = 1/2$. In this case, the above approximation ratio is lower bounded by $\frac{1 - e^{-\frac{1}{2}}}{4}$.

Performance Evaluation

In this section, we conduct experiments to evaluate the performance of our proposed algorithms. The performance of considered algorithms is evaluated in terms of their revenue, seeding costs and rate of return. All algorithms are implemented using Java and all experiments are run on a Linux server with Intel Xeon 2.40GHz CPU and 128GB memory.

Experimental Setting

Datasets. We run our experiments on four large-scale benchmark social networks: *Wikivote*, *NetHEPT*, *NetPHY* and *Epinions* (http://snap.stanford.edu/data/). We capture each social network by a directed weighted graph. *Wikivote* dataset captures $103,663$ votes cast by $7,066$ active participants in Wikipedia's electoral processes. Each node in this network represents a user, with edges indicating voting connections between users. *NetHEPT* is sourced from arXiv's High Energy Physics Theory section, featuring $15,233$ nodes representing authors. The network encompasses $62,774$ edges, symbolizing papers co-authored by pairs of authors. *NetPHY*, also sourced from arXiv, originates from the Physics section and comprises $37,154$ nodes and $231,584$ edges, representing collaborative efforts among authors. *Epinion* stands as a trust network from Epinions.com, encompassing $75,879$ users represented as nodes and $508,837$ trust relationships represented as edges.

Seed Incentive Models. We use three seed incentive models (i.e., node seeding cost models) in our experiments. As discussed in previous section, $g_{exp}(\{v\})$ denotes the number of expected engagements generated by node $v \in \mathcal{V}$ and is called the (expected) spread of v. Given a fixed constant $\alpha > 0$ and any node v, these models set the cost of node v as follows:

- Natural Log incentive model: the cost of v is proportional to the natural log of three times of its influence spread, i.e., $c(v) = \alpha \cdot \ln(3 \cdot g_{exp}(\{v\}))$. By taking three times of a node's expected spread, we make sure the cost of isolated nodes in the network is not zero.
- Linear incentive model: the cost of v is proportional to its influence spread, i.e., $c(v) = \alpha \cdot g_{exp}(\{v\})$.
- Random incentive model: the cost of v is randomly sampled from a continuous uniform distribution $\mathcal{U}(0, 10)$.

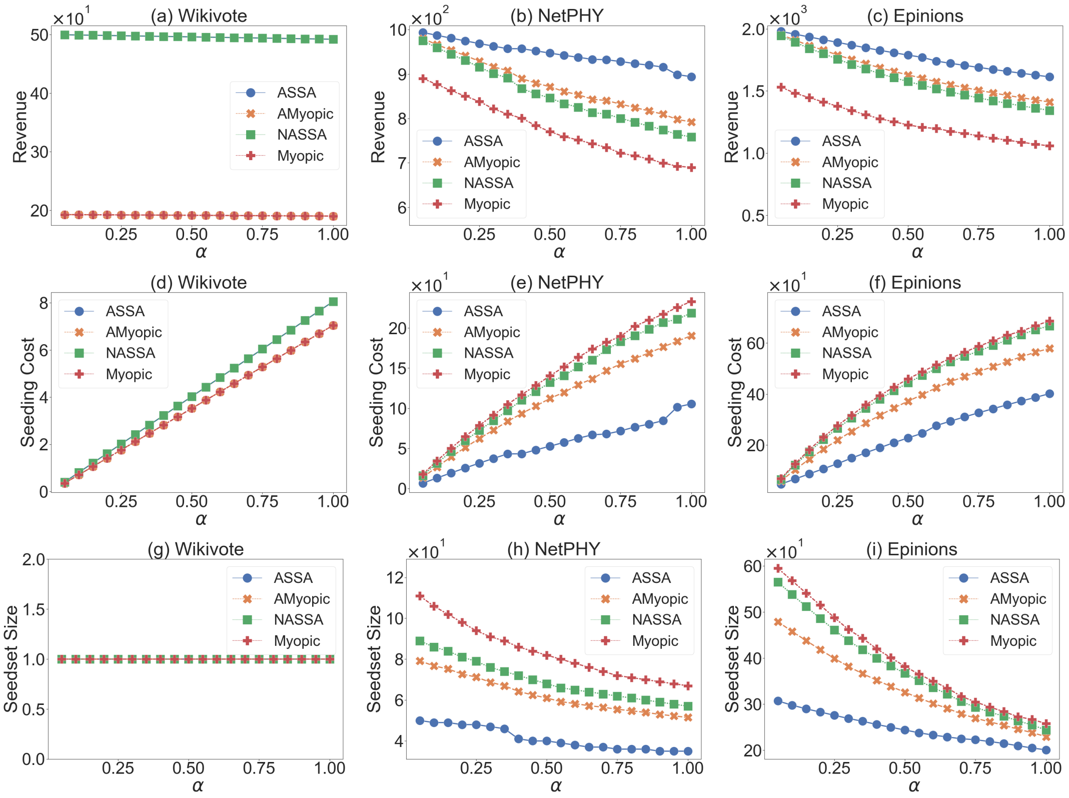

Figure 3: ASSA achieves superior revenue with the lowest seeding cost among all four algorithms under the natural log incentive model.

Algorithms. We evaluate the performance of the following four algorithms under various parameter settings: Non-adaptive Seed Selection Algorithm (NASSA for short, Algorithm 2), Adaptive Seed Selection Algorithm (ASSA for short, π^s), along with two benchmarks for comparison purpose. We implement Myopic algorithm (Aslay et al. 2016) as our non-adaptive benchmark. Myopic first calculates the sum of a node's cost and its expected spread for each node in the network, as a criteria to rule out unqualified nodes. Then it forms a candidate set with all nodes with above value less than or equal to the total budget B. It then runs on this candidate set and iteratively selects a seed node with the largest marginal expected spread to total cost ratio, until the sum of the seed set's cost and its expected spread exceeds B. Here the total cost of a node is calculated as the sum of the node's cost and its marginal expected spread. It is worth noting that Myopic does not allow to offer any free impressions to the advertiser, this is in sharp contrast to our design, which allows for free-riding. We also implement Adaptive Myopic algorithm (AMyopic for short), an adaptive version of Myopic, as our adaptive benchmark. AMyopic iteratively selects a seed node with the largest conditional marginal utility to total cost ratio. Here we measure the conditional marginal utility of a node as the expected increase in influence size based on the realized influence of the current seed set. The total cost of a node is calculated as the sum of the node's cost and its conditional marginal utility. AMyopic runs until the total cost of the seed set exceeds the budget. Here the total cost of a seed set is calculated as the sum of its cost and its conditional expected spread.

Parameter Settings. In our experiments, we adopt the Independent Cascade model as diffusion model and assign an influence probability of $p = 0.05$ to each edge. We set the budget $B = 500$ for *Wikivote* and *NetHEPT*, $B = 1000$ for *NetPHY*, and $B = 2000$ for *Epinions*, respectively. We measure the revenue of NASSA and Myopic by running 10^5 Monte Carlo simulations, generated independently of the considered algorithms. For ASSA, we measure the conditional marginal utility as the expected increase in revenue based on observations of the actual influence spread of the current seed set. For AMyopic, we measure the conditional marginal utility as the expected increase in influence size based on the realized influence triggered by the current seed set.

Experimental Results

General Comparisons. In this section, we compare the algorithms under different node cost models on all datasets. We measure the performance of the algorithms in terms of their revenue, seeding costs and size of seed set with respect to changes in the value of α. Note that increasing α under a fixed budget has a similar impact as decreasing budget with a fixed α. They both impose a tighter budget for selecting the seed set. Due to space limitation and the results being

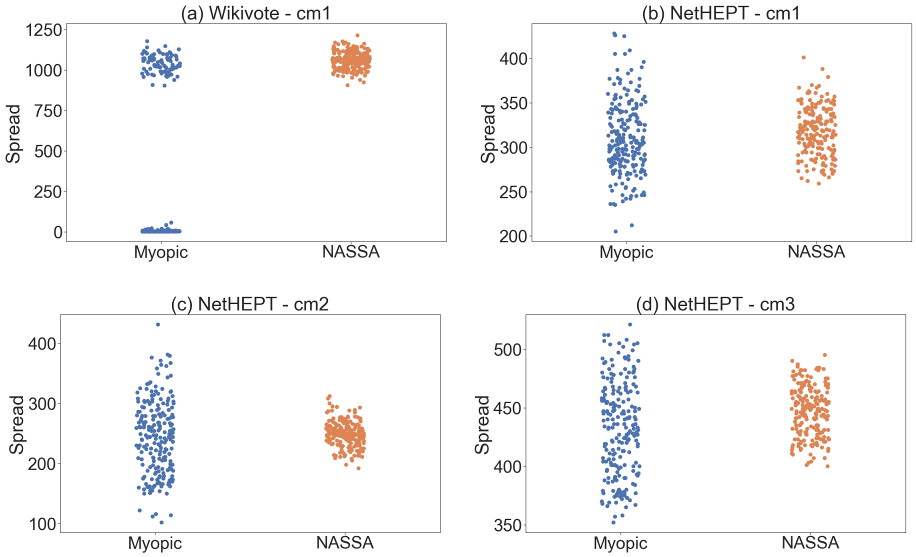

Figure 4: NASSA produces seed sets that generate high spread on all samples, leading to a higher expected revenue; while Myopic produces seed sets with more scattered spread on different samples, leading to a lower expected revenue.

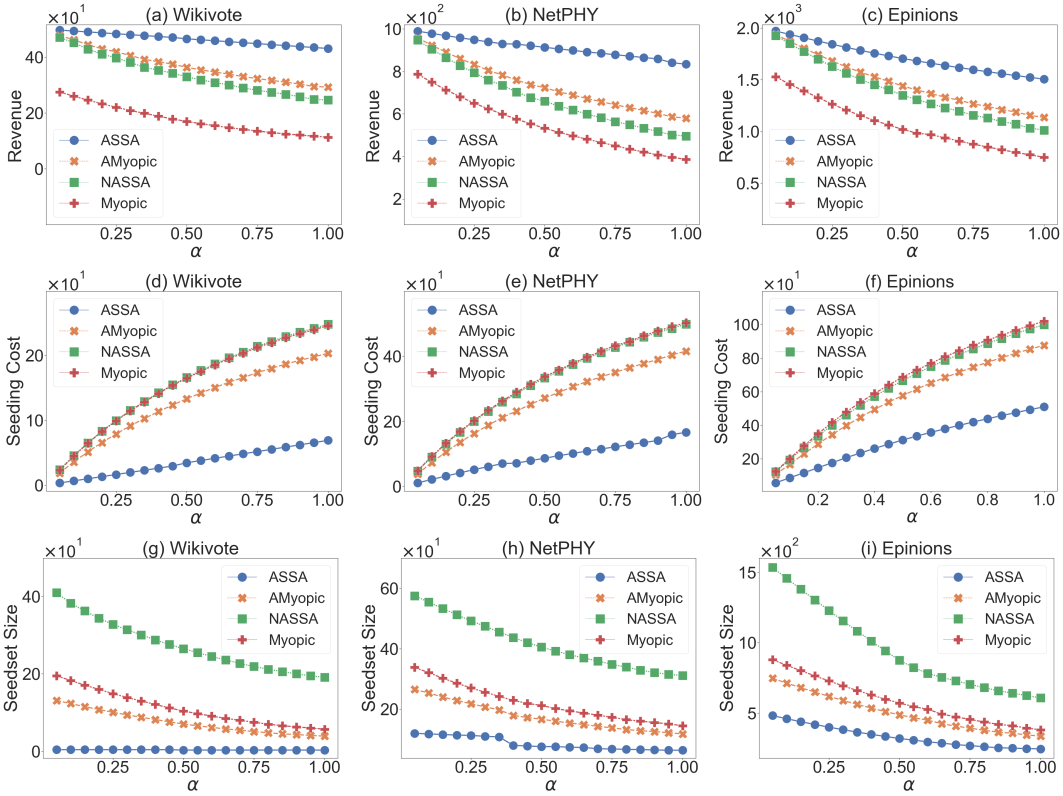

Figure 5: ASSA achieves superior revenue with the lowest seeding cost among all four algorithms under the linear incentive model.

similar for datasets *NetHEPT* and *NetPHY*, we report the results on *Wikivote*, *NetPHY* and *Epinions* for all algorithms under the natural log incentive model (Figure 3) and the linear incentive model (Figure 5). We report the results on all

Dataset	Algorithm	Seeding Cost	Seedset Size	RoR	Exp. Revenue
Wikivote	ASSA	5.3252	1	2.1041	494.6747
	NASSA	5.3252	1	2.1041	494.6747
	AMyopic	0.1452	2	0.9126	224.6579
	Myopic	0.1489	2	0.8921	215.7456
NetHEPT	ASSA	22.9150	143	0.956	477.0849
	NASSA	37.3939	243	0.9027	450.3408
	AMyopic	33.8678	176	0.9460	454.9986
	Myopic	59.3261	191	0.9032	402.8399
NetPHY	ASSA	32.1797	205	0.968	967.8202
	NASSA	49.1644	343	0.9193	927.3450
	AMyopic	44.5829	232	0.9629	932.3742
	Myopic	87.9435	237	0.9199	879.3034
Epinions	ASSA	150.7687	1003	0.9314	1861.2253
	NASSA	174.9430	1138	0.9155	1817.3458
	AMyopic	172.1141	1019	0.9211	1819.7628
	Myopic	195.5282	1024	0.9168	1642.4061

Table 1: Performance under Random Cost Model

four datasets under the random cost model in Table 1.

As shown in Figure 3(a)-(c), the revenue generated by the algorithms decreases as α increases on all datasets. For *Wikivote*, we observe that under natural log incentive model, all algorithms output a seedset of a single node on *Wikivote*, as shown in Figure 3(g). We find that *Wikivote* has a relatively dense network structure, where a single node can trigger a large number of engagements (even over the budget) in the network. In this case, ASSA selects the same single node as NASSA does, no observations of actual spread will be made since the budget is exhausted (by the engagements triggered) after the first node has been selected. Therefore, ASSA and NASSA yield the same revenue and seeding cost. On the other hand, Myopic and AMyopic select a different single node, since they rule out the nodes, whose expected spread is over the budget, from the candidate set at the very beginning. Then they select a node with the highest expected spread to total cost ratio from the reduced candidate set. It shows in Figure 3(a) that ASSA and NASSA outperform the benchmarks on Wikivote. While the benchmarks yield an expected revenue around 190, ASSA and NASSA both yield an expected revenue over 490, a 157% increase. The increase of α leads to a slightly higher node cost, resulting in a subtle decrease in the revenue. This result validates the superiority of our proposed algorithms over the benchmarks.

To reveal the details behind the superiority of our approaches, we plot in Figure 4 the spread yielded by NASSA and Myopic, respectively, on 200 samples randomly selected from the 10^5 Monte Carlo simulations on different datasets. Note that in the figures we also include the data points with maximum and minimum spread from all 10^5 simulations. We abbreviate the natural log node cost model, linear node cost model and random node cost model as cm1, cm2 and cm3, respectively. Figure 4(a) illustrates the results on *Wikivote* under natural log node cost model with $\alpha = 0.5$. We observe that while NASSA outputs a seed node that triggers around 1000 engagements on all samples, the seed node produced by Myopic yields very low engagements in some samples. This is because NASSA selects a node that triggers a large spread on each sample $\phi \in \Phi$, by maximizing the expected value of $\min\{g(S, \Phi), B - c(S)\}$. Myopic, however, selects a node with the highest *expected* spread on all samples without looking into spread on each individual sample. The low engagements on those individual samples are the culprit of the low overall revenue. Figure 4(b)-(d) plot the spread on *NetHEPT* under three node cost models with $\alpha = 1$ for cm1 and cm2. We observe similar patterns on the structure of NASSA and Myopic on these datasets. In particular, the spread triggered by the seed set of Myopic is more scattered than that of NASSA, leading to a lower yielded revenue. This result validates the efficacy of our approaches on selecting high-quality seed nodes that yield high spread among all samples, leading to an improved revenue.

We illustrate in Figure 3(b) and (c) the expected revenue yielded by all four algorithms on *NetPHY* and *Epinions*. We observe that ASSA outperforms the adaptive benchmark AMyopic, and the performance gap becomes larger as the cost of nodes increases. In our experiments, we observe that ASSA outperforms its non-adaptive counterpart NASSA on all datasets. Moreover, NASSA outperforms the non-adaptive benchmark Myopic by at least 20% in terms of expected revenue under all test settings. This result confirms the effectiveness of our proposed algorithms. It also validates the power of adaptiveness in selecting a high quality set of seed nodes to maximize the revenue.

Figure 3(d)-(f) present the results on the seeding cost produced by all algorithms on different datasets with respect to changes in the value of α. As expected, as α increases, the seeding cost increases accordingly. We observe that ASSA achieves superior revenue with the lowest seeding cost among all four algorithms. It verifies that ASSA performs the best on allocating the budget for seeding and generated engagements for revenue maximization. Figure 3(g)-(i) illustrate the results on the size of seed set produced by all algorithms with respect to changes in the value of α. As expected, as α increases, the size of seed set decreases. We observe that among all four algorithms, ASSA yields the highest revenue with the smallest seed set. This again shows the superiority of ASSA on selecting high-quality seed nodes for revenue maximization.

We report our results under the linear incentive model in Figure 5. We observe similar patterns on the structure of the algorithms. In particular, ASSA outperforms the adaptive benchmark AMyopic and its non-adaptive counterpart NASSA in terms of yielded revenue on all datasets. Moreover, NASSA outperforms the non-adaptive benchmark Myopic by at least 20% in terms of revenue under all test settings. Furthermore, we report our results under the random incentive model in Table 1. We define the rate of return (RoR) of a seedset S as $g_{exp}(S)/B$. The value of RoR captures the level of profitability from the advertisers' perspective. A higher RoR indicates that the corresponding algorithm generates a larger number of user engagement in expectation, ie. more profitable for the advertisers. Again, we observe that ASSA achieves higher revenue and higher RoR on all datasets compared with AMyopic and NASSA. We also observe that ASSA and NASSA both convert over 90% of the budget into revenue on all datasets.

References

Abbassi, Z.; Bhaskara, A.; and Misra, V. 2015. Optimizing display advertising in online social networks. In *Proceedings of the 24th International Conference on World Wide Web*, 1–11.

Alon, N.; Gamzu, I.; and Tennenholtz, M. 2012. Optimizing budget allocation among channels and influencers. In *Proceedings of the 21st international conference on World Wide Web*, 381–388.

Alperin, J. P.; Hanson, E. W.; Shores, K.; and Haustein, S. 2017. Twitter bot surveys: A discrete choice experiment to increase response rates. In *Proceedings of the 8th International Conference on Social Media & Society*, 1–4.

Aslay, C.; Bonchi, F.; Lakshmanan, L. V.; and Lu, W. 2016. Revenue maximization in incentivized social advertising. *arXiv preprint arXiv:1612.00531*.

Aslay, C.; Lu, W.; Bonchi, F.; Goyal, A.; and Lakshmanan, L. V. 2014. Viral marketing meets social advertising: Ad allocation with minimum regret. *arXiv preprint arXiv:1412.1462*.

Bakshy, E.; Eckles, D.; Yan, R.; and Rosenn, I. 2012. Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM EC*, 146–161. ACM.

Brady, E.; Morris, M. R.; and Bigham, J. P. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1055–1064.

Chalermsook, P.; Das Sarma, A.; Lall, A.; and Nanongkai, D. 2015. Social network monetization via sponsored viral marketing. *ACM SIGMETRICS Performance Evaluation Review*, 43(1): 259–270.

Feldman, M. 2020. Guess free maximization of submodular and linear sums. *Algorithmica*, 1–26.

Golovin, D.; and Krause, A. 2011. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42: 427–486.

Grau, P.; Naderi, B.; and Kim, J. 2018. Personalized motivation-supportive messages for increasing participation in crowd-civic systems. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–22.

Han, K.; Wu, B.; Tang, J.; Cui, S.; Aslay, C.; and Lakshmanan, L. V. 2021. Efficient and Effective Algorithms for Revenue Maximization in Social Advertising. In *Proceedings of the 2021 International Conference on Management of Data*, 671–684.

Harshaw, C.; Feldman, M.; Ward, J.; and Karbasi, A. 2019. Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. In *International Conference on Machine Learning*, 2634–2643. PMLR.

Kazemi, E.; Minaee, S.; Feldman, M.; and Karbasi, A. 2020. Regularized Submodular Maximization at Scale. *arXiv preprint arXiv:2002.03503*.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Krause, A.; and Golovin, D. 2014. Submodular function maximization. *Tractability*, 3: 71–104.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions-I. *Mathematical programming*, 14(1): 265–294.

Savage, S.; Monroy-Hernandez, A.; and Höllerer, T. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 813–822.

Sviridenko, M.; Vondrák, J.; and Ward, J. 2017. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4): 1197–1218.

Tang, S.; and Yuan, J. 2016. Optimizing ad allocation in social advertising. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1383–1392.

Tang, S.; and Yuan, J. 2020. Influence maximization with partial feedback. *Operations Research Letters*, 48(1): 24–28.

Tang, S.; and Yuan, J. 2021. Submodular Optimization Beyond Nonnegativity: Adaptive Seed Selection in Incentivized Social Advertising. *arXiv preprint arXiv:2109.15180*.

Tucker, C. 2012. Social advertising. *Available at SSRN 1975897*.

Paper Checklist

1. For most authors...
   - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A
   - (e) Did you describe the limitations of your work? Yes
   - (f) Did you discuss any potential negative societal impacts of your work? N/A
   - (g) Did you discuss any potential misuse of your work? N/A
   - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? N/A
   - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   - (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   - (b) Have you provided justifications for all theoretical results? N/A
   - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   - (e) Did you address potential biases or limitations in your theoretical framework? N/A
   - (f) Have you related your theoretical results to the existing literature in social science? N/A
   - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   - (a) Did you state the full set of assumptions of all theoretical results? Yes
   - (b) Did you include complete proofs of all theoretical results? Yes
4. Additionally, if you ran machine learning experiments...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   - (a) If your work uses existing assets, did you cite the creators? N/A
   - (b) Did you mention the license of the assets? N/A
   - (c) Did you include any new assets in the supplemental material or as a URL? N/A
   - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A
   - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A
   - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A
   - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? N/A
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   - (a) Did you include the full text of instructions given to participants and screenshots? N/A
   - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A
   - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A
   - (d) Did you discuss how data is stored, shared, and deidentified? N/A