Co-Activity Maximization in Online Social Networks

Dongyu Mao, Weili Wu, *Senior Member, IEEE*, and Ding-Zhu Du

***Abstract*— Social media with online social networks has risen to be a prevalent force in information diffusion and public discourse. Despite its popularity and convenience, social media has been criticized for contributing to societal and ideological polarization as the result of trapping users in an echo chamber and filter bubbles. An emerging line of research focuses on ways to redesign content or link recommendation algorithms to mitigate the polarization phenomenon. However, existing works mainly concentrate on node-level balancing, while omitting the balancing effect that can be incurred by edge interaction in social networks. In this article, we take the first step to study the problem (CoAM) that assuming two campaigns are present in a network, how we should select seeds for each so as to maximize the interaction/activity between the followers of two campaigns (co-activity) after the diffusion process is finished. We begin our analysis by showing the hardness of CoAM under two diffusion models that are generalized from wildly used diffusion models and its objective function is neither submodular nor supermodular. This encourages us to design a submodular function that acts as a lower bound to the objective, by exploiting which we are able to devise a greedy algorithm with a provable approximation guarantee. To overcome the #P-hardness of diffusion calculation, we further extend the notion of random reverse-reachable (RR) set to devise a scalable instantiation of our approximation algorithm. We experimentally demonstrate the quality of our approximation algorithm on datasets collected from real-world social networks.**

***Index Terms*— Approximation algorithm, echo chamber, filter bubbles, martingale, social network.**

I. Introduction

WITH undoubtedly a large number of advantages, social media and social network service have gone popular enormously in the last decades, as shown in a recent survey that 71% of U.S. adults get news on social media in 2020 [1]. Together with its big success, social media service is under blame for its possible linkage to the increase of societal and ideologically polarization [2], [3], [4]. The criticism here mainly goes that the combination of the viral nature of information propagation and personal-curated content recommendation algorithms used by social media platforms will create and amplify the phenomenon of echo chamber [5], [6] and filter bubble [7], [8] on a social network.

The echo chamber is a phenomenon where users' information exposures are dominated by like-minded individuals, similar opinions or views are shared and can bounce off each other which will eventually reinforce users' own voices, causing it more difficult for individuals to understand opposing viewpoints. A filter bubble is a space where recommendation algorithms used by social media platforms that are trained based on users' previous online behaviors, such as searching, likes, shares, and interactions history, only present personalized contents that agree with ones' interests or viewpoints in the feeds. Echo chambers and filter bubble are even more harmful when considering the presence of misinformation [9], [10], [11].

Considering that many controversial issues, for example, the 2016 U.S. Presidential Election [12], the EU referendum in the U.K. [13], and the COVID-19 Pandemic [14], have stirred fierce debates in the online world, and those debates usually accompanied with spreading of fake or extremely biased news, many researchers have realized the importance of diversifying users' information exposure for fighting against polarization in social network and conducted related studies [15], [16], [17], [18], [19], [20].

The effect of an echo chamber can be weakened by adding different voices to users' chambers, i.e., diversifying users' friend lists/clusters so that each person can enjoy a higher probability of hearing belief-challenging opinions. Popular studies [15], [16], [17] in this line consider using link recommendation or link weight adjustment to bridge chambers to reduce the overall degree of polarization in a social network. Algorithmic curation and personalized recommendation are designed to increase metrics like user engagement or ad revenue, but at the same time, create an echo chamber and trap the user into a filter bubble [8]. In order to break those bubbles, we should change the design of recommendation algorithms in a manner that makes them value variety more, especially for controversial issues. Recent studies consider using direct recommendation [15], [17] or information propagation method [18], [19], [20] to increase users' likelihood of encountering ideologically cross-cutting news content.

Inspired by the ideas mentioned above, in this article, we take a step in this direction and study the problem of breaking filter bubbles in a social network by maximizing the total strength of inter-group contact (co-activity) under the stochastic information propagation model. Specifically, we consider the condition where two opposing campaigns

Manuscript received 30 June 2022; revised 26 August 2022; accepted 30 September 2022. This work was supported in part by the National Science Foundation under Grant 1822985 and Grant 1907472. *(Corresponding author: Dongyu Mao.)*

The authors are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: maody@utdallas.edu; weiliwu@utdallas.edu; dzdu@utdallas.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSS.2022.3213260, provided by the authors.

Digital Object Identifier 10.1109/TCSS.2022.3213260

around one controversial issue are propagating on a social network simultaneously, for example, the U.S. Presidential Election, and the objective is to maximize the co-activity between the followers of two opposing campaigns when the diffusion process finished by allocating seeds to each campaign in the beginning. Here, co-activity occurs when two ends (users) of an edge (denotes friendship or following relation) hold a pair of opposing opinions toward one controversial issue. The inherent rationality of our idea is based on the studies [21], [22], [23] showing contact between opposing groups/individuals helps alleviate group polarization and increase the likelihood of deliberation and political compromise.

Note that our work differs from previous works as follows. First, while several previous works [18], [19], [20] consider using stochastic information propagation model and seed allocation for diversifying, their aim is diversifying the exposure of nodes, our work considers diversifying the exposure of edges. Second, while several previous works [16], [17] consider the differences in edge weights and propose methods like link suggestion and link weight adjustment to fight against polarization, their approaches are essentially built on opinion dynamics model.

Technically, we consider the following problem setting. We assume two campaigns of a controversial issue are propagating in the network according to a specific propagation model independently. Each campaign is associated with a seed budget, within which we can choose initial seed nodes for campaign propagation. Each edge is associated with a weight denoting activity strength through it. The objective is to recruit initial seeds for the two competing campaigns within their budgets, such that the total activity strength between followers of the two campaigns when diffusion process is finished is maximized.

Although derived from a large volume of work on information propagation and breaking filter bubbles, by combining the merits of using stochastic propagation and edge weight, our paper shows the following significant differences and novelties.

1) This is the first paper trying to diversify information exposure in edge level by maximizing co-activity for breaking filter bubbles using information propagation method.
2) The problem of maximizing co-activity (CoAM) is formally defined, after which its hardness and approximability, the properties of objective function are studied.
3) Due to the non-sub/super-modularity of the objective function, its submodular lower bound is devised and based on which an algorithm with approximation guarantee is provided.
4) The quality of the proposed approximation algorithm is evaluated on dataset extracted from real-world social network.

The remaining part of this article is organized as follows. Section II briefly introduces related works concerning balancing information exposure and breaking filter bubbles on social networks. Section III formulates the diffusion models and formally defines the CoAM problem. Section IV shows the hardness result, the approximability, and properties of the objective function of CoAM problem, and then a lower bound of the objective, with which we propose an approximation algorithm of CoAM, is introduced and analyzed. Section V gives an efficient implementation of the proposed approximation algorithm. Section VI is dedicated to show experiments and Section VII concludes this article. Note that all proofs and several tables are shown in Appendixes in supplementary material available online.

II. Related Works

With the growing popularity of social media [1], online polarization receives ascending attention from researchers in many fields, as this polarization is observed to link to society across many issues in politics [4], [12], [13], public policy [24], and healthcare [14], [25]. Our work belongs to an emerging line of research on fighting against online polarization by breaking filter bubbles. Specifically, we consider maximizing the strength of contact between opposing groups which is shown to be beneficial for reaching compromise [21], [22], [23]. Many studies have been done on the effect of echo chamber [5], [6] and filter bubble [7], [8] that may contribute to polarization. It is shown that opinion-challenging information spread less than others [6] and content filtering by social media platform for higher user engagement can increase polarization significantly [16], which impels social media companies to highlight tradeoff between revenue and polarization. Popular approaches consider using opposing content recommendation [17], [18], [19], [20], [26], [27], [28], or bridging opposing views by link suggestion [15], [16], [17] to diversify users' exposure and reduce polarization.

Among the works mentioned, closest to ours lies in a line of research using stochastic information propagation method to balance or diversify node-level information exposure in social network [18], [19], [20]. Garimella et al. [18] consider balancing users' information exposure by recruiting additional seeds for two opposing campaigns so that the number of nodes accepts either both or none of the campaigns when diffusion finished is maximized. Within the algorithms they proposed, only one provides an approximation guarantee for campaigns with different propagation statistics, while others rely on limited assumptions like forcing additional seeds selected to be the same or campaigns sharing common propagation statistics. Aslay et al. [20] consider the leanings of users and news articles propagating through the network and formulate the problem of diversifying users' information exposure by recommending articles to selected users. Tu et al. [19] consider maximizing the number of users that accept both sides of two opposing campaigns. Technically, it is the closest one to our work. However, as our work focuses on edge-level balancing, the sampling design is totally different. The other line of research close to ours is based on opinion dynamics that model social learning process [16], [17]. For example, Musco et al. [17] quantify both the disagreement and polarization and try to optimize their sum by graph topological optimization or content recommendation under Friedkin–Johnsen model [29].

More broadly, our work relates to a series of research on viral marketing for multiple items in online social networks via information propagation lens [30], [31], [32], [33], [34], [35], [36]. While in these works the adoption of items by users is considered to contribute to the final revenue, the connection strength through edges is also considered as an important factor in several variant problems of influence maximization (IM) and misinformation containment [37], [38], [39]. Technically, we borrow the concept of reverse influence sampling (RIS) [40] for overcoming the #P-hardness of influence calculation [41], [42] in the design of our sampling algorithm. A series of more efficient and scalable randomized algorithms for IM are proposed based on RIS, including TIM/TIM+ [43] and IMM [44]. Specifically, we follow the martingale analysis used in IMM in our algorithm design.

III. Problem Definition

In the following, we describe the information propagation models (Section III-B) and formally define the CoAM problem by using possible-world semantics (Section III-C). A table listing the frequently used notations is shown in Appendix B.

A. Preliminaries

Inputs to our problem consist of the following.

1) A directed graph $G = (V, E)$ with $|V| = n$ nodes and $|E| = m$ edges, where a directed edge (u, v) means v follows u so v can see posts of u.
2) Two opposing campaigns for a controversial issue, e.g., two candidates for an election, referred as campaign r (*red*) and campaign b (*blue*), with their seed set budgets $k_r \in \mathbb{Z}_+$ and $k_b \in \mathbb{Z}_+$, respectively.
3) Each $(u, v) \in E$ is associated with a activity strength, denoted as $a_{(u,v)} \geq 0$.
4) Campaign-specific propagation model and parameters.

B. Propagation Models

First, assume the propagation of one campaign is independent of the propagation of the other, then consider two specific models.

1) Each campaign follows the independent cascade model [45]. Specifically, for each $i \in \{r, b\}$, the propagation probability of i through an edge (u, v) is $p^i_{(u,v)}$. In each discrete time step t, a newly i-activated node u has a single chance to propagate its influence to node v that is not i-activated through $(u, v) \in E$ with probability $p^i_{(u,v)}$, if succeed, v will be i-active in time $t + 1$.
2) Each campaign follows the linear threshold (LT) model [45]. Specifically, for each $i \in \{r, b\}$, each edge (u, v) is associated with an influence weight $b^i_{(u,v)}$, each node $v \in V$ selects a threshold θ^i_v in $[0, 1]$ uniformly random and $\sum_{u \in N^-(v)} b^i_{(u,v)} \leq 1$, $N^-(v) = \{u | (u, v) \in E\}$ is the set of incoming neighbor of v in E. In each discrete time step t, a i-inactive node v checks the sum of influence weights from its i-active neighbors and change its state to i-active in time $t + 1$ only if this sum exceeds θ^i_v.

As each campaign is associated with a different color (*red or blue*), denote: 1) as multicolor independent cascade (MCIC) model and 2) as multicolor LT (MCLT) model. MCIC is used in related works [18], [19], [20].

C. Possible-World Semantics

Given two seed sets S_r and S_b for campaign r and b, respectively, a single possible world w represents an outcome of the stochastic propagation process starting from S_r and S_b in network G. We follow previous works [19], [20] to use an edge-colored multigraph representation for formulating the probability of a possible world w. Accordingly, for MCIC model, define a directed multigraph $\tilde{G} = (V, \tilde{E}, \tilde{p})$ from $G = (V, E)$, specifically, for each $(u, v) \in E$, creating a parallel edge $(u, v)^i$ with probability $p^i_{(u,v)}$ for each $i \in \{r, b\}$, then the probability of a possible world w by sampling $(u, v)^i$ from $\tilde{G}$ can be represented as

$$\Pr[w] = \prod_{i \in \{r,b\}} \prod_{(u,v)^i \in w} p^i_{(u,v)} \prod_{(u,v)^i \in \tilde{E} \setminus w} \left(1 - p^i_{(u,v)}\right).$$

For MCLT model, we can also define a similar *directed multigraph* $\tilde{G} = (V, \tilde{E}, \tilde{b})$ from $G = (V, E)$ by creating parallel edges with campaign-specific influence weights and assign each node campaign-specific thresholds; similarly, by the equivalent view (Claim 2.6) established in Kempe et al. [45], we can view the process of generating a possible world w as following: v chooses at most one of $u \in N^-(v)$ s.t. $(u, v)^i$ appears in w with probability $b^i_{(u,v)}$ exclusively, and there is no incoming edge of v in w for campaign i with probability $1 - \sum_{u \in N^-(v)} b^i_{(u,v)}$, denote $\bar{V}^i = \{v : \nexists (u, v)^i \in w\}$, then

$$\Pr[w] = \prod_{i \in \{r,b\}} \prod_{(u,v)^i \in w} b^i_{(u,v)} \prod_{v \in \bar{V}^i} \left(1 - \sum_{u \in N^-(v)} b^i_{(u,v)}\right).$$

Note that w is a "*hard-wired*" graph, denote $I_w(S_r)/I_w(S_b)$ as the set of nodes reachable from seed set S_r/S_b in w, let $\mathrm{CO}_w(S_r, S_b) = \{(u, v) | (u, v) \in E \wedge (u \in I_w(S_r) \wedge v \in I_w(S_b) \vee u \in I_w(S_b) \wedge v \in I_w(S_r))\}$ denote set of edges that can pass "co-activity" here "passing co-activity" means an edge may contribute to balance/diversify the information exposure and break the filter bubbles, thus even if both ends are activated by both campaign, we still count it in CO_w), then define $A_w(S_r, S_b) = \sum_{(u,v) \in \mathrm{CO}_w(S_r, S_b)} a_{(u,v)}$ as the total co-activity in w by seeding S_r and S_b, we have

$$\mathbb{E}[A(S_r, S_b)] = \sum_{w \sqsubseteq \tilde{G}} \Pr[w] A_w(S_r, S_b).$$

Note that we allow $S_r \cap S_b \neq \emptyset$. Then we are ready to formally define CoAM.

Problem 1 [Co-Activity Maximization (CoAM)]: Given a directed social graph $G = (V, E)$, two opposing campaigns r and b, a propagation model MCIC/MCLT with campaign-specific propagation parameters, initial seed set budgets $k_r \in \mathbb{Z}_+$ and $k_b \in \mathbb{Z}_+$, find seed sets S_r and S_b such that $|S_r| \leq k_r$ and $|S_b| \leq k_b$ and the expected co-activity is maximized when propagation process finished

$$\max_{S_r \subseteq V, S_b \subseteq V} \mathbb{E}[A(S_r, S_b)], \quad \text{s.t. } |S_r| \leq k_r, |S_b| \leq k_b.$$

IV. Theoretical Analysis

In this section, we first analyze the complexity of CoAM (Section IV-A), and then properties of its object function are studied, with which a submodular lower bound of the objective is devised for approximation algorithm design (Section IV-B). We start by establishing hardness results.

A. Hardness

Theorem 1: CoAM problem is NP-hard under MCIC/MCLT model.

Theorem 2: Under MCIC/MCLT model: 1) CoAM does not admit a PTAS unless P = NP and 2) in terms of parameterized complexity, CoAM is not FPT with respect to k_r even assuming $k_b = |V|$ or vice versa.

Theorems 1 and 2 show the difficulties of directly optimizing and approximating the CoAM problem, respectively, which encourages us to design approximation algorithm by analyzing the properties of its objective $\mathbb{E}[A]$.

B. Approximation

First, notice that $\mathbb{E}[A(S_r, S_b)]$ is a *biset* function and IM with a single seed set under IC/LT model is *monotone nondecreasing submodular* [45]. If we can link $\mathbb{E}[A(S_r, S_b)]$ with *bisubmodularity* [46], then we can use *bisubmodular maximization* technique to deal with it. However, Lemma 1 negates this way.

Lemma 1: $\mathbb{E}[A(S_r, S_b)] : 2^V \times 2^V \to \mathbb{R}_{\geq 0}$ is a monotone non-decreasing biset function but not bisubmodular under MCIC/MCLT model.

As optimizing $\mathbb{E}[A]$ in *biset* form seems groundless, inspired by univariate transformation used in Tu et al. [19], in the rest of this section, we continue our analysis by building an equivalent univariate formulation of $\mathbb{E}[A]$.

Let (S_r^*, S_b^*) denote the optimal seed sets that maximize co-activity $\mathbb{E}[A]$, as $\mathbb{E}[A]$ is a monotone nondecreasing *biset* function (Lemma 1), we can attain the maximal objective when $|S_r^*| = k_r$ and $|S_b^*| = k_b$. Without loss of generality, assume $k_r \leq k_b$ in the following analysis. Let $O_1 = \{(S_r, S_b) || S_r| = k_r, |S_b| = k_b, S_r \subseteq V, S_b \subseteq V\}$ denote the set of feasible seed set pairs of maximal size, we have $(S_r^*, S_b^*) \in O_1$. For any $(S_r, S_b) \in O_1$, it follows that $1 \leq |S_b|/|S_r| \leq \lceil k_b/k_r \rceil$. This inspires us to construct a set of pairings between nodes in S_r and S_b such that each node in S_r corresponding to at least $\lfloor k_b/k_r \rfloor$ and at most $\lceil k_b/k_r \rceil$ nodes. Denote $\mathcal{E} = \{V \times V\}$ as the ground set of all ordered node pairs, where $(u, v) \in \mathcal{E}$ represents the pairing of a node u, selected as a seed node for campaign r, with a node v, selected as a seed node for campaign b. We can define a *set-of-pairs* system on $\mathcal{E}$, and establish its relation to O_1 (Lemma 2).

Definition 1 (Set-of-Pairs System): Let $(\mathcal{E}, \mathcal{I})$ be a set system where $\mathcal{E} = \{V \times V\}$ is the ground set and $\mathcal{I}$ is a collection of subsets of $\mathcal{E}$. For any $Y \in \mathcal{I}$, let $Y_r = \bigcup\{r|(r, b) \in Y\}$ and let $Y_b = \bigcup\{b|(r, b) \in Y\}$. We say that $(\mathcal{E}, \mathcal{I})$ is a *set-of-pairs* system iff for any set $Y \in \mathcal{I}$, the following conditions hold.

1) $|Y_r| \leq k_r$.
2) $|Y_b| = |Y| \leq k_b$.
3) For each $r_0 \in Y$, $|\bigcup\{b|(r_0, b) \in Y\}| \leq \lceil k_b/k_r \rceil$.
4) If $\lfloor k_b/k_r \rfloor < \lceil k_b/k_r \rceil$, then $|\{r|(r, b) \in Y \wedge |\bigcup\{b|(r, b) \in Y\}| = \lceil k_b/k_r \rceil\}| \leq k_b \mod k_r$ (pairing Y_r and Y_b as evenly as possible for the case $|Y_r| = k_r$ and $|Y_b| = k_b$).

Lemma 2: Let $O_2 = \{(Y_r, Y_b)|Y \in \mathcal{I}\}$, then $O_1 \subseteq O_2$.

Let $f : 2^{\mathcal{E}} \to \mathbb{R}_{\geq 0}$ be a function defined as $f(Y) = \sum_{(u,v)\in \mathrm{CO}_{\mathrm{CoAM}}(Y)} a_{(u,v)}$ where $\mathrm{CO}_{\mathrm{CoAM}}(Y) = \{(u, v)|(u, v) \in E \wedge (u \in I(Y_r) \wedge v \in I(Y_b) \vee u \in I(Y_b) \wedge v \in I(Y_r))\}$, $I(Y_r)$ and $I(Y_b)$ are random variables representing the set of nodes infected by campaign r and b when propagation finishes, respectively. Note that $I(\cdot)$ is *monotone nondecreasing submodular* under IC and LT model [45]. By Lemma 2, for an optimal pair $(S_r^*, S_b^*) \in O_1$ that maximize $\mathbb{E}[A]$, we can construct a corresponding $Y^* \in \mathcal{I}$ that maximize f, thus CoAM problem can be reformulated as

$$\max_{Y\in\mathcal{I}} \mathbb{E}[f(Y)]$$

converting a *two variables* (S_r, S_b) problem to a *single variable* (Y) problem. Next, we proceed by analyzing the properties of f.

Lemma 3: $f : 2^{\mathcal{E}} \to \mathbb{R}_{\geq 0}$ is a monotone nondecreasing set function, which is neither submodular nor supermodular under MCIC/MCLT model, besides, its submodularity ratio is 0.

Again, Lemma 3 shows that directly approximate f is hard even under cardinality constraint, as existing approximation techniques for submodular set function and non-submodular set function with proper submodular ratio (>0) and curvature [47] turn out to be not suitable here. In addition, it's easy to see that this problem does not follow cardinality constraint, we should also analyze the properties of $(\mathcal{E}, \mathcal{I})$ to deal with it.

Using submodular upper/lower bound for nonsubmodular function maximization is a widely used technique [19], [37], [38], [39]. Specifically, the nonsubmodularity of f comes from "combination effect" between activated nodes by different seed pairs, which is given more discussion in several works [37], [48]. Then, we can design a function $g : 2^{\mathcal{E}} \to \mathbb{R}_{\geq 0}$, with $g(Y) = \sum_{(u,v)\in \mathrm{CO}_{\mathrm{lower}}(Y)} a_{(u,v)}$, and $\mathrm{CO}_{\mathrm{lower}}(Y) = \{(u, v)|(u, v) \in E \wedge (r, b) \in Y \wedge (u \in I(r) \wedge v \in I(b) \vee u \in I(b) \wedge v \in I(r))\}$, compared to $\mathrm{CO}_{\mathrm{CoAM}}(Y)$, here in $\mathrm{CO}_{\mathrm{lower}}(Y)$, only edges that can be balanced by a pair of seeds in Y are included, thus $f(Y) \geq g(Y)$ strictly. It can be shown that $g(Y)$ is submodular (Lemma 4), and differs from $f(Y)$ within a multipicative factor in any possible world ω generated by stochastic propagation process (Lemma 5).

Lemma 4: $g : 2^{\mathcal{E}} \to \mathbb{R}_{\geq 0}$ is a monotone nondecreasing submodular set function, maximizing g is NP-hard.

Lemma 5: Let $Y^0 = \arg\max_{Y\in\mathcal{I}} \mathbb{E}[g(Y)]$, $Y^* = \arg\max_{Y\in\mathcal{I}} \mathbb{E}[f(Y)]$, $k_r \leq k_b$, then $\mathbb{E}[f(Y^*)] \leq k_r \mathbb{E}[g(Y^0)]$.

Lemma 5 suggests that any algorithm that provides an approximation guarantee for maximizing $\mathbb{E}[g]$ over set system $(\mathcal{E}, \mathcal{I})$, provides also a bounded approximation guarantee for maximizing $\mathbb{E}[f]$. Now it is the turn to analyze properties of the set-of-pairs system $(\mathcal{E}, \mathcal{I})$. We first provide the preliminary definitions (Definitions 2 and 3), which will help us establish the properties (Lemmas 6 and 7).

Definition 2 (Independence System): A set system $(\mathcal{E}, \mathcal{I})$ is an independence system if $\mathcal{I} \neq \emptyset$ and satisfies the downward-closure property, i.e., if $Y \in \mathcal{I}$ and $X \subseteq Y$, then $X \in \mathcal{I}$.

Definition 3 (p-System [49]): An independence system $(\mathcal{E}, \mathcal{I})$ is a p-system, if

$$\max_{X \subseteq \mathcal{E}} \frac{\max_{J:J \text{ is a base of } X} |J|}{\min_{J:J \text{ is a base of } X} |J|} \leq p$$

where any subset J of X is a base of X if $J \in \mathcal{I}$ and $\forall e \in X \backslash J,\ J \cup e \notin \mathcal{I}$.

Lemma 6: The set-of-pairs system $(\mathcal{E}, \mathcal{I})$ is an independence system, but not a matroid.

Lemma 7: The set-of-pairs system $(\mathcal{E}, \mathcal{I})$ is a $4\lceil (k_b/k_r) \rceil$-system.

Thus far, we know $(\mathcal{E}, \mathcal{I})$ is a $4\lceil (k_b/k_r) \rceil$-system and $\mathbb{E}[g]$ is monotone non-decreasing submodular, then greedy algorithm starting from $\emptyset$ will provides an $(1/1+4\lceil (k_b/k_r) \rceil)$-approximation [49]. Denote $Y^G \subseteq \mathcal{E}$ as the solution returned by greedy algorithm, $Y_r^G = \bigcup \{r | (r,b) \in Y^G\}$, $Y_b^G = \bigcup \{b | (r,b) \in Y^G\}$. Then we have (Y_r^G, Y_b^G) is a $1/((1 + 4\lceil (k_b/k_r) \rceil)k_r)$-approximation to optimal solution of CoAM problem.

Theorem 3: Let Y^G be the solution returned by a greedy algorithm running for $\mathbb{E}[g]$ starting from $\emptyset$, then

$$\mathbb{E}\big[f\big(Y^G\big)\big] \geq \frac{1}{\Big(1 + 4\lceil \frac{k_b}{k_r} \rceil\Big)k_r} \mathbb{E}\big[f\big(Y^*\big)\big].$$

Unfortunately, directly using greedy algorithm is unrealistic since evaluating $\mathbb{E}[g(X)]$ for a $X \subseteq \mathcal{E}$ is #P-hard, as shown in Lemma 8. This problem will be tackled by using RIS [40] technique in Section V.

Lemma 8: Given a $Y \subseteq \mathcal{E}$, computing $\mathbb{E}[g(Y)]$ is #P-hard under MCIC/MCLT model.

V. Algorithms

Efficient implementation of a greedy algorithm is challenging since calculating lower bound $\mathbb{E}[g(\cdot)]$ is #P-hard as shown in Lemma 8. Naïvely, using a large number of Monte Carlo (MC) simulations is a choice here [45]. Consider using r rounds MC simulations, the time complexity of the greedy algorithm can be $\mathcal{O}(k_b n^2 m r)$. Clearly, it is prohibitively expensive, making it difficult to compromise between efficiency and accuracy.

On the other hand, among the efforts devoted to scalable IM, Borgs et al. [40] first proposed the concept of RIS and designed a quasi-linear time randomized algorithm based on reverse-reachable (RR) sets with approximation guarantee. Using RIS technique, Tang et al. designed TIM/TIM+ Algorithm [43] that achieve near-optimal time complexity and subsequently proposed IMM Algorithm [44] which further improved the performance using martingale based analysis.

Random RR sets are a vital tool for estimating expected influence spread used in IM algorithms mentioned above, and its variants are shown useful in many related tasks [19], [20], [37], [38], [39]. Considering the inherent similarity between computing influence spread and $\mathbb{E}[g(\cdot)]$, in this section, we first introduce a non-trivial generalization of RR set, named as *edge RR pairs (ER^{2}P) set*, based on which we devise an unbiased estimator of $\mathbb{E}[g(\cdot)]$ (Section V-A). Then we focus on essentially how many ER^{2}P sets are needed in our approximation algorithm (Sections V-B and V-C).

A. Lower Bound Estimator

Definition 4 (Random edge RR pairs (ER^{2}P) set): Let $w \sqsubseteq \tilde{G}$ be any possible world, $B = \sum_{(u,v) \in G} a_{(u,v)}$ denotes the summed activity strength of edges in G, $R^T_{w-r}(v)/R^T_{w-b}(v)$ denotes the RR set [43] for a node v in $\tilde{G}$ that can propagate campaign r/b to it in a possible world w. A ER^{2}P set can be generated by the following steps.

1) Generate a possible world w of $\tilde{G}$.
2) Select an edge $(u, v) \in E$ with probability $a_{(u,v)}/B$.
3) Collect $R^T_{w-r}(u)$, $R^T_{w-b}(v)$, $R^T_{w-r}(v)$, $R^T_{w-b}(u)$.
4) Calculate $R1_{w,(u,v)} = \{(r,b)\colon r \in R^T_{w-r}(u) \wedge b \in R^T_{w-b}(v)\}$, $R2_{w,(u,v)} = \{(r,b)\colon r \in R^T_{w-r}(v) \wedge b \in R^T_{w-b}(u)\}$.
5) Return $R_{w,(u,v)} = R1_{w,(u,v)} \cup R2_{w,(u,v)}$.

The efficiency of sampling an ER^{2}P set can be improved by using randomized *breadth-first search (BFS)* instead of generating a whole possible world w. Let $\mathcal{R}$ denote a pool of random ER^{2}P-sets, let $F_{\mathcal{R}}(Y) = (1/|\mathcal{R}|) \sum_{R \in \mathcal{R}} \mathbb{1}[R \cap Y \neq \emptyset]$ denote the fraction of ER^{2}P-sets that intersect with $Y \subseteq \mathcal{E}$, $\mathbb{1}(\cdot)$ is indicator function. Then we can show that $\mathbb{E}[g(Y)]$ can be estimated using $F_{\mathcal{R}}(Y)$.

Lemma 9: For any $Y \subseteq \mathcal{E}$, we have $\mathbb{E}[g(Y)] = B\mathbb{E}[F_{\mathcal{R}}(Y)]$, the expectation is taken over the randomness in $w \sim \tilde{G}$ and $(u, v) \sim E$.

B. Two-Phase Approximation

Shown that $B\mathbb{E}[F_{\mathcal{R}}(Y)]$ is an unbiased estimator of $\mathbb{E}[g(Y)]$, we can follow previous works [19], [20], [44] to design our two-phase co-activity maximization (TCoAM) algorithm (Algorithm 1) that provides an approximation-guaranteed solution $\tilde{Y}^G$ to the problem of maximizing $\mathbb{E}[g(Y)]$ using a sample $\mathcal{R}$ of random ER^{2}P sets. TCoAM operates in two phases as follows.

1) *Sampling phase*, which determines the size of $\mathcal{R}$ needed for accurately estimating $\mathbb{E}[g(Y)]$ and generates $\mathcal{R}$.
2) *Greedy pair selection phase*, which selects a feasible pair y maximizing marginal gain $F_{\mathcal{R}}(\tilde{Y}^G \cup y) - F_{\mathcal{R}}(\tilde{Y}^G)$ and adds it to $\tilde{Y}^G$ in each iteration.

Algorithm 1 TCoAM($\tilde{G}$, $(\mathcal{E}, \mathcal{I})$, λ, λ^a, ϵ_2, LB_0)

```
R ← Sampling(G̃, (E, I), λ, λ^a, ε_2, LB_0);
Ỹ^G ← ER^2-Pairs-Greedy(R, (E, I));
return Ỹ^G
```

For *greedy pair selection phase* (Algorithm 2), we show in Theorem 4 that if we can obtain accurate-enough estimation of $\mathbb{E}[F_{\mathcal{R}}(Y)]$ for all $Y \in \mathcal{I}_{\text{base}}$, $\mathcal{I}_{\text{base}} \subseteq \mathcal{I}$ is the maximum independent sets in $(\mathcal{E}, \mathcal{I})$ and its size is given in Lemma 10, with high probability given a sample $\mathcal{R}$ of random ER^{2}P sets, then we can approximate CoAM to a factor with high probability. Let $\text{OPT} = \mathbb{E}[g(Y^0)]$.

Algorithm 2 ER^2-Pairs-Greedy($\mathcal{R}$, $(\mathcal{E}, \mathcal{I})$)

$\tilde{Y}^G \leftarrow \emptyset$;
while $y \leftarrow \arg\max_{y:\tilde{Y}^G \cup \{y\} \in \mathcal{I}} F_{\mathcal{R}}(\tilde{Y}^G \cup \{y\}) - F_{\mathcal{R}}(\tilde{Y}^G)$ **do**
  $\tilde{Y}^G \leftarrow \tilde{Y}^G \cup \{y\}$;
return $\tilde{Y}^G$

Theorem 4: Assume in greedy pair section phase TCoAM receives as input a sample $\mathcal{R}$ of random ER^{2}P-sets such that

$$|\mathrm{BF}_{\mathcal{R}}(Y) - \mathbb{E}[(g(Y)]| < \frac{\epsilon}{2}\mathrm{OPT} \tag{1}$$

holds for any $Y \in \mathcal{I}_{\text{base}}$ with probability at least $1 - B^{-\ell}/|\mathcal{I}_{\text{base}}|$, then TCoAM returns a $((1/(1+4\lceil(k_b/k_r)\rceil)k_r) - \epsilon)$-approximate solution to the CoAM problem with probability at least $1 - B^{-\ell}$ and runs in $\mathcal{O}(\sum_{R\in\mathcal{R}}|R|)$.

Lemma 10: The size of $\mathcal{I}_{\text{base}}$ satisfies $|\mathcal{I}_{\text{base}}| = \binom{n}{k_r}\binom{n}{k_b}k_b!$.

Next, we are interested in finding a lower bound of $|\mathcal{R}|$ such that Theorem 4 holds. Specifically, let

$$\lambda = \frac{4B}{\epsilon^2}\Big(\frac{\epsilon}{3} + 2\Big)(\ln 2 + \ell \ln B + \ln |\mathcal{I}_{\text{base}}|)$$

we have the following Lemma 11. The proof of Lemma 11 is based on martingale analysis which has been successfully implemented in related works [19], [20], [44].

Lemma 11: If $|\mathcal{R}| \geq \lambda/\mathrm{OPT}$, then (1) holds for all $X \in \mathcal{I}_{\text{base}}$ with probability at least $1 - B^{-\ell}$.

C. Sampling

However, it is not easy to get this lower bound as maximizing $\mathbb{E}[g(\cdot)]$ is NP-hard (Lemma 4). Alternatively, following existing approaches [19], [20], [44], we try to find a lower bound of OPT and use it to get the required $|\mathcal{R}|$. The key part of our approach includes a statistical test $\mathcal{T}(z)$ such that if $\mathrm{OPT} < z$, then $\mathcal{T}(z) = \text{False}$ holds with high probability. As $\mathrm{OPT} \in [1, B]$, this test is performed iterative on $\mathcal{O}(\log_2 B)$ values of $z = B/2, B/4, \ldots, 1$. Note that we restrict $\mathrm{OPT} \geq 1$, which can be easily set up in preprocessing.

We now explain the workflow of TCoAM's sampling phase (Algorithm 3), which first identifies a lower bound LB of $|\mathcal{R}|$ by implementing a $\mathcal{T}(\cdot)$ adaptively, then it generates a sample of ER^{2}P sets $\mathcal{R}$ such that $|\mathcal{R}| \geq \lambda/\mathrm{LB}$. The algorithm starts by initializing $\mathcal{R}$ to $\emptyset$, LB to LB_0, while LB_0 can be naïvely set to be $\max_{(u,v)\in E} a_{(u,v)}$, and an error parameter ϵ_2. Then it enters for a loop at most $\log_2 B$ times. In the ith iteration, the algorithm compute a $z = B/2^i$ and use z to derive a $\theta_i = \lambda^\alpha/z$, where

$$\lambda^\alpha = \frac{B}{\epsilon_2^2}\left(\frac{2\epsilon_2}{3} + 2\right)\big(\ln \log_2 B + \ell \ln B + \ln |\mathcal{I}_{\text{base}}|\big).$$

The algorithm keeps generating new random ER^{2}P set and add it to $\mathcal{R}$ until $|\mathcal{R}| \geq \theta_i$, then it computes a greedy solution $\tilde{Y}_i^G$ on this new sample $\mathcal{R}$. If the new $\mathcal{R}$ satisfies the condition

$$\mathrm{BF}_{\mathcal{R}}\big(\tilde{Y}_i^G\big) \geq (1+\epsilon_2)z$$

then set $\mathrm{LB} = (\mathrm{BF}_{\mathcal{R}}(\tilde{Y}_i^G)/1+\epsilon_2)$ and break the for loop, otherwise continue with the $(i+1)$th iteration. After the for loop, we regenerate λ/LB random ER^{2}P sets and return them as $\mathcal{R}$. The correctness of the sampling phase is shown by Theorem 5.

Algorithm 3 Sampling($\tilde{G}$, $(\mathcal{E}, \mathcal{I})$, λ, λ^α, ϵ_2, LB_0)

$\mathcal{R} \leftarrow \emptyset$;
$LB \leftarrow LB_0$;
for $i = 1, \cdots, \log_2 B$ **do**
  $z \leftarrow B/2^i$;
  $\theta_i = \lambda^\alpha/z$;
  while $\mathcal{R} \leq \theta_i$ **do**
    $\mathcal{R} \leftarrow \mathcal{R} \cup$ GenerateER^{2}P-Set; `// Follows Definition 4`
  $\tilde{Y}_i^G \leftarrow$ ER^2-Pairs-Greedy;
  if $BF_{\mathcal{R}}(\tilde{Y}_i^G) \geq (1+\epsilon_2)z$ **then**
    $LB \leftarrow \frac{BF_{\mathcal{R}}(\tilde{Y}_i^G)}{1+\epsilon_2}$;
    break;
$\mathcal{R} \leftarrow \emptyset$;
$\theta \leftarrow \lambda/LB$;
while $|\mathcal{R}| \leq \theta$ **do**
  $\mathcal{R} \leftarrow \mathcal{R} \cup$ GenerateER^{2}P-Set;
return $\mathcal{R}$

Theorem 5: With probability at least $1 - B^{-\ell}$, the sampling phase of ToCAM returns a sample $\mathcal{R}$ such that $|\mathcal{R}| \geq \lambda/\mathrm{OPT}$.

With all the above discussions, we have the following Theorem 6.

Theorem 6: TCoAM returns a $((1/(1+4\lceil(k_b/k_r)\rceil)k_r) - \epsilon)$-approximate solution to the CoAM problem with probability at least $1 - 2B^{-\ell}$.

VI. Experiments

In this section, we evaluate the performance of our proposed TCoAM algorithm on several real-world network datasets. We first introduce datasets and experimental settings used in our evaluation, and then experimental results are presented and analyzed.

A. Settings

1) Datasets: The experiments are carried out on five network datasets with varied statistics as shown in Table I. All of five network datasets are publicly available [50], *rt-copen* and *rt-assad* are selected from retweet networks category, *ca-netscience* and *ca-GrQc* are within collaboration networks category and *soc-wiki-Vote* is from social networks category. Note that *ca-netscience* and *ca-GrQc* are undirected in nature, so for each undirected edge we replace it with two directed edges to make the datasets directed.

2) Propagation Models: We consider both MCIC and MCLT model in the experiments. For each dataset, two methods are used to assign propagation-related parameters.

1) *Weighted Cascade:* For each $(u, v) \in E$, set $p^i_{(u,v)} = 1/N^-(v)$ for MCIC model and $b^i_{(u,v)} = 1/N^-(v)$ for MCLT model where $i \in \{r, b\}$; in this case, two campaigns share the same propagation-related parameters.

TABLE I
STATISTICS OF THE DATASETS

Dataset	n	m	Avg. degree
ca-netscience	379	1828	9.65
rt-copen	761	1029	2.70
soc-wiki-Vote	889	2914	6.56
rt-assad	2139	2803	2.62
ca-GrQc	4152	26844	6.47

2) *Random:* For each $(u, v) \in E$ and $i \in \{r, b\}$, $p^i_{(u,v)}$ and $b^i_{(u,v)}$ are firstly chosen from [0, 1] uniformly at random, then we normalize the parameters of all incoming edges of a node v if the sum exceeds 1.0 in order to fulfill the requirements of MCLT model.

Similar methods have been used in related works [19], [37], [38], [39], [41], [42]. For activity strength on each $(u, v) \in E$, we simply set it as $a_{(u,v)} = 1/N^-(v)$. The resulting networks are denoted by adding suffix ".wc" or ".rd" to dataset name, e.g., soc-wiki-Vote.wc. Note that getting exact values of those parameters is orthogonal to the interests of this article.

3) Baselines: Besides TCoAM, we implement four heuristic algorithms as baselines for comparison.

a) *Random:* Randomly select k_r seeds for campaign r and k_b seeds for campaign b.
b) *MaxODeg:* Sort the nodes by descending order of out-degrees, choose the first k_r nodes for seeds of campaign r and the first k_b nodes for seeds of campaign b.
c) *MaxOAct:* Sort the nodes by descending order of out-activity strengths, choose the first k_r nodes for seeds of campaign r and the first k_b nodes for seeds of campaign b.
d) *PageRank:* Use PageRank Algorithm [51] to rank the nodes in G_r and G_b and select the top nodes to seed sets.
e) *TCEM*[1]: A node-level balancing Algorithm [19] tries to find S_r and S_b that maximize the number of node exposed to both campaign r and b when diffusion finished.

4) Parameters: Unless otherwise stated, set $\ell = 1$, $\epsilon = \epsilon_2 = 0.2$, MC simulation round $r = 200$ is used to evaluate the selected seeds. Heuristic is used when the memory limit (8 GB) is met.

B. Main Results

Co-activity results under MCIC/MCLT propagation models with weighted cascade/random parameters and $k_r = 5$, $k_b \in [5, 10]$ are shown in Figs. 1 and 2. Note that Random is omitted from the reports as its corresponding co-activity results are at least two orders of magnitude worse than the other competitors throughout all settings. It means that blindly selecting seeds for competing campaigns is not a good idea for contributing intergroup activities between opposing follower groups. From these two figures, it is easy to observe that while different settings and properties of the networks influence the results, TCoAM algorithm demonstrates its comparably good performance in all of the networks with different experimental settings. Other baselines, e.g., TCEM, although demonstrate good performance in some settings, their performance is observed as not stable across different settings. For example, TCEM shows comparable performance to TCoAM when applied to *rt-assad.wc*, but its performance drops when applied to *rt-assad.rd*.

One interesting finding is that when applied to larger dataset *ca-GrQc* under MCLT model, MaxODeg slightly outperforms TCoAM under seed budgets $k_r = 5$, $k_b \in [5, 10]$. The reason may lie in that under the MCLT model, compared to the MCIC case, the balancing ability of seeds seems to be stronger (Figs. 1 vs. 2), directly choosing "popular" seeds with maximum out-degrees tends to be a good choice when seed budgets are limited, especially for larger network *ca-GrQc* with more highly influential nodes. Based on the above analysis, when more seeds are allowed, the advantages of MaxODeg should be wiped off. We verify this thought by enlarging seed budgets ($k_r = k_b \in [10, 35]$), as shown in Fig. 3.

In addition, for each experimental setting and each algorithm used, we recorded the average of the degrees of seeds selected (SeedsAD) and the average of the activity strength associated with the incident edges of seeds selected (SeedsAA), an example records table for dataset ca-netscience.wc under MCIC model with $k_r = 5$ and $k_b \in [5, 10]$ is shown in Appendix B. Over the total 120 groups of comparisons, only in 2 seeds selected by TCoAM showed a highest SeedsAD, and in 1 seeds selected by TCoAM showed the highest SeedsAA, with ties allowed. It shows the potential of TCoAM to select seeds that may contribute more to co-activity, not just consider the popularity of propagating or receiving cascade.

C. Further Analysis

Based on the above analysis, natural thinking is that when the viralities of campaigns decreased, the balancing results induced by the baselines relying on selecting "popular" nodes as seeds will also decrease as the propagation abilities of those "popular g " nodes are cut. This phenomenon is verified by the following experiments. We select four denser networks *ca-netscience.wc*, *ca-netscience.rd*, *soc-wiki-Vote.wc* and *soc-wiki-Vote.rd* in which baselines demonstrate relatively better performances than in sparser ones. Then all of the propagation parameters in these networks are divided by a factor to produce new networks with more restricted propagation abilities. Denote the original network with suffix ".ORI" and the corresponding new one with suffix ".DIV5"/".DIV10" (we choose the factor $= 5$ or 10). We compared the performance gain in terms of final co-activity by using TCoAM rather than the best of baselines (BoB) in each experimental setting, Fig. 4 shows examples of the performance gain comparison on dataset *ca-netscience* under MCIC model. Clearly, it can be observed that when the viralities of campaigns decreased, TCoAM shows a trend of becoming more advantageous. Furthermore,

[1]Note that TCEM algorithm can be extended to MCLT model with seeds-intersection allowed, as the hardness results of its corresponding CoEM problem can be simply established by reduction from IM problem under LT model, and the following analysis can be adapted from this article.

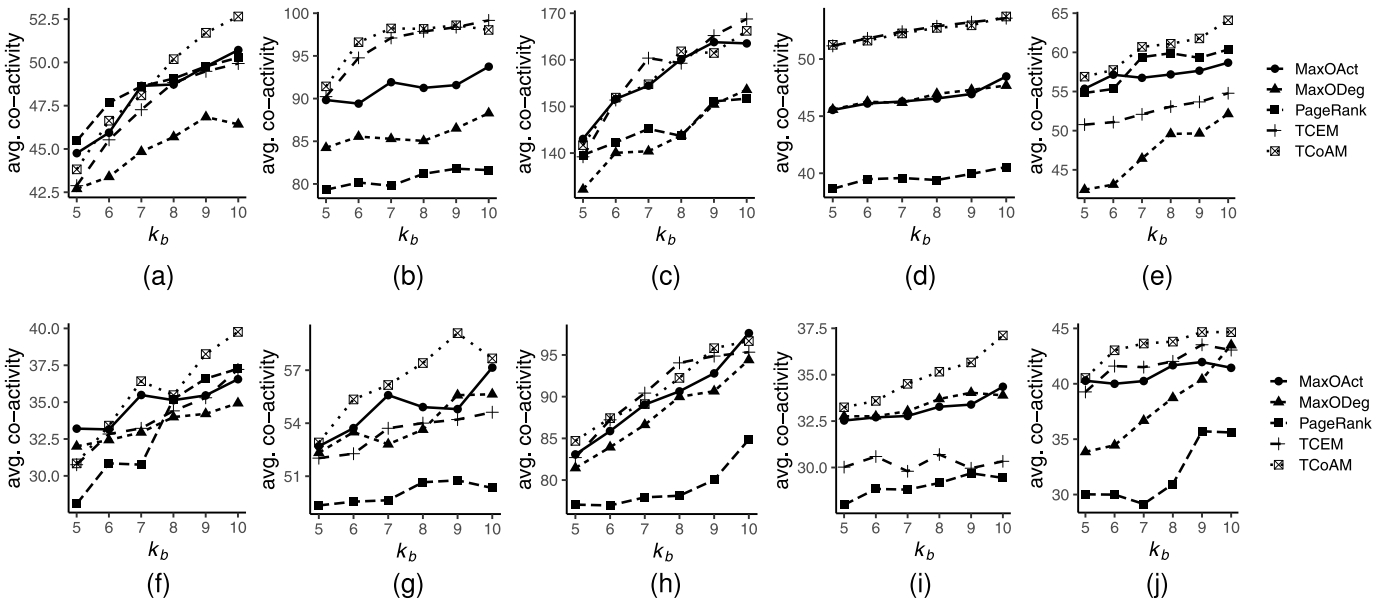

Fig. 1. Under MCIC model, weight cascade/random parameters and $r = 200$, co-activity results on networks for $k_r = 5$ and varying $k_b \in [5, 10]$. (a) ca-netscience.wc. (b) rt-copen.wc. (c) soc-wiki-Vote.wc. (d) rt-assad.wc. (e) ca-GrQc.wc. (f) ca-netscience.rd. (g) rt-copen.rd. (h) soc-wiki-Vote.rd. (i) rt-assad.rd. (j) ca-GrQc.rd.

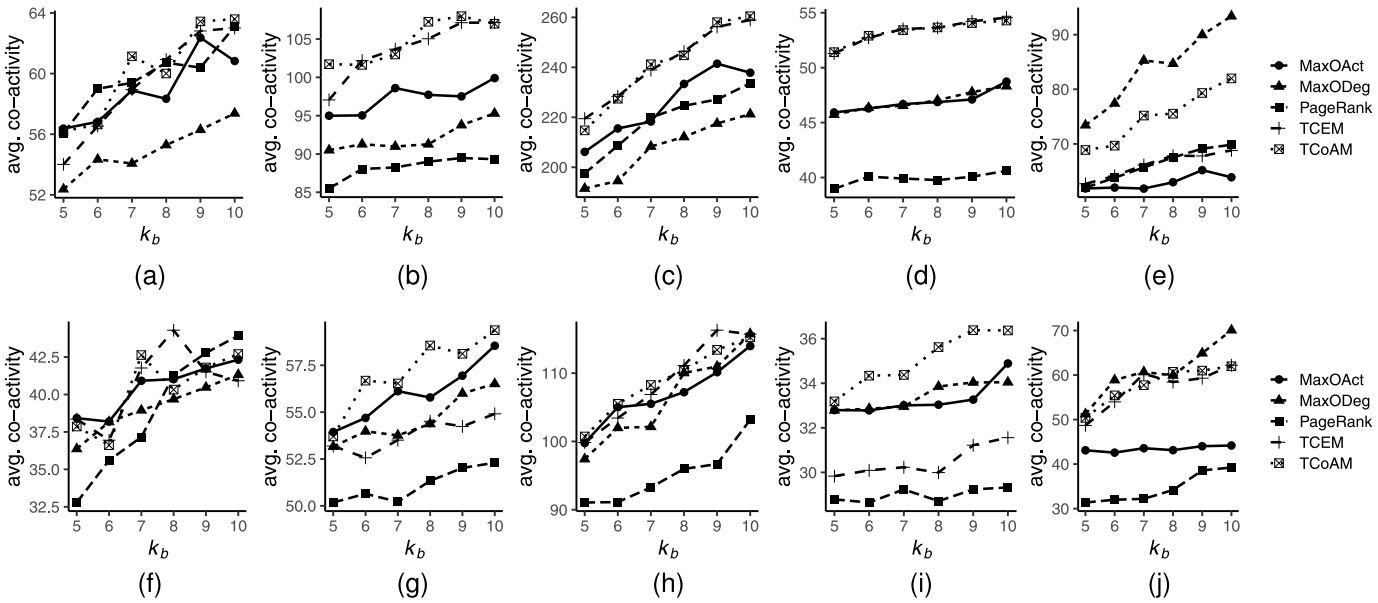

Fig. 2. Under MCLT model, weight cascade/random parameters and $r = 200$, co-activity results on networks for $k_r = 5$ and varying $k_b \in [5, 10]$. (a) ca-netscience.wc. (b) rt-copen.wc. (c) soc-wiki-Vote.wc. (d) rt-assad.wc. (e) ca-GrQc.wc. (f) ca-netscience.rd. (g) rt-copen.rd. (h) soc-wiki-Vote.rd. (i) rt-assad.rd. (j) ca-GrQc.rd.

for each comparing tuple of *.ORI, *.DIV5 and *.DIV10, the balanced edges directly caused by seeds of campaign r and b (i.e., $\mathrm{CO}_{\mathrm{seeds}} = \{(u, v) | (u, v) \in E \wedge u \in S_r \wedge v \in S_b \vee u \in S_b \wedge v \in S_r\}$) are extracted and the corresponding total co-activity (i.e., $\sum_{(u,v)\in \mathrm{CO}_{\mathrm{seeds}}} a_{(u,v)}$) of those edges are recorded. Table II shows an example of comparison of total co-activity directly caused by seeds S_r and S_b selected by TCoAM algorithm run in *ca-netscience.wc.ORI*, *ca-netscience.wc.DIV5*, and *ca-netscience.wc.DIV10* under MCIC model with $k_r = 5$ and $k_b \in [5, 10]$. From Table II, we find that when the propagation effect weakened, TCoAM tends to choose seeds that can directly balance more edges without the help of propagation, in fact, for *ca-netscience.wc.DIV10*, in the case of $k_b = 5$ and $k_b = 10$, these directly balanced edges count about 94.3% in the final co-activity. Similar results are also observed in the other comparisons. These results show that TCoAM is more adaptable to various settings of CoAM, i.e., more suitable for this task.

VII. Conclusion

In this article, a novel problem, called CoAM, which aims at maximizing co-activity in social networks, is investigated.

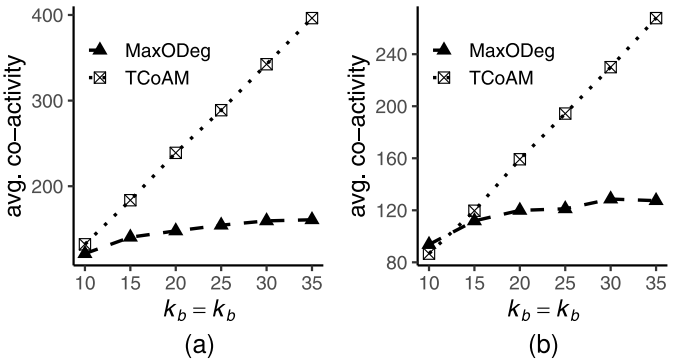

Fig. 3. Under MCLT model, weight cascade/random parameters and $r = 200$, co-activity results comparisons by using TCoAM versus MaxODeg on dataset *ca-GrQc* for $k_r = k_b \in [10, 35]$. (a) ca-GrQc.wc. (b) ca-GrQc.rd.

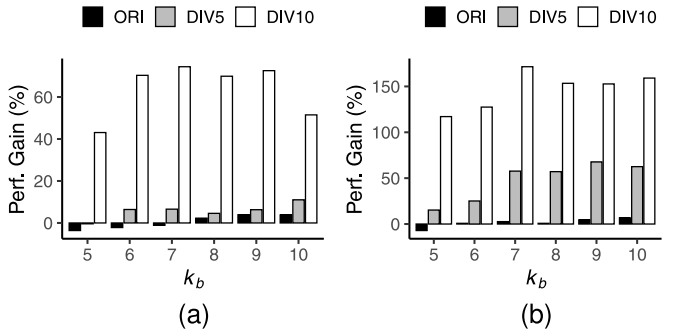

Fig. 4. Under MCIC model, weight cascade/random parameters and $r = 200$, comparison of performance gain in terms of final co-activity by using TCoAM rather than the BoB on dataset *ca-netscience* for $k_r = 5$ and $k_b \in [5, 10]$. "ORI" denotes the original network and "DIV5"/"DIV10" denotes the corresponding virality-reduced one. (a) ca-netscience.wc. (b) ca-netscience.rd.

TABLE II

UNDER MCIC MODEL, COMPARISON OF TOTAL CO-ACTIVITY DIRECTLY CAUSED BY SEEDS S_r AND S_b SELECTED BY TCOAM ALGORITHM RUN IN CA-NETSCIENCE.WC.ORI, CA-NETSCIENCE.WC.DIV5 AND CA-NETSCIENCE.WC.DIV10 WITH $k_r = 5$ AND $k_b \in [5, 10]$

Network \ k_b	5	6	7	8	9	10
*.ORI	0.10	0.16	0.32	0.53	0.63	0.79
*.DIV5	1.26	2.43	3.63	3.36	5.59	6.72
*.DIV10	4.61	6.16	6.51	7.10	7.24	8.50

We show the NP-hardness of CoAM and also its hardness of approximation, furthermore, the objective function of CoAM is neither submodular nor supermodular. In view of this, we design a submodular lower bound of the objective function and with which we devise an approximation algorithm with a provable accuracy guarantee. By extending the idea of RIS and IMM, we propose the TCoAM algorithm, which gives a scalable instantiation of the devised approximation algorithm. Despite the good quality of TCoAM demonstrated in the experiments, several future directions are worth investigating. First, it is eager to improve the current approximation guarantee as when the seeds size increase it drops quickly now. Second, it would be interesting to consider other constraints other than the cardinality one as in reality the cost of seeding varies among nodes. Third, relaxing the mutual independence assumption and introducimg competing to opposing campaigns, or considering multiple campaigns with different leanings and levels of competition, is also a meaningful direction.

REFERENCES

[1] E. Shearer and A. Mitchell, "News use across social media platforms in 2020," Pew Res. Center, Washington, DC, USA, 2021. Accessed: Jun. 30, 2022. [Online]. Available: https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

[2] I. Dylko, I. Dolgov, W. Hoffman, N. Eckhart, M. Molina, and O. Aaziz, "The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure," *Comput. Hum. Behav.*, vol. 73, pp. 181–190, Aug. 2017.

[3] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public Opinion Quart.*, vol. 80, no. S1, pp. 298–320, 2016.

[4] M. D. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on Twitter," in *Proc. ICWSM*, 2011, pp. 89–96.

[5] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 9, Mar. 2021, Art. no. e2023301118.

[6] R. K. Garrett, "Echo chambers online?: Politically motivated selective exposure among internet news users," *J. Comput.-Mediated Commun.*, vol. 14, no. 2, pp. 265–285, Jan. 2009.

[7] D. Nikolov, D. F. M. Oliveira, A. Flammini, and F. Menczer, "Measuring online social bubbles," *PeerJ Comput. Sci.*, vol. 1, p. e38, Dec. 2015.

[8] E. Pariser, *The Filter Bubble: What the Internet is Hiding From You*. London, U.K.: Penguin Group, 2011.

[9] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, vol. 13, no. 2, pp. 1–22, May 2019.

[10] D. Spohr, "Fake news and ideological polarization: Filter bubbles and selective exposure on social media," *Bus. Inf. Rev.*, vol. 34, no. 3, pp. 150–160, Sep. 2017.

[11] M. D. Vicario et al., "The spreading of misinformation online," *Proc. Nat. Acad. Sci. USA*, vol. 113, pp. 554–559, Jan. 2016.

[12] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017.

[13] M. D. Vicario, F. Zollo, G. Caldarelli, A. Scala, and W. Quattrociocchi, "Mapping social dynamics on Facebook: The brexit debate," *Social Netw.*, vol. 50, pp. 6–16, Jul. 2017.

[14] M. Cinelli et al., "The COVID-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, Oct. 2020.

[15] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 81–90.

[16] U. Chitra and C. Musco, "Analyzing the impact of filter bubbles on social network polarization," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 115–123.

[17] C. Musco, C. Musco, and C. E. Tsourakakis, "Minimizing polarization and disagreement in social networks," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 369–378.

[18] K. Garimella, A. Gionis, N. Parotsidis, and N. Tatti, "Balancing information exposure in social networks," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 4666–4674.

[19] S. Tu, C. Aslay, and A. Gionis, "Co-exposure maximization in online social networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 3232–3243.

[20] C. Aslay, A. Matakos, E. Galbrun, and A. Gionis, "Maximizing the diversity of exposure in a social network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 863–868.

[21] T. F. Pettigrew and L. R. Tropp, "A meta-analytic test of intergroup contact theory," *J. Personality Social Psychol.*, vol. 90, no. 5, pp. 751–783, 2006.

[22] D. C. Mutz, "Cross-cutting social networks: Testing democratic theory in practice," *Amer. Political Sci. Rev.*, vol. 96, no. 1, pp. 111–126, Mar. 2002.

[23] K. Grönlund, K. Herne, and M. Setälä, "Does enclave deliberation polarize opinions?" *Political Behav.*, vol. 37, no. 4, pp. 995–1020, Dec. 2015.

[24] A. M. McCright and R. E. Dunlap, "The politicization of climate change and polarization in the American Public's views of global warming, 2001–2010," *Sociol. Quart.*, vol. 52, no. 2, pp. 155–194, May 2011.
[25] H. Holone, "The filter bubble and its effect on online personal health information," *Croatian Med. J.*, vol. 57, no. 3, pp. 298–301, 2016.
[26] A. Matakos and A. Gionis, "Tell me something my friends do not know: Diversity maximization in social networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 327–336.
[27] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Factors in recommending contrarian content on social media," in *Proc. ACM Web Sci. Conf.*, Jun. 2017, pp. 263–266.
[28] Q. V. Liao and W.-T. Fu, "Can you hear me now?: Mitigating the echo chamber effect by source position indicators," in *Proc. 17th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2014, pp. 184–196.
[29] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *J. Math. Sociol.*, vol. 15, nos. 3–4, pp. 193–206, 1990.
[30] C. Aslay, W. Lu, F. Bonchi, A. Goyal, and L. V. S. Lakshmanan, "Viral marketing meets social advertising: Ad allocation with minimum regret," *Proc. VLDB Endowment*, vol. 8, no. 7, pp. 814–825, Feb. 2015.
[31] C. Aslay, F. Bonchi, L. V. Lakshmanan, and W. Lu, "Revenue maximization in incentivized social advertising," *Proc. VLDB Endowment*, vol. 10, no. 11, pp. 1238–1249, Aug. 2017.
[32] P. Chalermsook, A. D. Sarma, A. Lall, and D. Nanongkai, "Social network monetization via sponsored viral marketing," *SIGMETRICS Perform. Eval. Rev.*, vol. 43, no. 1, pp. 259–270, Jun. 2015.
[33] S. Datta, A. Majumder, and N. Shrivastava, "Viral marketing for multiple products," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 118–127.
[34] A. Khan, B. Zehnder, and D. Kossmann, "Revenue maximization by viral marketing: A social network host's perspective," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 37–48.
[35] K. Han, B. Wu, J. Tang, S. Cui, C. Aslay, and L. V. S. Lakshmanan, "Efficient and effective algorithms for revenue maximization in social advertising," in *Proc. Int. Conf. Manage. Data*, Jun. 2021, pp. 671–684.
[36] N. Du, Y. Liang, M.-F. Balcan, M. Gomez-Rodriguez, H. Zha, and L. Song, "Scalable influence maximization for multiple products in continuous-time diffusion networks," *J. Mach. Learn. Res.*, vol. 18, no. 2, pp. 1–45, 2017.
[37] Z. Wang, Y. Yang, J. Pei, L. Chu, and E. Chen, "Activity maximization by effective information diffusion in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2374–2387, Nov. 2017.
[38] T. Chen, W. Liu, Q. Fang, J. Guo, and D.-Z. Du, "Minimizing misinformation profit in social networks," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 6, pp. 1206–1218, Dec. 2019.
[39] J. Guo, T. Chen, and W. Wu, "Continuous activity maximization in online social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2775–2786, Dec. 2020.
[40] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. SODA*, 2014, pp. 946–957.
[41] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining* New York, NY, USA: Association for Computing Machinery, 2010, pp. 1029–1038.
[42] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 88–97.
[43] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 75–86.
[44] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A Martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data* New York, NY, USA: Association for Computing Machinery, 2015, pp. 1539–1554.
[45] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2003, pp. 137–146.
[46] A. Singh, A. Guillory, and J. Bilmes, "On bisubmodular maximization," in *Proc. 15th Int. Conf. Artif. Intell. Statist.*, N. D. Lawrence and M. Girolami, Eds., vol. 22. La Palma, Canary Islands: PMLR, Apr. 2012, pp. 1055–1063.
[47] A. A. Bian, J. M. Buhmann, A. Krause, and S. Tschiatschek, "Guarantees for greedy maximization of non-submodular functions with applications," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 498–507.
[48] W. Lu, W. Chen, and L. V. Lakshmanan, "From competition to complementarity: Comparative influence diffusion and maximization," *VLDB Endowment*, vol. 9, no. 2, pp. 60–71, 2015.
[49] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM J. Comput.*, vol. 40, no. 6, pp. 1740–1766, Dec. 2011.
[50] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI*, 2015, pp. 4292–4293. [Online]. Available: https://networkrepository.com
[51] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999.

Dongyu Mao received the M.S. degree in computer science from The University of Texas at Dallas, Richardson, TX, USA, in 2018.

He is a Student Researcher at The University of Texas at Dallas. He is broadly interested in data management, optimization, software engineering, machine learning, and deep learning. He is currently exploring research topics in computational social systems and approximation algorithm design working with Dr. D.-Z. Du and Dr. W. Wu.

Weili Wu (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from the University of Minnesota, Minneapolis, MN, USA, in 1998 and 2002, respectively.

She is currently a Full Professor with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. Her current research interests include data communication and data management, and design and analysis of algorithms for optimization problems that occur in wireless networking environments and various database systems.

Ding-Zhu Du received the M.S. degree in operations research from the Chinese Academy of Sciences, Beijing, China, in 1981, and the Ph.D. degree in mathematics with a concentration in theoretical computer science from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 1985.

He has been involved in research on the design and analysis of approximation algorithms for 30 years. He has authored/coauthored 177 journal articles, 60 conference and workshop papers, 22 editorship, nine reference works, and 11 informal publications.