A Scalable Algorithm for Fair Influence Maximization with Unbiased Estimator

Xiaobin Rui, Zhixiao Wang, Hao Peng, Wei Chen, *Fellow, IEEE*, Philip S. Yu, *Life Fellow, IEEE*

Abstract—This paper studies the fair influence maximization problem with efficient algorithms. In particular, given a graph G, a community structure $\mathcal{C}$ consisting of disjoint communities, and a budget k, the problem asks to select a seed set S ($|S| = k$) that maximizes the influence spread while narrowing the influence gap between different communities. This problem derives from some significant social scenarios, such as health interventions (*e.g.* suicide/HIV prevention) where individuals from racial minorities or LGBTQ communities may be disproportionately excluded from the benefits of the intervention. To depict the concept of fairness in the context of influence maximization, researchers have proposed various notions of fairness, where the welfare fairness notion that better balances fairness level and influence spread has shown promising effectiveness. However, the lack of efficient algorithms for optimizing the objective function under welfare fairness restricts its application to networks of only a few hundred nodes. In this paper, we modify the objective function of welfare fairness to maximize the exponentially weighted sum and the logarithmically weighted sum over all communities' influenced fractions (utility). To achieve efficient algorithms with theoretical guarantees, we first introduce two unbiased estimators: one for the fractional power of the arithmetic mean and the other for the logarithm of the arithmetic mean. Then, by adapting the Reverse Influence Sampling (RIS) approach, we convert the optimization problem to a weighted maximum coverage problem. We also analyze the number of reverse reachable sets needed to approximate the fair influence at a high probability. Finally, we present an efficient algorithm that guarantees $1 - 1/e - \varepsilon$ (positive objective function) or $1 + 1/e + \varepsilon$ (negative objective function) approximation for any small $\varepsilon > 0$. Experiments demonstrate that our proposed algorithm could efficiently handle large-scale networks with good performance.

Index Terms—Influence maximization, scalable algorithm, fairness, reverse influence sampling, unbiased estimator.

1 Introduction

Derived from social advertising, influence maximization (IM) is a widely studied problem in social network analysis. The formal definition of the problem can be described as follows: given a graph G and a budget k (positive integer), the objective is to find a node set S ($|S| \leq k$) that can disseminate information to trigger the largest expected number of remaining nodes. In real-world applications, diverse scenarios have generated varying demands for IM algorithms, resulting in the development of several variants of the classic influence maximization problem. These variants address specific challenges, such as adaptive influence maximization [1], multi-round influence maximization [2], competitive influence maximization [3], budgeted influence maximization [4], limited access influence maximization [5] and time-critical influence maximization [6]. Influence maximization and its variants have a broad range of applications in social contexts, including viral marketing, health interventions, rumor control, etc [7].

Given the context of when disseminating public health interventions, such as suicide or HIV prevention [8] programs or promoting community preparedness against natural disasters, selecting individuals to act as peer leaders can help maximize outreach by leveraging the principles of influence maximization. However, solutions to the classic IM problem may often lead to discriminatory outcomes, as individuals from racial minorities or LGBTQ communities might be disproportionately excluded from the benefits of the intervention [9] when community structure is disregarded. Consequently, fairness in influence maximization, derived from such significant social scenarios, has become a focus of attention for recent researchers [6], [10]–[12].

Generally, fair influence maximization seeks to improve the fraction of individuals influenced within communities where coverage may otherwise be disproportionately low. Currently, a universally accepted definition of fair influence maximization remains elusive. Recent studies have attempted to incorporate fairness into influence maximization by proposing various notions of fairness, such as maximin fairness [10], diversity constraints [10], equity fairness [13], equality fairness [13], and welfare fairness [9]. Among these notions, welfare fairness has demonstrated several attractive features. Its objective functions are in the form of $\sum_{c\in\mathcal{C}} n_c \boldsymbol{u}_c^{\alpha}/\alpha$ for $\alpha < 1$, $\alpha \neq 0$ and $\sum_{c\in\mathcal{C}} n_c \log(\boldsymbol{u}_c)$ for $\alpha = 0$, where $\mathcal{C}$ denotes the community structure, n_c denotes the number of nodes in c, $\boldsymbol{u}_c$ represents the utility (influenced fraction) of c, and α is the inequality aversion parameter. Based on the Cardinal Welfare theory [14], both objective functions satisfy monotonicity and submodularity, which naturally enables (if positive) a greedy algorithm

- *Xiaobin Rui and Zhixiao Wang are with the School of Computer Science, China University of Mining and Technology, Xuzhou Jiangsu, 221116, China, and also with the Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou Jiangsu 221116, China. E-mail: {ruixiaobin, zhxwang}@cumt.edu.cn. E-mail: lis221@lehigh.edu.*
- *Hao Peng is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China. E-mail: penghao@buaa.edu.cn*
- *Wei Chen is with the Theory Center of Microsoft Research Asia, Beijing 100080, China. E-mail: weic@microsoft.com.*
- *Philip S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA. E-mail: psyu@uic.edu.*

(Corresponding author: Zhixiao Wang.)

0000–0000/00$00.00 © 2021 IEEE

with a $1-1/e$ approximation factor [15]. However, despite Rahmattalabi *et al.*'s [9] comprehensive analysis of welfare fairness, no efficient algorithm currently exists to optimize its objective function with a provable guarantee. This limitation restricts the practical application of their approach to small-scale networks, typically containing only a few hundred nodes.

In this study, we focus on $0 \leq \alpha < 1$ and adapt the objective of welfare fairness to $F_\alpha(S) = \sum_{c\in\mathcal{C}} n_c \boldsymbol{u}_c(S)^\alpha$ for $0 < \alpha < 1$ and $F_0(S) = \sum_{c\in\mathcal{C}} n_c \ln(\boldsymbol{u}_c(S))$ for $\alpha = 0$, which we both refer to as the fair influence of the seed set S. Fair influence can be viewed as the weighted sum (with community size as the weight) of the fractional power or the logarithm of the expected proportion of activated nodes within each community. The fractional exponent α is the inequality aversion parameter, allowing one to balance between fairness and influence spread, with α tending to 1 for influence spread and α tending to 0 for fairness. When $\alpha = 0$, it becomes a peculiar case that asks all communities to be reached from S. If any community remains unreached, the objective function becomes negative infinity.

To efficiently maximize fair influence with theoretical guarantees, one may leverage Reverse Influence Sampling (RIS), a common approach adopted in many efficient IM algorithms [16]–[19]. Nevertheless, adapting the RIS approach to the welfare fairness objective presents two key challenges: (1) how to achieve an unbiased estimation of the fractional power and the logarithm of the expected proportion of activated nodes in each community, as directly obtaining an unbiased estimate of the expected proportion and then taking its fractional power or logarithm results is biased (revealed by Jensen's inequality); and (2) how the designed unbiased estimation can be integrated into the RIS framework to estimate fair influence using reverse reachable sets, enabling straightforward seed selection based on maximum coverage? In this paper, we address both challenges and propose a new scalable fair influence maximization algorithm with theoretical guarantees.

Our contributions can be summarized as follows:

- We propose two unbiased estimators: one for the fractional power of the arithmetic mean and another for the logarithm of the arithmetic mean, by leveraging Taylor expansion techniques. These estimators enable us to accurately estimate the fair influence under welfare fairness.
- Based on the above unbiased estimators, we adapt the RIS approach to approximate the fair influence with Reverse Reachable (RR) sets and propose the FIMM algorithm that works efficiently while guaranteeing the $(1-1/e-\varepsilon)$-approximation for the positive objective or $(1+1/e+\varepsilon)$-approximation for the negative objective. Our theoretical analysis needs to address the concentration of the unbiased estimator of fractional power or logarithm, which is much more involved than the standard RIS analysis.
- We carry out a detailed experimental analysis on eight real social networks to investigate the trade-off between fairness and total influence spread. Our experiments evaluate the performance of the proposed algorithms under varying fairness parameters, influence probabilities, seed budgets, and community structures. Moreover, the inclusion of a large-scale network demonstrates the scalability and robustness of our proposed algorithms.

This paper is an in-depth extension of our conference paper [20]. We have extended both of our algorithms' technical contribution and empirical evaluation. The main extensions in this paper are summarized as follows: (1) First, in addition to the objective function $\sum_{c\in\mathcal{C}} n_c \boldsymbol{u}_c^\alpha, 0 < \alpha < 1$ studied in [20], this paper introduces the other objective function $\sum_{c\in\mathcal{C}} n_c \ln(\boldsymbol{u}_c)$ specifically for $\alpha = 0$. Note that this represents a fundamentally different scenario where the objective requires all communities to be reached. (2) Second, to address this new case, we propose a new unbiased estimator for $\sum_{c\in\mathcal{C}} n_c \ln(\boldsymbol{u}_c)$ and incorporate it into our algorithms. We also analyze the lower bound under the new objective, which complements the theoretical study of fair influence maximization under social welfare. (3) Third, we evaluate our algorithm across a broader range of datasets, varying in scale and explicit group structures, with the entire experimental framework redesigned. This allows us to assess the performance of our proposed algorithms under more complex parameter combinations, leading to a comprehensive discussion and a deeper understanding of fair influence maximization within the context of welfare fairness. (4) Additionally, we provide a more extensive review of related works, particularly focusing on algorithms for the fair influence maximization problem.

2 Related Work

2.1 Influence Maximization

Influence maximization (IM) is first studied as an algorithmic problem by Domingos and Richardson [21], [22]. Then, Kempe *et al.* [23] mathematically formulate IM as a discrete optimization problem and prove it is NP-hard. They also provide a naive Greedy algorithm with $1-1/e$ approximation based on the submodularity and monotonicity of the problem. Hence, many works have been proposed to improve the efficiency and scalability of influence maximization algorithms [16]–[18], [24]–[33]. For example, Leskovec *et al.* [24] proposed the CELF algorithm, which exploits the submodularity of the IM problem and accelerates the naive Greedy about 700 times faster. Goyal *et al.* [25] extended CELF to CELF++, which additionally maintains the marginal gain for the node that had the previous best marginal gain. Wang *et al.* [26] derived an upper bound for the spread function and proposed the UBLF algorithm that enhances CELF by 2-10 times. To improve the efficiency of estimating influence spread, Chen *et al.* [27] proposed NewGreedy, and Cheng *et al.* [30] introduced StaticGreedy. Both algorithms utilize the idea of estimating influence through a series of pre-generated snapshots. Subsequently, Ohsaka *et al.* [31] proposed the PMC algorithm which applies intelligent pruning strategies to further improve scalability. Beyond these methods, the most recent and the state-of-the-art is the reverse influence sampling (RIS) approach [16]–[18], [32], [33], where the IMM algorithm [18] is one of the representative implementations. The idea of RIS approaches is to generate a suitable number of reverse reachable sets (*a.k.a.* RR sets), and then the influence spread can be approximated at a high probability based on these RR sets. Therefore, the greedy approach can be easily applied by iteratively selecting the node that could bring the maximum marginal gain in terms of influence spread as a seed node.

2.2 Fairness in Influence Maximization

In recent years, the study of fair influence maximization has attracted considerable attention. Researchers have attempted to incorporate their designed notions of fairness into the influence maximization framework. Stocia *et al.* [34] propose the prototype of equality constraints, which require that the number of seed nodes in each community is proportional to the population ratio of that community. It focuses on the fair distribution among the seed nodes without considering whether the influence is fairly allocated among communities. Based on the Rawlsian theory, the maximin fairness [10], [35] aims to maximize the influence fraction of the worst-off community. Inspired by the game theoretic notion of core, diversity constraints [10] require that every community obtains an influenced fraction higher than when it receives resources proportional to its size and allocates them internally. Equity-based notion [36] strives for equal influenced fraction across all communities. However, these notions can hardly balance fairness and total influence and usually lead to a high influence reduction. Especially, strict equity [36] is rather hard to achieve in influence maximization. To address these shortcomings, Rahmattalabi *et al.* [9] propose the welfare fairness that can control the trade-off between fairness and total influence by an inequality aversion parameter.

Based on the cardinal welfare theory [14], the objective function of welfare fairness is to maximize the weighted sum over the exponential influenced fraction of all communities. Similarly, Fish *et al.* [37] also follow welfare functions and propose ϕ-mean fairness, where the objective function becomes MMF when $\phi = -\infty$. However, they only consider fairness at the individual level and do not address the challenge of unbiased estimation of the fractional power.

2.3 Algorithms for Fair Influence Maximization

Currently, most algorithms addressing the fair influence maximization problem are based on either mixed integer programming or multi-objective optimization.

Rahmattalabi *et al.* [35] express the objective of maximin fairness as the optimal objective value of a covering problem, which is equivalent to the two-stage linear robust problem. They try to address the problem with a mixed integer bilinear program. Farnadi *et al.* [13] develop a general formalism for different notions of fairness (including maximin, equality, equity, and diversity), where the problem defined in this formalism can be then solved using efficient mixed integer programming solvers. Moreover, Becker *et al.* [11] study two different variants of maximin fairness, allowing for randomized strategies in choosing seeds rather than being restricted to deterministic strategies (*i.e.*, sets of size k). They further show that the problem can be approximated to within a constant factor using a specific kind of linear programming algorithm.

Tsang *et al.* [10] show that optimizing either the utility function of maximin fairness or diversity constraints reduces to multi-objective submodular maximization. Their proposed algorithm employs a Frank-Wolfe style approach to simultaneously optimize the multilinear extensions of the discrete objectives. Similarly, Rahmattalabi *et al.* [9] view welfare fairness as a multi-objective submodular optimization with the utility of each community being a separate objective and follow the algorithm proposed by Tsang *et al.* [10]. However, the multi-objective solver is time-consuming, restricting their algorithms to only hundred-scale networks.

Recent studies have attempted to propose efficient algorithms for fair influence maximization. In 2023, Lin *et al.* [19] address the scalability problem by applying attribute-aware reverse influence sampling. Although their proposed algorithms demonstrated efficiency, they did not provide a guarantee on the approximation ratio between their solutions and the optimal one. In the same year, Feng *et al.* [12] propose an approach based on learning node representations (embeddings) from diffusion cascades for fair spread, rather than social connectivity. In this way, the method can handle very large graphs but requires a substantial number of diffusion cascades as prior knowledge. Thus, to the best of our knowledge, we are the first to study scalable algorithms with theoretically guaranteed solutions in the context of fair influence maximization.

3 Preliminaries and Problem Definition

3.1 Information Diffusion Model

In this paper, we adopt the well-studied *Independent Cascade (IC) model* [23] as the basic information diffusion model. Under the IC model, a social network is viewed as a directed influence graph $G = (V, E, p)$, where V is the set of vertices (nodes) and $E \subseteq V \times V$ is the set of directed edges that connect pairs of nodes. For an edge $(v_i, v_j) \in E$, $p(v_i, v_j)$ indicates the probability that v_i influences v_j. In the IC model, the diffusion of information or influence proceeds in discrete time steps. At time $t = 0$, the *seed set* S is selected to be active, denoted as A_0. At each time $t \geq 1$, all nodes in A_{t-1} try to influence their inactive neighbors with a one-shot attempt for each neighbor, following influence probability $p(v_i, v_j) \in V \times V$. The set of activated nodes at step t is denoted as A_t. The diffusion process ends when there is no more node activated in a time step, *i.e.*, $A_t = \varnothing$ after diffusion.

An important metric in influence maximization is the *influence spread*, denoted as $\sigma(S)$, which is defined as the expected number of active nodes when the propagation from the given seed set S ends. For the IC model, $\sigma(S) = \mathbb{E}[|A_0 \cup A_1 \cup A_2 \cup \dots|] = \mathbb{E}[|A_0| + |A_1| + |A_2| + \dots]$. In this paper, we use $ap(v, S)$ to represent the probability that node v is activated given the seed set S, and then it establishes that $\sigma(S) = \sum_{v \in V} ap(v, S)$.

3.2 Live-edge Graph

Given the influence probability $p(v_i, v_j) \in V \times V$, we can construct the live-edge graph $\mathcal{L} = (V, E(\mathcal{L}))$, where each edge (v_i, v_j) is selected independently to be a *live-edge* with the probability $p(v_i, v_j)$. The influence diffusion in the IC model is equivalent to the deterministic propagation via bread-first traversal in a random live-edge graph $\mathcal{L}$ [23]. Let $\Gamma(G, S)$ denote the set of nodes in graph G that can be reached from the node set S. By the above live-edge graph model, we have $\sigma(S) = \mathbb{E}_{\mathcal{L}}[|\Gamma(\mathcal{L}, S)|] =$

$\sum_{\mathcal{L}} Pr[\mathcal{L}|G] \cdot |\Gamma(\mathcal{L}, S)|$, where the expectation is taken over the distribution of live-edge graphs, and $Pr[\mathcal{L}|G]$ is the probability of sampling a live-edge graph $\mathcal{L}$ in graph G.

3.3 Reverse Influence Sampling

Many efficient influence maximization algorithms such as IMM [18] are based on the Reverse Influence Sampling (RIS) approach, which generates a suitable number of reverse-reachable (RR) sets for influence estimation. Let $\mathcal{L}$ be a random live-edge graph generated from $G = (V, E, p)$, which maps to an arbitrary diffusion instance. An RR set $RR(v)$ (rooted at node $v \in V$) can be generated by reversely simulating the influence diffusion process starting from the root v, and then adding all nodes reached by reversed simulation into this RR set. Assume both the above-mentioned $\mathcal{L}$ and the simulated $RR(v)$ come from the same diffusion instance, then $RR(v)$ is equivalent to collecting all nodes that can reach v in that random live-edge graph $\mathcal{L}$, denoted by $\Gamma'(\mathcal{L}, v)$.

Intuitively, each node $u \in RR(v)$ if selected as a seed would activate v in this random diffusion instance. We say that S covers an RR set $RR(v)$ if $S \cap RR(v) \neq \varnothing$. If we use $RR(v)$ to represent a randomly generated RR set when v is not specified, *i.e.*, $RR(v) = \Gamma'_{\mathcal{L}\sim\mathcal{U}(P_{\mathcal{L}})}(\mathcal{L}, v)$ where $P_{\mathcal{L}}$ is the space of all live-edge graphs and $\mathcal{U}(\cdot)$ denotes the uniform distribution. Then, the expected activated probability $ap(v, S)$ is equivalent to the probability that S covers a randomly generated v-rooted RR set [16], *i.e.*, $ap(v, S) = \mathbb{E}[Pr\{RR(v) \cap S \neq \varnothing\}]$.

3.4 Approximation Solution

A set function $f : V \to \mathbb{R}$ is called *submodular* if for all $S \subseteq T \subseteq V$ and $u \in V \setminus T$, $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$. Intuitively, submodularity characterizes the diminishing return property often occurring in economics and operation research. Moreover, a set function f is called *monotone* if for all $S \subseteq T \subseteq V$, $f(S) \leq f(T)$. It is shown in [23] that the influence spread $\sigma(\cdot)$ for the IC model is a monotone submodular function. A non-negative monotone submodular function allows a greedy solution to its maximization problem with $1 - 1/e$ approximation [15], which provides the technical foundation for most influence maximization tasks.

3.5 Fair Influence Maximization

For a given graph G with n_G nodes, the classic influence maximization problem is to choose a seed set S consisting of at most k seeds to maximize the influence spread $\sigma(S, G)$. Assuming each node belongs to one of the disjoint communities $c \in \mathcal{C} := \{1, 2, \ldots, C\}$, such that $V_1 \cup V_2 \cup \cdots \cup V_C = V$ where V_c ($n_c = |V_c|$) denotes the set of nodes that belongs to community c. Generally, *fair influence maximization (FIM)* aims to narrow the influence gap between different communities while maintaining the total influence spread as much as possible. In this paper, we adapt the notion of fairness proposed by Rahmattalabi *et al.* [9], where the welfare function is used to aggregate the cardinal utilities of different communities. The goal is to select at most k seed nodes, such that the objective functions $F_\alpha(S)$ or $F_0(S)$ (also referred to as fair influence in this paper) are maximized, where $F_\alpha(S) = \sum_{c\in\mathcal{C}} n_c \boldsymbol{u}_c(S)^\alpha, 0 < \alpha < 1$, and $F_0(S) = \sum_{c\in\mathcal{C}} n_c \ln(\boldsymbol{u}_c(S)), \alpha = 0$. The utility $\boldsymbol{u}_c(S)$ denotes the expected proportion of influenced nodes in the community c with the seed set S. The exponent α is the inequality aversion parameter that controls the trade-off between fairness and total influence, with α approaching 1 favoring influence spread and α approaching 0 favoring fairness. When $\alpha = 0$, it becomes a unique case in which all communities are asked to be reached from S. We thus define the fair influence maximization problem in this paper as follows:

Definition 1. *The Fair Influence Maximization (FIM) under the independent cascade model is the optimization task where the input includes the directed influence graph $G = (V, E, p)$, the non-overlapping community structure $\mathcal{C}$, and the budget k. The goal is to find a seed set S^* to maximize the fair influence,* i.e., $S^* = \operatorname{argmax}_{S:|S|=k} F_\alpha(S)$ *for* $0 < \alpha < 1$, *and* $S^* = \operatorname{argmax}_{S:|S|=k} F_0(S)$ *for* $\alpha = 0$.

According to [9], the adapted fair influence $F_\alpha(S)$ and $F_0(S)$ in this paper still satisfies both monotone and submodular, which provides the theoretical basis for our efficient algorithm design, to be presented in the next section.

4 Methodology

The fair influence objective function possesses monotonicity and submodularity, which allows the natural way of utilizing a greedy approach for maximization. However, as frequently noted in influence maximization studies, implementing a greedy strategy directly leads to significant computational costs. This is primarily due to the need for a large number of Monte Carlo simulations to accurately estimate the influence spread. In this section, we aim to

TABLE 1
Important symbols appeared in this paper.

Symbol	Explanation
$G = (V, E, p)$	A network;
V	Node set of the network;
E	Edge set of the network;
n_G	The number of nodes in G, i.e. $n_G = \|V\|$;
$p(v_i, v_j)$	The probability that v_i influence v_j;
$\mathcal{C} = \{c_1, c_2, \cdots\}$	Community structure;
C	The number of communities in $\mathcal{C}$;
V_c	The node set in community c;
n_c	The number of nodes in community c, i.e. $n_c = \|V_c\|$;
S	A seed set;
S^*	The optimal seed set for fair influence maximization;
$ap(v, S)$	The expected probability that v is activated by S;
$\sigma(S)$	Influence spread of S, i.e. $\sigma(S) = \sum_{v\in V} ap(v, S)$;
$\boldsymbol{u}_c$	The utility of c (expected influenced fraction in c);
$F_\alpha(S)$	The fair influence of S for $0 < \alpha < 1$;
$F_0(S)$	The fair influence of S for $\alpha = 0$;
$\mathcal{R}$	A set of RR sets;
$\mathcal{R}_c$	The set of RR sets rooted in community c;
$\hat{F}_\alpha(S, \mathcal{R})$	The unbiased estimator for $F_\alpha(S)$ based on $\mathcal{R}$;
θ	The total number of RR sets;
θ_c	The number of RR sets rooted in community c;
α	The aversion parameter regarding fairness;
Q	The approximation parameter for Taylor expansion;
ε	The accuracy parameter;
ℓ	The confidence parameter.

significantly speed up the greedy approach by adapting the reverse influence sampling (RIS) [16]–[18], which provides both theoretical guarantee and high efficiency. We propose the FIMM algorithm that is efficient when the number of communities is relatively small, which is hopefully a common situation such as gender and ethnicity. Basically, our proposed FIMM algorithm follows the idea of the IMM algorithm [18], with a new challenge lying in accurately estimating two new objective functions of fair influence without introducing bias.

For the convenience of reading, we list the most important symbols featured in this paper in Table 1.

4.1 Unbiased Fair Influence

To estimate the influence spread, we may generate a number of live-edge graphs $\mathcal{L} = \{L_1, L_2, \cdots, L_t\}$ as samples. Then, for a given seed set S, $\hat{\sigma}(\mathcal{L}, S) = \frac{1}{t}\sum_{i=1}^{t} |\Gamma(L_i, S)|$ is an unbiased estimator of $\sigma(S)$.

However, situations are completely different for fair influence. For each community c, its corresponding fair influence is $n_c \boldsymbol{u}_c^{\alpha}$. If we still generate a number of live-edge graphs and estimate $\boldsymbol{u}_c$ by $\hat{\boldsymbol{u}}_c(\mathcal{L}, S) = \frac{1}{t}\sum_{i=1}^{t} |\Gamma(L_i, S) \cap V_c|/|V_c|$, then $\hat{\boldsymbol{u}}_c(\mathcal{L}, S)$ is an unbiased estimator for $\boldsymbol{u}_c$, but $\hat{\boldsymbol{u}}_c(\mathcal{L}, S)^{\alpha}$ is a biased estimator of $\boldsymbol{u}_c^{\alpha}$ for $0 < \alpha < 1$, and $\ln(\hat{\boldsymbol{u}}_c(\mathcal{L}, S))$ is a biased estimator of $\ln(\boldsymbol{u}_c)$ for $\alpha = 0$. In fact, the value of $\boldsymbol{u}_c^{\alpha}$ and $\ln(\boldsymbol{u}_c)$ would be generally higher than the true value, which is revealed by the following Jensen's inequality.

Fact 1. *(Jensen's inequality) If X is a random variable and ϕ is a concave function, then*

$$\mathbb{E}[\phi(X)] \leq \phi(\mathbb{E}[X]).$$

Therefore, our first challenge in dealing with the welfare fairness objective is to provide unbiased estimators for the fractional power value of $\boldsymbol{u}_c^{\alpha}$ and the logarithm value of $\ln(\boldsymbol{u}_c)$. We meet this challenge by incorporating Taylor expansion as in Lemma 1. It is worth noting that the bias introduced by Jensen's inequality may not significantly impact the seed selection process. Adopting such a biased estimator can still yield empirically good performance with high probability. However, it does not offer any theoretical guarantee, which is precisely the aspect this paper aims to address.

Lemma 1. *For a given seed set S and an inequality aversion parameter α, the fair influence*

$$F_{\alpha}(S) = \sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\infty} \eta(n, \alpha)\big(1 - \boldsymbol{u}_c(S)\big)^n\right), \quad (1)$$

$$\eta(n, \alpha) = \begin{cases} 1, & n = 1, \\ \frac{(1-\alpha)(2-\alpha)\ldots(n-1-\alpha)}{n!}, & n \geq 2, \end{cases}$$

and

$$F_0(S) = -\sum_{c \in \mathcal{C}} n_c \sum_{n=1}^{\infty} \frac{1}{n}\big(1 - \boldsymbol{u}_c(S)\big)^n. \quad (2)$$

Proof. By Taylor expansion of the binomial series, we have

$$(1+x)^{\alpha} = 1 + \sum_{n=1}^{\infty} \binom{\alpha}{n} x^n, \binom{\alpha}{n} = \frac{\alpha(\alpha-1)\ldots(\alpha-n+1)}{n!}.$$

By definition of the fair influence $F_{\alpha}(S)$, we have

$$\begin{aligned} F_{\alpha}(S) &= \sum_{c \in \mathcal{C}} n_c \big(1 + \big(\boldsymbol{u}_c(S) - 1\big)\big)^{\alpha} \\ &= \sum_{c \in \mathcal{C}} n_c \left(1 + \sum_{n=1}^{\infty} \binom{\alpha}{n} \big(\boldsymbol{u}_c(S) - 1\big)^n\right) \\ &= \sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\infty} \eta(n, \alpha)\big(1 - \boldsymbol{u}_c(S)\big)^n\right), \quad (3) \end{aligned}$$

where

$$\eta(n, \alpha) = \begin{cases} 1, & n = 1, \\ \frac{(1-\alpha)(2-\alpha)\ldots(n-1-\alpha)}{n!}, & n \geq 2. \end{cases}$$

Using the Taylor expansion of the logarithm series, we have

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{1}{n}(-1)^{n-1} x^n.$$

By the definition of the fair influence $F_0(S)$, we have

$$\begin{aligned} F_0(S) &= \sum_{c \in \mathcal{C}} n_c \ln\big(1 + \big(\boldsymbol{u}_c(S) - 1\big)\big) \\ &= \sum_{c \in \mathcal{C}} n_c \left(\sum_{n=1}^{\infty} \frac{1}{n}(-1)^{n-1}\big(\boldsymbol{u}_c(S) - 1\big)^n\right) \\ &= -\sum_{c \in \mathcal{C}} n_c \sum_{n=1}^{\infty} \frac{1}{n}\big(1 - \boldsymbol{u}_c(S)\big)^n. \quad (4) \end{aligned}$$

This concludes the proof. □

Lemma 1 demonstrates that the calculation of fair influence with fractional powers can be transformed into the summation of integral powers. Further, we can get an unbiased estimator for integral powers of arithmetic mean as given in the following Lemma 2.

Lemma 2. *[38] Suppose that a simple random sample of size m is to be drawn, with replacement, in order to estimate the mean μ^n. An unbiased estimator for μ^n $(n \leq m)$ is*

$$\hat{\mu}^n = \frac{(m-n)!}{m!}\{\textstyle\sum x_{i_1} x_{i_2} \cdots x_{i_n}\}(i_1 \neq i_2 \neq \cdots \neq i_n), \quad (5)$$

where the summation extends over all permutations of all sets of n observations in a sample subject only to the restriction noted.

4.2 Unbiased Fair Influence with RR sets

As mentioned above, many efficient influence maximization algorithms such as IMM [18] are based on the Reverse Influence Sampling (RIS) approach, which generates a suitable number of reverse-reachable (RR) sets for influence estimation. Recall that an RR set $RR(v)$ is generated by reversely simulating the influence diffusion (a random diffusion instance) from the root v. The seed set S covering $RR(v)$ implies that S could reach (influence) v in that random diffusion instance.

Let X_c be the random event that indicates whether a randomly selected node in community c would be influenced in a diffusion instance by the given seed set S. As given in Section 3.3, an RR set maps to a random diffusion instance. Assuming we generate $\mathcal{R}$ consisting of θ RR sets in total and each community c gets θ_c RR sets. Let $\mathcal{R}_c$ be the set of RR sets that are rooted in the community c, *i.e.*, $|\mathcal{R}_c| = \theta_c$. Let X_c^i ($i \in [\theta_c]$) be a random variable for each RR set $R_i \in \mathcal{R}_c$, such that $X_c^i = 1$ if $\mathcal{R}_c^i \cap S \neq \varnothing$, and $X_c^i = 0$ otherwise. Then, we have $\mathbb{E}[X_c] = \boldsymbol{u}_c$ and $\mathbb{E}[\overline{X_c}] = 1 - \boldsymbol{u}_c$.

Based on Lemma 1 and Lemma 2, we can get the unbiased estimator of $\mathbb{E}[X_c]^\alpha$ through RR sets as

$$\begin{aligned}\mathbb{E}[X_c]^\alpha &= 1 - \alpha \sum_{n=1}^{\infty} \eta(n,\alpha)(1-\mathbb{E}[X_c])^n \\ &= 1 - \alpha \Big(\sum_{n=1}^{\theta} \eta(n,\alpha)(1-\mathbb{E}[X_c])^n + \xi(\theta_c) \Big), \end{aligned} \tag{6}$$

where

$$\begin{aligned}\xi(\theta_c) &= \sum_{n=\theta_c+1}^{\infty} \eta(n,\alpha)(1-\mathbb{E}[X_c])^n \\ &\leq \sum_{n=\theta_c+1}^{\infty} \frac{1}{n} \cdot (1-\mathbb{E}[X_c])^n \\ &\leq \frac{1}{\theta_c+1} \cdot \frac{(1-\mathbb{E}[X_c])^{\theta_c+1}}{\mathbb{E}[X_c]} \end{aligned} \tag{7}$$

Thus, we get the unbiased estimator[1] of $\mathbb{E}[X_c]^\alpha$ as

$$\hat{E}[X_c]^\alpha = 1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \frac{(\theta_c - n)!}{\theta_c!} \sum \prod_{d=1}^{n} \overline{X_c^{i_d}} \tag{8}$$

Further, we can get the unbiased estimator of $F_\alpha(S)$ as

$$\begin{aligned}\hat{F}_\alpha(S,\mathcal{R}) &= \sum_{c\in\mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \frac{(\theta_c - n)!}{\theta_c!} \sum \prod_{d=1}^{n} \overline{X_c^{i_d}} \right) \\ &= \sum_{c\in\mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} \right), \end{aligned} \tag{9}$$

where $\pi_c = \theta_c - \sum_{i\in[\theta_c]} X_c^i$.

Similarly, we can get the unbiased estimator of $F_0(S)$ as

$$\hat{F}_0(S,\mathcal{R}) = -\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i}, \tag{10}$$

In the following, we consider Eq.(9) and Eq.(10) as our objective functions to deal with the fair IM problem.

4.3 The Proposed Algorithm FIMM

For a given seed set S, let $\varphi[c]$ denote the number of all u-rooted ($u \in V_c$) RR sets covered by S, and $\kappa[v][c]$ denote the number of all u-rooted ($u \in V_c$) RR sets that covered by v ($v \in V \setminus S$) but not by S, then the marginal fair influence gain of v w.r.t. $F_\alpha(S)$ and $F_0(S)$ is

$$\begin{aligned}\hat{F}_\alpha(v|S) &= \sum_{c\in\mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \prod_{i=0}^{n-1} \frac{\theta_c - \kappa[v][c] - \varphi[c] - i}{\theta_c - i} \right) \\ &\quad - \sum_{c\in\mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \prod_{i=0}^{n-1} \frac{\theta_c - \varphi[c] - i}{\theta_c - i} \right) \\ &= \sum_{c\in\mathcal{C}} \alpha n_c \sum_{n=1}^{\theta_c} \eta(n,\alpha) \\ &\quad \cdot \left(\prod_{i=0}^{n-1} \frac{\theta_c - \varphi[c] - i}{\theta_c - i} - \prod_{i=0}^{n-1} \frac{\theta_c - \kappa[v][c] - \varphi[c] - i}{\theta_c - i} \right), \end{aligned} \tag{11}$$

$$\begin{aligned}\hat{F}_0(v|S) &= -\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\theta_c - \kappa[v][c] - \varphi[c] - i}{\theta_c - i} \\ &\quad + \sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\theta_c - \varphi[c] - i}{\theta_c - i}, \\ &= \sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \\ &\quad \cdot \left(\prod_{i=0}^{n-1} \frac{\theta_c - \varphi[c] - i}{\theta_c - i} - \prod_{i=0}^{n-1} \frac{\theta_c - \kappa[v][c] - \varphi[c] - i}{\theta_c - i} \right). \end{aligned} \tag{12}$$

If we only consider $n = 2$, the marginal fair influence gain becomes

$$\begin{aligned}\hat{F}_\alpha^2(v|S) &= \sum_{c\in\mathcal{C}} \alpha n_c \kappa[v][c] \cdot \left(\frac{(4-2\alpha)\theta_c - 3 + \alpha}{\theta_c(\theta_c - 1)} \right. \\ &\quad \left. - \frac{(1-\alpha)\varphi[c] + (1-\alpha)(\varphi[c] + \kappa[v][c])}{\theta_c(\theta_c - 1)} \right), \end{aligned} \tag{13}$$

$$\hat{F}_0^2(v|S) = \sum_{c\in\mathcal{C}} n_c \kappa[v][c] \cdot \frac{4\theta_c - 3 - \varphi[c] - (\varphi[c] + \kappa[v][c])}{\theta_c(\theta_c - 1)}. \tag{14}$$

Note that $\kappa[v][c]$ is monotonically decreasing, and both $\varphi[c]$ and $\varphi[c] + \kappa[v][c]$ are monotonically increasing with the expansion of S. Therefore, both $F_\alpha^2(v|S)$ and $F_0^2(v|S)$ are still monotonically decreasing, which enables a lazy-updating strategy to effectively select seeds with the maximal marginal fair influence gain.

Following the calculation of the marginal fair influence gain, when generating RR sets, we have to count $\kappa[v][c]$ which indicates the community-wise coverage for v and record $\eta[v]$ which indicates the linked-list from v to all its covered RR sets, as shown in Algorithm 1. As shown in lines 6∼9, when generating a random v-rooted RR set $RR(v)$, we count all nodes $u \in RR(v)$ and raise all $\kappa[u][c(v)]$ by 1, where $c(v)$ indicates v's community label. It should be noted that modifying $\kappa[u][c(v)]$

Algorithm 1: RR-Generate: Generate RR sets

```
Input: Graph $G = (V, E, p)$, community $\mathcal{C}$, budget $k$,
       number of RR sets for each community $\theta_c$
Output: RR sets $\mathcal{R}$, community-wise coverage $\kappa$,
        linked-list $\eta$ from nodes to covered RR sets
1  Initialize $\kappa[v][c] = 0$ for all $v \in V$, $c \in \mathcal{C}$;
2  Initialize $\eta[v] = \varnothing$ for all $v \in V$;
3  $\mathcal{R} = \varnothing$;
4  for $c \in \mathcal{C}$ do
5      for $i = 1$ to $\theta_c$ do
6          Select a random node $v$ in community $c$;
7          Sample a random RR set $R_i = RR(v)$;
8          for $u \in R$ do
9              $\kappa[u][c(v)] = \kappa[u][c(v)] + 1$;
10         $\mathcal{R} = \mathcal{R} \cup \{R_i\}$;
11         $\eta[v] = \eta[v] \cup \{R_i\}$;
```

[1] $\xi(\theta_c)$ is sufficiently small and can be viewed as 0 since θ is usually large enough. Thus, $\hat{\mathbb{E}}[X_c]^\alpha$ can be treated as an unbiased estimator.

can be accomplished simultaneously when generating $RR(v)$ by the reverse influence sampling.

Based on the RR sets generated by Algorithm 1, we present our FIMM algorithm (Algorithm 2) to select k seed nodes that maximize Eq.(9) and Eq.(10) through a greedy approach, *i.e.*, iteratively selecting a node with the maximum alternative marginal fair influence gain as presented in Eq.(11) and Eq.(12). Apparently, it costs $O(C)$ to calculate $\hat{F}_\alpha(v|S)$ or $\hat{F}_0(v|S)$ for any v where C is the number of communities. When C is small (*i.e.*, a constant), it would be efficient to compute $\hat{F}(v|S)$ for all $v \in V$ in $O(Cn_G)$. Additionally, since $\hat{F}(S,\mathcal{R})$ is submodular and monotone, we can adopt a lazy-update strategy [24] that selects v with the maximal $\hat{F}(v|S)$ as a seed node if $\hat{F}(v|S)$ is still the maximum after updating. This lazy-update strategy (lines 10∼12) can significantly reduce redundant time costs, achieving empirical speeds up to 700 times faster than a simple greedy algorithm [24].

There are two vital counting arrays in Algorithm 2, *i.e.*, $\varphi[c]$ and $\kappa[v][c]$. $\varphi[c]$ records and updates the number of RR sets covered by S in community-wise. By lines 20∼24, $\kappa[v][c]$ keeps updating and always indicates the extra coverage of v on all u-rooted ($u \in V_c$) RR sets besides S. It establishes a convenient way for updating $\varphi(c)$ that only needs to increase $\varphi(c)$ by $\kappa[v][c]$ where v is the newly selected node for all $c \in \mathcal{C}$. If we denote the original community-wise coverage as κ', which means $(\sim, \kappa', \sim)$ =RR-Generate$(G, \mathcal{C}, k, \theta_c)$, then it holds $\kappa'[v][c] = \kappa[v][c] + \varphi[c]$ for all $v \in V$ and $c \in \mathcal{C}$.

Now we discuss the time complexity of Algorithm 2. Apparently, the initialization of $\gamma(v)$ in line 2 can be conducted simultaneously with generating RR sets in line 1. Lines 3∼5 take $O(C + \theta + n_G)$ for initialization where θ is the number of RR sets in $\mathcal{R}$. Line 7 takes k rounds to select k seed nodes. In each round, lines 8∼17 take $O(n_G)$ to select v with maximal $\gamma(v)$, and take $O(Cn_G)$ to calculate $\gamma(v)$ for at most n nodes. Lines 18∼19 take $O(C)$ to update $\varphi[c]$. Since $covered[R]$ will be set to $false$ if ever covered by S, each RR set would only be traversed at most once during lines 20∼24, which results in a $O(\sum_{R\in\mathcal{R}}|R|)$ time cost independent of k. In summary, the seed selection process in Algorithm 2 takes $O(C+\theta+n_G)+O(kn_G)+O(kCn_G)+O(kC)+O(\sum_{R\in\mathcal{R}}|R|) = O(kCn_G + \sum_{R\in\mathcal{R}}|R|)$.

Algorithm 2: FIMM: Fair Influence Maximization

```
   Input: Graph $G = (V, E, p)$, community $\mathcal{C}$, budget $k$,
          approximation parameter $Q$
   Output: Seed set $S$
1  $(\mathcal{R}, \kappa, \eta) =$ RR-Generate$(G, \mathcal{C}, k, \theta_c)$
2  Initialize $\gamma(v)$ according to Eq.(11-12) for all $v \in V$;
   //indicating initial marginal gain
3  Initialize $\varphi[c] = 0$ for all $c \in \mathcal{C}$; //indicating the
   number of covered RR sets rooted in $c$
4  Initialize $covered[R] = false$ for all $R \in \mathcal{R}$;
5  Initialize $updated[v] = true$ for all $v \in V$;
6  $S = \varnothing$;
7  for $i = 1$ to $k$ do
8      while true do
9          $v = \arg\max_{u \in V \setminus S} \gamma(u)$;
10         if $updated(v) == false$ then
11             Updating $\gamma(v)$ according to Eq.(11-12);
12             $updated(v) = true$;
13         else
14             $S = S \cup \{v\}$;
15             for $v \in V$ do
16                 $updated(v) = false$;
17             break;
18     for $c \in \mathcal{C}$ do
19         $\varphi[c] = \varphi[c] + \kappa[v][c]$;
20     for all $R \in \eta[v] \wedge covered[R] == false$ do
21         $covered[R] = true$;
22         $r = root(R)$;
23         for all $u \in R \wedge u \neq v$ do
24             $\kappa[u][c(r)] = \kappa[u][c(r)] - 1$;
```

4.4 Number of RR sets

In this subsection, we discuss the number of RR sets needed to approximate the fair influence with high probability. Let OPT_α and OPT_0 denote the optimal solution of the FIM problem with S^* denote the corresponding optimal seed set, *i.e.*, $OPT_\alpha = F_\alpha(S^*) = \sum_{c\in\mathcal{C}} n_c \boldsymbol{u}_c(S^*)^\alpha = \sum_{c\in\mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\infty} \eta(n,\alpha)\big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)$ for $0 < \alpha < 1$, and $OPT_0 = F_0(S^*) = \sum_{c\in\mathcal{C}} n_c \ln(\boldsymbol{u}_c(S^*)) = -\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\infty} \frac{1}{n}\big(1 - \boldsymbol{u}_c(S^*)\big)^n$ for $\alpha = 0$. It should be noted that S^* can be different for $OPT_\alpha \in (0, +\infty)$ and $OPT_0 \in (-\infty, 0)$. Since this paper deals with the fair IM problem, we thus assume that the maximal community utility $max_{c\in\mathcal{C}}\boldsymbol{u}_c(S^{\#}) \geq max_{c\in\mathcal{C}}\boldsymbol{u}_c(S^*)$ of an arbitrary seed set $S^{\#}$ would not be too big.

Lemma 3. *Let $\delta_1 \in (0,1)$, $\varepsilon_1 \in (0,1)$, and $\theta_1 = \frac{12Q^2\ln(C/\delta_1)}{\varepsilon_1^2(1-b)}$ where Q is the approximation parameter, $b = max(\boldsymbol{u}_c(S^*)), \forall c \in \mathcal{C}$, and $S^* = \text{argmax}_{S:|S|\leq k} F_\alpha(S)$ denotes the optimal solution for the FIM problem based on $\mathcal{R}$, then $\hat{F}_\alpha(S^*, \mathcal{R}) \geq (1 - \varepsilon_1)\cdot OPT_\alpha$ holds with at least $1 - \delta_1$ probability if $\theta \geq C\theta_1$.*

Lemma 4. *Let $\delta_2 \in (0,1)$, $\varepsilon_2 = (\frac{e}{e+1})\varepsilon - \varepsilon_1$, and $\theta_2 = \frac{8Q^2\ln(C\binom{n_G}{k}/\delta_2)}{\varepsilon_2^2(1-b_0)}$ where Q is the approximation parameter, $b_0 = max(\boldsymbol{u}_c(S^{\#})), \forall c \in \mathcal{C}$ where $S^{\#}$ could be an arbitrary fair solution. For each bad S (which indicates $F_\alpha(S) < (1 - 1/e - \varepsilon)\cdot OPT_\alpha$), $\hat{F}_\alpha(S, \mathcal{R}) \geq (1 - 1/e)(1 - \varepsilon_1)\cdot OPT_\alpha$ holds with at most $\delta_2/\binom{n_G}{k}$ probability if $\theta \geq C\theta_2$.*

Please refer to the Appendix for the detailed proof of Lemma 3 and Lemma 4.

Theorem 1. *For every $\varepsilon > 0$, $\ell > 0$, $0 < \alpha < 1$, and $Q \geq 2$, by setting $\delta_1 = \delta_2 = 1/2n_G^\ell$ and $\theta \geq C \cdot max(\theta_1, \theta_2)$, the output S of FIMM satisfies $F_\alpha(S) \geq (1 - 1/e - \varepsilon)\, F_\alpha(S^*)$, where S^* denotes the optimal solution with probability at least $1 - 1/n_G^\ell$.*

Proof. Combining Lemma 3 and Lemma 4, we have $\hat{F}_\alpha(S, \mathcal{R}) \geq (1 - 1/e - \varepsilon)\cdot OPT_\alpha$ at least $1 - \delta_1 - \delta_2$ probability based on the union bound. If we set $\delta_1 = \delta_2 = 1/2n_G^\ell$, then, following the standard analysis of IMM, our FIMM algorithm provides $(1 - 1/e - \varepsilon)$-approximation with probability at least $1 - 1/n_G^\ell$. □

Since $F_0(S)$ is a non-positive monotone submodular function, it leads to a greedy solution with a $1 - 1/e$ approximation, plus an additional $F_0(\varnothing)/e$. However, as $ln(0)$ is undefined, $F_0(\varnothing)$ cannot be computed. To address the problem, we assume the existence of an activated virtual node that connects to certain marginal nodes in each community, where those marginal nodes are unlikely to be influenced by S^*. Consequently, $u_c(S)$ would be positive when $S = \varnothing$. Let $u_c(\varnothing) = u_c(S^*)^2$ (*e.g.*, if $u_c(S^*)$ is 0.2, $u_c(\varnothing)$ is only 0.04), then we have $F_0(\varnothing) = 2F_0(S^*)$, which yields a greedy solution with a $1 + 1/e$ approximation.

Lemma 5. *Let $\delta_1 \in (0,1)$, $\varepsilon_1 \in (0,1)$, and $\theta_1 = \frac{12Q^2\ln(C/\delta_1)}{\varepsilon_1^2(1-b)}$ where Q is the approximation parameter, $b = max(\boldsymbol{u}_c(S^*)), \forall c \in \mathcal{C}$, and $S^* = \text{argmax}_{S:|S|\leq k} F_0(S)$ denotes the optimal solution for the*

FIM problem based on $\mathcal{R}$, then $\hat{F}_0(S^*, \mathcal{R}) \geq (1+\varepsilon_1) \cdot OPT_0$ holds with at least $1-\delta_1$ probability if $\theta \geq C\theta_1$.

Lemma 6. *Let $\delta_2 \in (0,1)$, $\varepsilon_2 = (\frac{e}{e-1})\varepsilon - \varepsilon_1$, and $\theta_2 = \frac{8Q^2 \ln(C\binom{n_G}{k}/\delta_2)}{\varepsilon_2^2(1-b_0)}$ where Q is the approximation parameter, $b_0 = max(\boldsymbol{u}_c(S^{\#})), \forall c \in \mathcal{C}$ where $S^{\#}$ could be an arbitrary fair solution. For each bad S (which indicates $F_0(S) < (1+1/e+\varepsilon) \cdot OPT_0$), $\hat{F}_0(S,\mathcal{R}) \geq (1+1/e)(1+\varepsilon_1) \cdot OPT_0$ holds with at most $\delta_2/\binom{n_G}{k}$ probability if $\theta \geq C\theta_2$.*

Similarly, please refer to the Appendix for the detailed proof of Lemma 5 and Lemma 6.

Theorem 2. *For every $\varepsilon > 0$, $\ell > 0$, $0 < \alpha < 1$, and $Q \geq 2$, by setting $\delta_1 = \delta_2 = 1/2n_G^{\ell}$ and $\theta \geq C \cdot max(\theta_1, \theta_2)$, the output S of* FIMM *satisfies $F_0(S) \geq (1+1/e+\varepsilon)\,F_0(S^*)$, where S^* denotes the optimal solution with probability at least $1-1/n_G^{\ell}$.*

Proof. Combining Lemma 5 and Lemma 6, we have $\hat{F}_0(S,\mathcal{R}) \geq (1+1/e+\varepsilon) \cdot OPT_0$ at least $1-\delta_1-\delta_2$ probability based on the union bound. If we set $\delta_1 = \delta_2 = 1/2n_G^{\ell}$, then, following the standard analysis of IMM, our FIMM algorithm provides $(1+1/e+\varepsilon)$-approximation with probability at least $1-1/n_G^{\ell}$. □

It should be noted that $F_\alpha(S)$ is always positive and $F_0(S)$ is always negative for $S \neq \varnothing$, thus yielding two different lower bounds in Theorem 1 and Theorem 2. In addition, for both Theorem 1 and Theorem 2, if we set $\delta_1 = \delta_2 = \frac{1}{2n_G^{\ell}}$ and $\varepsilon_1 = \varepsilon \cdot \frac{e}{e-1} \cdot \frac{\sqrt{3}\tau_1}{\sqrt{3}\tau_1+\sqrt{2}\tau_2}$ where $\tau_1 = \sqrt{\ln C + \ell \ln n_G + \ln 2}$ and $\tau_2^2 = \tau_1^2 + \ln\binom{n_G}{k}$, then a possible setting of θ could be $\theta = (\frac{e-1}{e})^2 \cdot \frac{4CQ^2(\sqrt{3}\tau_1+\sqrt{2}\tau_2)^2}{\varepsilon^2(1-b_0)}$, which satisfies $\theta \geq C \cdot max(\theta_1, \theta_2)$.

5 Experiments

5.1 Dataset

Email: The Email dataset [39] is generated using email data from a large European research institution, where every node is a member of the research institution and a directed edge (v,u) indicates that v has sent u at least one email. It contains 1,005 nodes and 25,571 directed edges. Moreover, this dataset also contains "ground-truth" community memberships of nodes, where each member belongs to exactly one of 42 departments at the research institute.

UVM: The UVM dataset [40] is generated using the Facebook social network data in UVM (University of Vermont), where every node is a member (Faculty or Student) of UVM. The network is preprocessed by Lin. *et al.* [19] who remove the nodes without user information in the network profile. It contains 7,322 nodes and 191,197 undirected edges and is either divided by the Status feature into Faculty (12%) and Student (88%) or divided by the Grade feature into Senior (40%) and Junior (60%).

UCSC: The UCSC dataset [40] is generated using the Facebook social network data in UCSC (University of California at Santa Cruz), where every node is a member of UCSC. The network contains 8,990 nodes and 224,545 undirected edges and is either divided by the Status feature into Faculty (10%) and Student (90%) or divided by the Gender feature into Male (45%) and Female (55%).

Flixster: The Flixster dataset [41] is a network of American social movie discovery services. To transform the dataset into a weighted graph, each user is represented by a node, and a directed edge from node u to v is formed if v rates one movie shortly after u does so on the same movie. It contains 29,357 nodes and 212,614 directed edges. It also provides the learned influence probability between each node pair, which can be incorporated into the IC model. Since it has no community information, we construct the biased community structure by categorizing individuals according to their susceptibility of being influenced to highlight the level of inequality and get two different divisions consisting of 10 and 100 communities.

Amazon: The Amazon dataset [42] is collected based on *Customers Who Bought This Item Also Bought* feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains an undirected edge between i to j. The dataset also provides the ground-truth community structure which indicates the product categories. The original network has 334,863 nodes and 925,872 undirected edges. After Pruning low-quality communities (whose size is no more than 10 nodes), the Amazon network tested in our experiments contains 9,239 nodes, 29,370 edges, and 229 communities.

Youtube: The Youtube dataset [42] is a network of the video-sharing web site that includes social relationships. Users form friendship each other and users can create groups which other users can join. The friendship between users is regarded as undirected edges and the user-defined groups are considered as ground-truth communities. The original network has 1,134,890 nodes and 2,987,624 undirected edges. After screening high-quality communities, it remains 20,707 nodes, 95,403 edges, and 379 communities.

DBLP: This paper adopts two different DBLP datasets, which will be referred to as DBLP1 and DBLP2 in the following context. The DBLP1 dataset [42] is the co-authorship network where two authors are connected if they have ever published a paper together. Publication venues, such as journals or conferences, define an individual ground-truth community, and authors who publish to a certain journal or conference form a community. The original network has 717,080 nodes and 1,049,866 undirected edges. We also perform the network pruning and finally obtain 59,028 nodes, 215,335 edges, and 193 communities. Similar to DBLP1, the DBLP2 dataset [43] is also the co-authorship network but attached with a gender feature. The network has 280,200 nodes (23% female and 77% male) and 750,601 undirected edges.

5.2 Evaluation Metrics

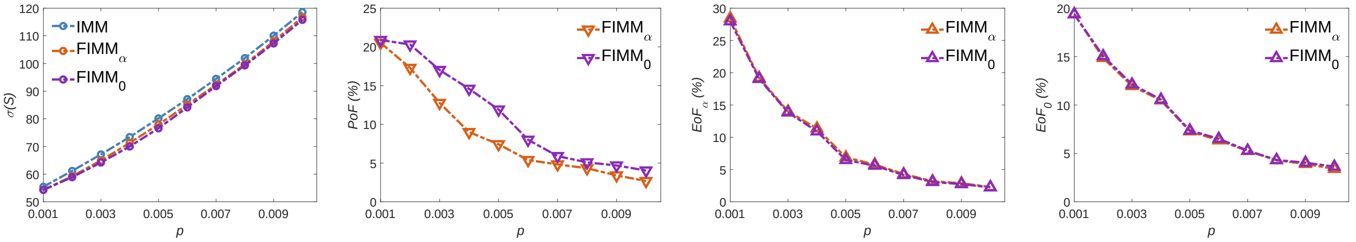
Let S_I denote the seed set returned by IMM [18] (which is one of the state-of-the-art IM algorithms with $1-1/e$ theoretical guarantee), S_F denote the seed set returned by FIMM, the performance of S_F towards fairness can be evaluated via the Price of Fairness (PoF) and the Effect of Fairness (EoF) as

$$PoF = \frac{\sigma(S_I) - \sigma(S_F)}{\sigma(S_I) - k},$$
$$EoF_\alpha = \frac{F_\alpha(S_F) - F_\alpha(S_I)}{F_\alpha(S_I) - k},$$
$$EoF_0 = \frac{F_0(S_F) - F_0(S_I)}{-F_0(S_I) - k}.$$

where $|S_I| = |S_F| = k$, $\sigma(\cdot)$ denotes the influence spread, and both $F_\alpha(\cdot)$ (≥ 0) and $F_0(\cdot)$ (≤ 0) are the fair influence. PoF measures the loss of influence spread $\sigma(S)$ by calculating the relative gap between a fair IM algorithm and the IMM algorithm. In contrast, EoF$_\alpha$ and EoF$_0$ measure the improvement of fair influence (corresponding to $F_\alpha(S)$ and $F_0(S)$, respectively) by calculating the relative gap between a fair IM algorithm and the IMM algorithm. Intuitively, PoF implies how much it costs to access fairness and two EoFs imply to what extent it steps towards fairness.

5.3 Results

We test IMM and our proposed FIMM algorithm in the experiment, where FIMM$_\alpha$ and FIMM$_0$ denote our algorithm framework combined with $F_\alpha(S)$ and $F_0(S)$, respectively. In all tests,



(a) $\sigma(S)$ (b) PoF (c) EoF_α (d) EoF_0

Fig. 1. Results on Email network (testing influence probability p).

we run 20,000 Monte-Carlo simulations to evaluate both the influence spread and the fair influence under the IC model. We test different influence probability p, inequality aversion parameter α, and seed budget k under different datasets.

5.3.1 Email

The Email network has only 1005 nodes but is allocated into 42 communities. The two largest communities have 109 and 92 nodes, while the three smallest communities have only 1 node. For the Email network, we set $\alpha = 0.5$, $k = 50$, and apply the Uniformed IC model where the influence probability is the same across all edges. We test different probabilities that range from 0.001 to 0.01 with the step of 0.001. The results include the influence spread $\sigma(S)$, the Price of Fairness PoF, and the Effect of Fairness EoF_α and EoF_0, which are shown in Fig. 1.

As the influence probability p increases, both PoF and EoF show a downward trend. This may be attributed to the increased challenges faced by disadvantaged communities in being influenced when p is small. Besides, though FIMM$_\alpha$ holds a lower PoF than FIMM$_0$ across different p, they give nearly the same performance on both EoF_α and EoF_0, where the mean EoF_α are 9.83% and 9.60%, and the mean EoF_0 are 8.72% and 8.83% for FIMM$_\alpha$ and FIMM$_0$, respectively. But still, FIMM$_\alpha$ is slightly better in terms of EoF_α while FIMM$_0$ is slightly better in terms of EoF_0, which accords with their corresponding objective functions.

Moreover, we compare our method with equality-based fairness which is in favor of fair result-aware seeding to explore the difference between welfare fairness and equality fairness. The Equality asks to divide the budget k proportionally to the cluster sizes, *i.e.*, $|S \cap V_c| \approx k \cdot n_c/n_G$. To select seeds under the Equality criterion, we adopt two strategies: (1) selecting the highest-degree node within each community, referred to as $S_{\text{C-HD}}$; and (2) running the IMM algorithm independently within each community, referred to as $S_{\text{C-IMM}}$. Since FIMM$_\alpha$ and FIMM$_0$ give rather similar performances, we only involve the results of FIMM$_\alpha$ and test PoF and EoF_α.

Table 2 below exhibits the comparison results, where S_{FIMM}, $S_{\text{C-HD}}$, and $S_{\text{C-IMM}}$ refer to seeds selected by our FIMM$_\alpha$ method, seeds selected by community-aware highest degree, and community-aware IMM, respectively. PoF and EoF_α are calculated based on the IMM algorithm.

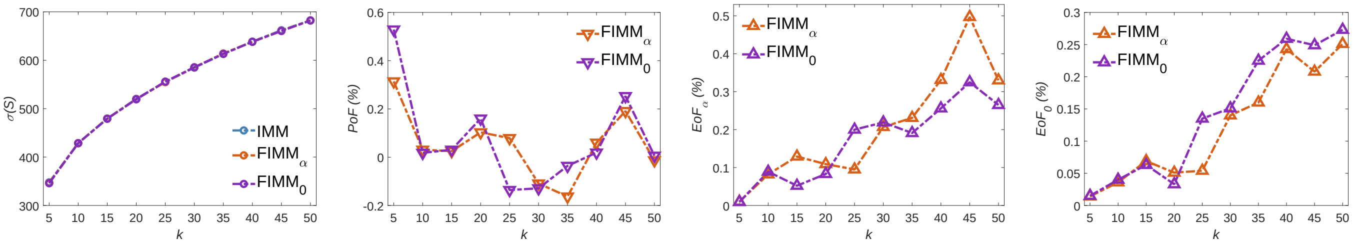
As can be seen from the results, both PoF (the lower the better) and EoF_α (the higher the better) of S_{FIMM} are always better than that of $S_{\text{C-HD}}$ and $S_{\text{C-IMM}}$. In other words, $S_{\text{C-HD}}$ pays more price of fairness yet achieves a lower degree of fairness. Compared with $S_{\text{C-HD}}$, $S_{\text{C-IMM}}$ yields a better performance. Moreover, as p increases, the fair influence of both $S_{\text{C-HD}}$ and $S_{\text{C-IMM}}$ is even lower (leading to negative EoF_α) than that of IMM, which does not even contribute to fairness at all. The reason is that community-aware seeding only highlights fairness in the process of seed allocation, while welfare fairness (also, maximin fairness and equity fairness) highlights fairness in the spreading results. The former could have an explicit fair distribution in seeding, but may still lead to unfair results.

TABLE 2
Comparison with equality-based methods

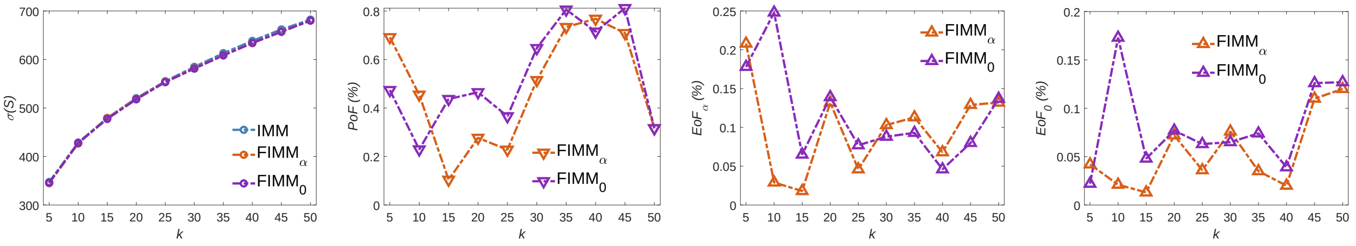
p	PoF			EoF_α		
	S_{FIMM}	$S_{\text{C-HD}}$	$S_{\text{C-IMM}}$	S_{FIMM}	$S_{\text{C-HD}}$	$S_{\text{C-IMM}}$
0.001	20.53%	31.99%	21.56%	27.97%	20.32%	20.94%
0.002	17.29%	31.50%	20.56%	19.04%	13.92%	15.14%
0.003	12.77%	30.73%	20.76%	13.84%	10.09%	11.72%
0.004	9.00%	29.63%	19.11%	10.92%	4.40%	6.73%
0.005	7.43%	28.85%	17.63%	6.48%	2.72%	5.58%
0.006	5.39%	28.25%	17.96%	5.60%	0.15%	3.26%
0.007	4.83%	27.06%	16.21%	4.15%	-1.75%	1.84%
0.008	4.33%	26.03%	15.77%	3.07%	-3.13%	0.64%
0.009	3.41%	24.65%	14.88%	2.74%	-4.63%	-0.73%
0.01	2.68%	24.07%	15.64%	2.24%	-5.18%	-1.51%

5.3.2 UVM & UCSC

These two networks can be partitioned in two distinct ways, each resulting in a structure of two communities. Specifically, the UVM network, comprising 7,322 nodes, can be divided either into Faculty (899 nodes) and Student (6,423 nodes) or into Senior (2,884 nodes) and Junior (4,438 nodes). Similarly, the UCSC network, with 8,990 nodes, can be split either into

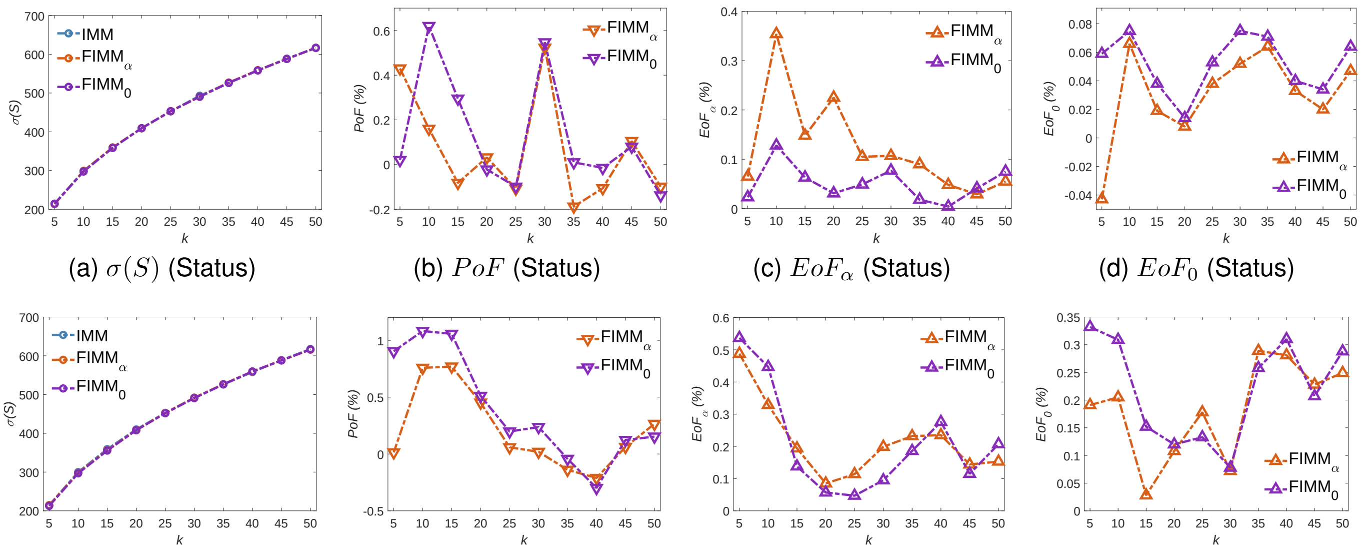


(a) $\sigma(S)$ (Status) (b) PoF (Status) (c) EoF_α (Status) (d) EoF_0 (Status)



(a) $\sigma(S)$ (Grade) (b) PoF (Grade) (c) EoF_α (Grade) (d) EoF_0 (Grade)

Fig. 2. Results on UVM network (testing seed budget k).



(a) $\sigma(S)$ (Status) (b) PoF (Status) (c) EoF_α (Status) (d) EoF_0 (Status)

(e) $\sigma(S)$ (Gender) (f) PoF (Gender) (g) EoF_α (Gender) (h) EoF_0 (Gender)

Fig. 3. Results on UCSC network (testing seed budget k).

Faculty (930 nodes) and Student (8,060 nodes) or into Male (4,006 nodes) and Female (4,984 nodes) groups. Since both networks have a large average degree, we set $p = 0.01$, $\alpha = 0.5$, and test the variation of the seed budget k from 5 to 50. The results on these two networks are shown in Fig. 2 and Fig. 3.

As shown in these two figures, the influence spread of the three methods remains consistently similar, regardless of how the communities are divided (*i.e.*, by Status or by Gender). Correspondingly, the PoF of FIMM$_\alpha$ and FIMM$_0$ is almost always lower than 1%, showing that both algorithms return high-quality seed sets even in terms of influence spread. This consistency may be due to the presence of only two communities, which are divided based on node attributes rather than community criteria such as modularity. As a result, the nodes are well-connected across different groups. In such a case, the high-quality seeds selected by IMM are also likely to influence a significant ratio of nodes across different communities. Therefore, both EoF_α and EoF_0 are always lower than 1% as well, revealing that the gap of fairness between IMM and FIMM is relatively small on these two networks.

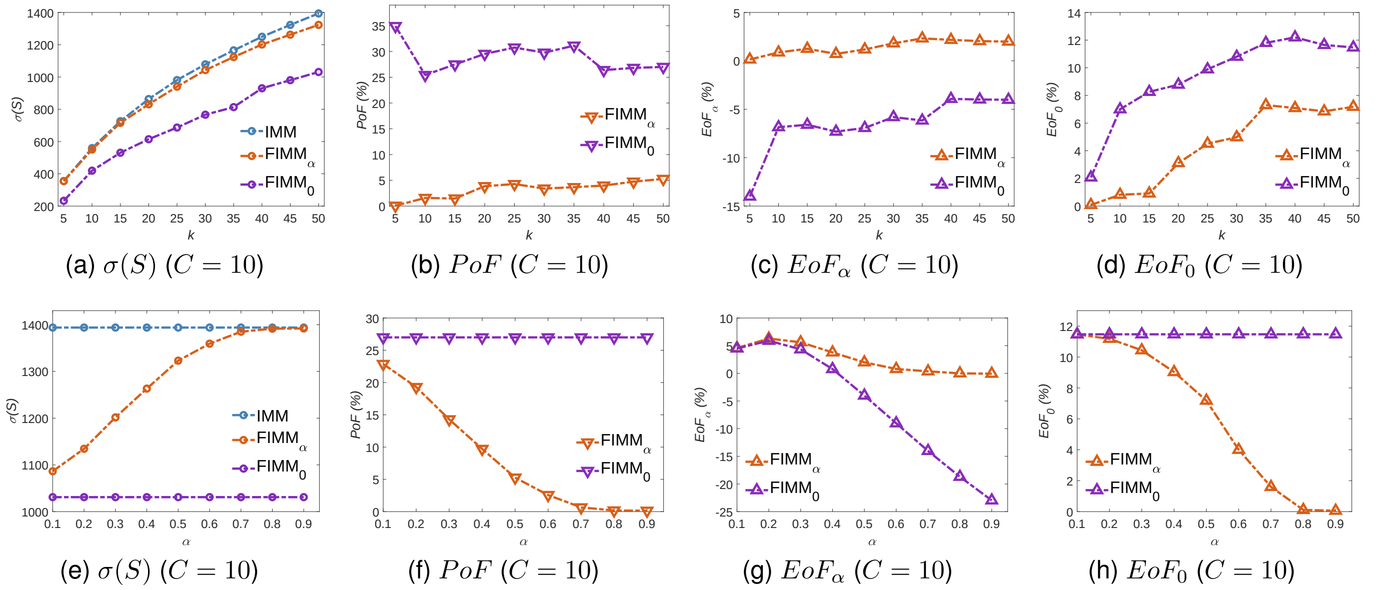
However, the results still show a trend that FIMM$_\alpha$ usually produces a better EoF_α while FIMM$_0$ achieves a higher EoF_0. Meanwhile, FIMM$_\alpha$ usually pays a lower price in influence spread. The reason is that FIMM$_0$ requires the utility of any community to be non-zero, which significantly affects the seed selection process at the initial stage, causing a lower influence spread.

5.3.3 *Flixster*

The Flixster dataset provides the learned influence probabilities. Thus, we test the inequality aversion parameter α which ranges from 0.1 to 0.9 with the step of 0.1 under $k = 50$. We also test the seed budget k that ranges from 5 to 50 with the step of 5 under $\alpha = 0.5$.

To highlight the level of inequality between different communities, we construct the biased community structure by dividing individuals according to their degree of being influenced. Since we generate the same number of RR sets for each node, we can calculate the total size of RR sets for each node and sort them decreasingly. The size indicates how easily a node can be influenced by random seeds. We then divide those sorted nodes into C communities where all communities have the same size. Consequently, nodes in the head communities are easier to be influenced and nodes in the tail communities can hardly be influenced. In practice, assuming we divide the network into C communities, the $\lceil n/C \rceil$ individuals that are easiest to be influenced are put into V_{c1}. Then, we put the $\lceil n/C \rceil$ individuals that are next easiest to be influenced are put into V_{c2}. Analogously, the network will ultimately be divided into C communities where individuals in different communities exhibit a gradient probability pattern of being influenced.

Following the above strategy, we divide the Flixster network into 10 and 100 communities, the results of which are shown in Fig. 4 and Fig. 5. It should be noted that the change of α does not affect IMM and FIMM$_0$, therefore their performance in Fig. 4 (e),(f),(h) and Fig. 5 (d),(e) remains a straight line. Besides, when the network is divided into 100 communities,



(a) $\sigma(S)$ $(C = 10)$ (b) PoF $(C = 10)$ (c) EoF_α $(C = 10)$ (d) EoF_0 $(C = 10)$

(e) $\sigma(S)$ $(C = 10)$ (f) PoF $(C = 10)$ (g) EoF_α $(C = 10)$ (h) EoF_0 $(C = 10)$

Fig. 4. Results on Flixster network with 10 communities (testing seed budget k and aversion parameter α).

the seeds selected by IMM fail to cover all the groups, which makes it impossible to calculate EoF_0 for FIMM$_\alpha$ and FIMM$_0$.

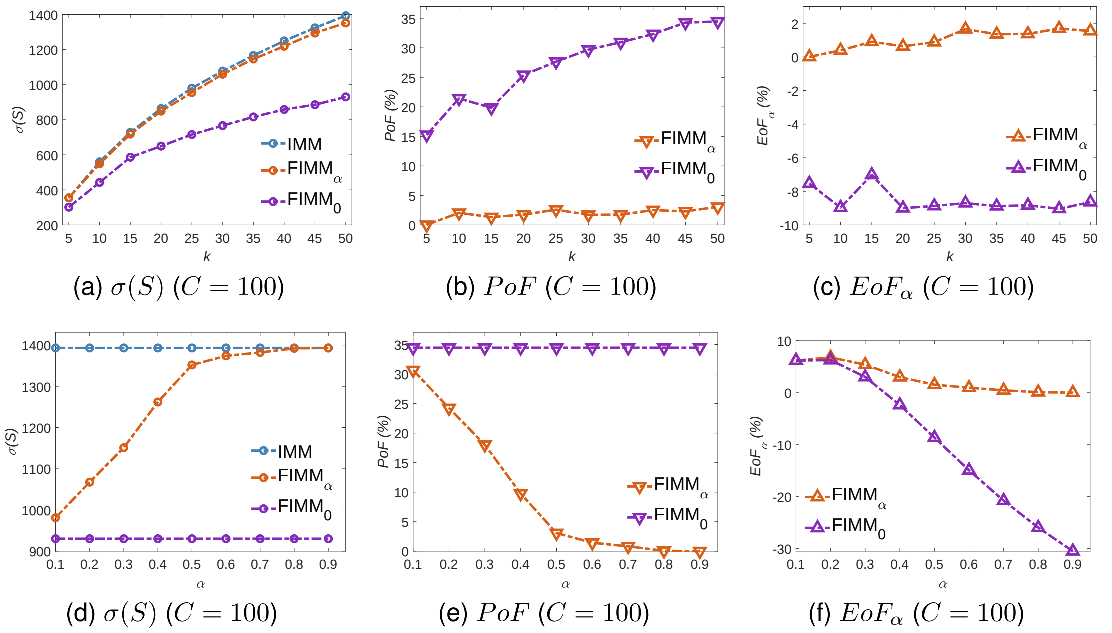
As shown in Fig. 4 and Fig. 5, the PoF of FIMM$_0$ is always significantly higher than that of FIMM$_\alpha$. The reason is that FIMM$_0$ must attempt to influence all the communities even when k is very small, resulting in a greater loss of influence spread. Meanwhile, FIMM$_\alpha$ and FIMM$_0$ exhibit clear superiority in their respective objective functions, *i.e.*, $F_\alpha(\cdot)$ and $F_0(\cdot)$.

As k increase, PoF, EoF_α, and EoF_0 of both FIMM$_\alpha$ and FIMM$_0$ all tend to rise. Since communities are divided based on their susceptibility to influence, selecting more seeds makes it easier to achieve a fairer utility distribution. When α increases, PoF, EoF_α, and EoF_0 of FIMM$_\alpha$ shows a downward trend. The reason lies that communities experience greater promotions in fair influence when α is smaller. Moreover, there is hardly any fairness when $\alpha \geq 0.7$ where the gap between $\boldsymbol{u}^\alpha$ and $\boldsymbol{u}$ is just too small.

5.3.4 Amazon

The Amazon network is a disconnected network that contains 42 connected components with 229 communities. Therefore, we only test FIMM$_\alpha$ and $F_\alpha(\cdot)$ on this dataset since $F_0(\cdot)$ is unavailable unless all communities are influenced. Like the above experiments, α is set to 0.5, and k ranges from 5 to 50 with step of 5. Since the network has a more complex topology, we set $p(v_i, v_j) = 1/d_{in}(v_j)$ where d_{in} denotes the in-degree as the influence probability following the weighted IC model [23]. Corresponding results are shown in Fig. 6.

Generally, FIMM tends to produce a noticeably fairer output when k is small. It reflects the idea that enforcing fairness as a

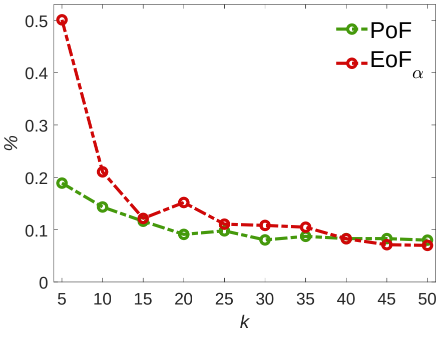


(a) $\sigma(S)$ $(C = 100)$ (b) PoF $(C = 100)$ (c) EoF_α $(C = 100)$

(d) $\sigma(S)$ $(C = 100)$ (e) PoF $(C = 100)$ (f) EoF_α $(C = 100)$

Fig. 5. Results on Flixster network with 100 communities (testing seed budget k and aversion parameter α).

constraint becomes easier when abundant resources are available. However, there are also some exceptions where smaller k leads to a lower EoF, *e.g.*, $k = 15$. This may be attributed to the fact that the seed selection in FIMM follows a pattern of remedying the previously fair solutions in each round.



(a) Amazon

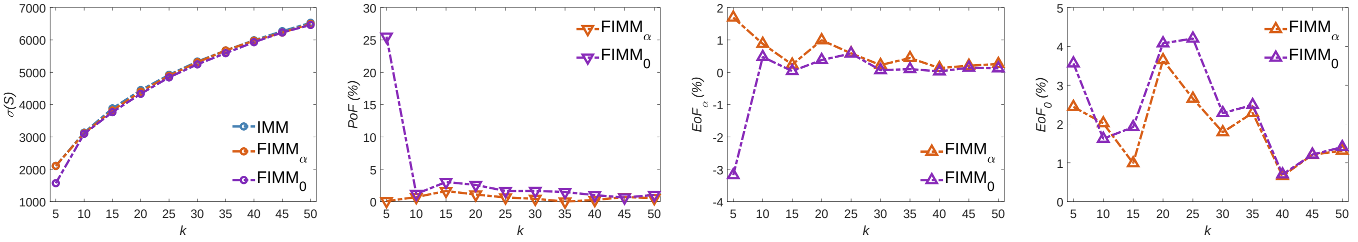
Fig. 6. Results on Amazon network (testing seed budget k).

5.3.5 Youtube & DBLP

Different from the Amazon network, Youtube, DBLP1, and DBLP2 are fully connected networks, where both $F_\alpha(\cdot)$ and $F_0(\cdot)$ are tested. Similarly, we set $\alpha = 0.5$ and $p(v_i, v_j) = 1/d_{in}(v_j)$. We test k ranging from 5 to 50 (step = 5) on Youtube and from 10 to 100 (step = 10) on both DBLP1 and DBLP2 as shown in Fig. 7, 8, and 9, respectively.

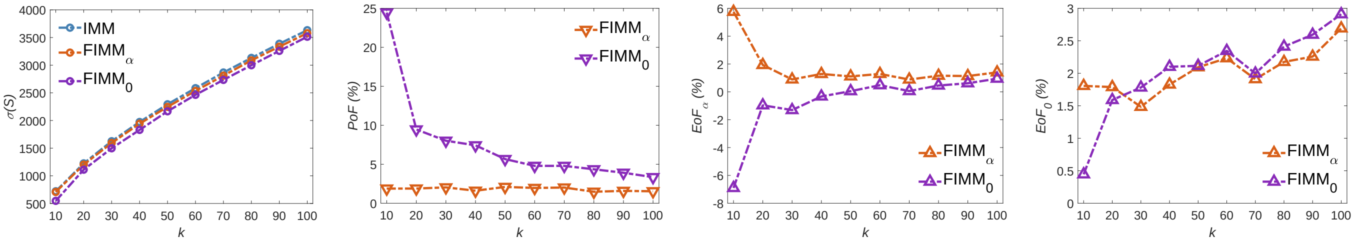
On Youtube and DBLP1, FIMM$_\alpha$ produces a comparatively higher EoF_α at a low PoF. On the contrary, FIMM$_0$ performs badly in terms of EoF_α, the value of which is even negative when k is small. Meanwhile, the PoF of FIMM$_0$ is also relatively higher, especially when $k = 5$ for Youtube and $k = 10$ for DBLP1. It is caused by the mandatory requirement of FIMM$_0$ to cover all the communities with the first seed. However, as k increases, FIMM$_0$ keeps remedying the previously fair solutions, leading to a similar PoF and EoF_α compared with FIMM$_\alpha$. For EoF_0, FIMM$_0$ is generally better than FIMM$_\alpha$. This accords with the aim of FIMM$_0$ to maximize $F_0(S) = \sum_{c \in \mathcal{C}} n_c \ln(\boldsymbol{u}_c(S))$. It seems counter-intuitive that FIMM$_\alpha$ gets a better EoF_0 when k is small. The reason is that though the first several seeds selected by FIMM$_\alpha$ may not be able to influence all the communities, the subsequent seeds gradually fill the gap. Consequently, FIMM$_\alpha$ could output a higher EoF_0 since it has a generally higher utility distribution.

Since DBLP2 includes only two groups divided by gender, its results are similar to those on the UVM and UCSC. As shown in Fig. 9, the PoF of FIMM$_\alpha$ and FIMM$_0$ remains below 1.5%, while both EoF_α and EoF_0 are always less than 0.2%. Interestingly, FIMM$_\alpha$ consistently achieves better EoF_α, while



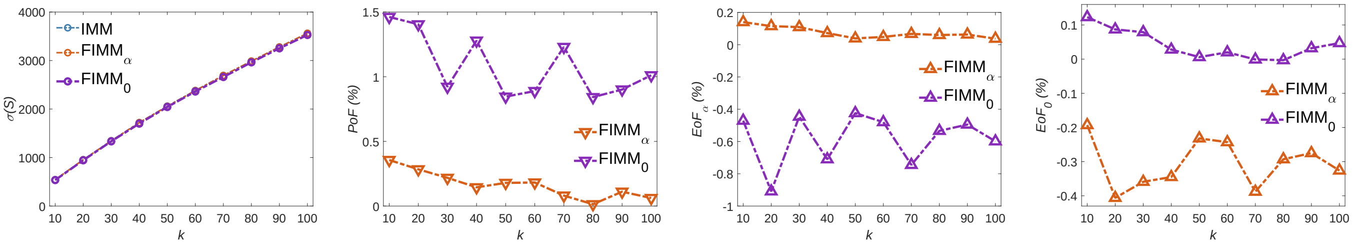
(a) $\sigma(S)$ (b) PoF (c) EoF_α (d) EoF_0

Fig. 7. Results on Youtube network (testing seed budget k).



(a) $\sigma(S)$ (b) PoF (c) EoF_α (d) EoF_0

Fig. 8. Results on DBLP1 network (testing seed budget k).



(a) $\sigma(S)$ (b) PoF (c) EoF_α (d) EoF_0

Fig. 9. Results on DBLP2 network (testing seed budget k).

FIMM$_0$ always performs better in terms of EoF_0. This observation confirms that FIMM$_\alpha$ and FIMM$_0$ strive towards similar yet fundamentally distinct objectives. Notably, the results on DBLP2 exhibit a different pattern compared to those on UVM and UCSC. This may be attributed to the fact that DBLP2 is significantly larger and has a much lower average degree. As a result, nodes from different communities are less likely to be connected, leading to a more distinct utility distribution.

5.3.6 Running Time

The number of RR sets is mainly determined by both the size of the network and the structure of communities. In the following, we exhibit the runtime of our algorithm with the scale of networks. Note that our algorithm is currently implemented in Matlab 2023a, thus it costs more time to generate RR sets (generating RR sets in C++ could be at least 100 times faster). In Table 3, **RRsets** refers to the time (seconds) used to generate RR sets, IMM, FIMM$_\alpha$ and FIMM$_0$ denote their corresponding time used to select seeds based on the generated RR sets, respectively.

TABLE 3
Running time (seconds).

Network	n_G	C	RRsets	IMM	FIMM$_\alpha$	FIMM$_0$
Email	1005	42	2.559	0.009	0.033	0.065
UVM	7322	2	61.284	0.040	0.360	0.457
UCSC	8990	2	43.703	0.049	0.462	0.675
Amazon	9239	229	31.415	0.006	0.146	0.438
Youtube	20707	379	13.252	0.008	0.544	1.570
Flixster	29357	10	49.259	0.036	2.659	7.519
DBLP1	59028	193	192.194	0.027	9.827	13.046
DBLP2	280200	2	650.581	0.114	15.821	19.958

The runtime on UVM and UCSC networks is averaged across two different community structures. The Flixster network is tested with the structure of 10 communities.

Generally, the runtime of FIMM$_\alpha$ and FIMM$_0$ scales with the size of networks, and FIMM$_0$ usually costs more time than FIMM$_\alpha$. When $\boldsymbol{u}$ is small, $\ln(\boldsymbol{u})$ experiences a much large change compared with $\boldsymbol{u}^\alpha$ as $\boldsymbol{u}$ varies. Therefore, FIMM$_0$ needs to update the marginal gain of more nodes even under the lazy-update strategy. The results demonstrate the strong scalability of our proposed algorithms, as evidenced by their efficient performance on a network with 280,200 nodes.

5.3.7 Difference between FIMM$_\alpha$ and FIMM$_0$

We summarize the above results of FIMM$_\alpha$ and FIMM$_0$ (except for UVM, UCSC, and DBLP2 where the results are all close to 0) and show their average value in Table 4.

TABLE 4
General results.

Method	PoF	EoF_α	EoF_0
FIMM$_\alpha$	3.68%	3.46%	4.31%
FIMM$_0$	13.25%	0.57%	5.76%

Overall, FIMM$_\alpha$ and FIMM$_0$ perform well in their respective scenarios. FIMM$_\alpha$ typically incurs a lower cost in influence spread while achieving a strong level of fairness, even across both objectives. In comparison, FIMM$_0$ has to pay a higher cost in influence spread to achieve notable fairness, but only within its specific objective.

Therefore, when seeking welfare fairness, FIMM$_\alpha$ is a more general choice, as it delivers better fairness at a lower cost. However, in scenarios where covering all communities is essential, FIMM$_0$ becomes the preferred option.

6 Conclusion

This paper focuses on the fair influence maximization problem with efficient approximation algorithms. Particularly, We study the problem under the notion of welfare fairness, the objective of which is the weighted sum of the fractional power of the expected proportion of activated nodes within every community. Existing methods that optimize the welfare objective lack efficiency, restricting their application to hundred-scale networks. In this paper, we first tackle the challenge of carrying out the unbiased estimation of the fractional power of the expected proportion of activated nodes in each community. Then, we deal with the challenge of integrating unbiased estimation into the Reverse Influence Sampling (RIS) framework. By meeting these two challenges, we propose an $(1-1/e-\varepsilon)$ approximation algorithm FIMM to maximize the fair influence. We further give a theoretical analysis that addresses the concentration of the unbiased estimator of the fractional power. The experiments validate that our algorithm is both scalable and effective, which is consistent with our theoretical analysis.

There are several future directions from this research. One direction is to find some other unbiased estimators for the fair influence that would be easier to calculate through RIS. Another direction is to develop a more efficient seed selection strategy that can handle cases with a large number of communities. Finally, exploring the fairness bound is a meaningful research direction that could enhance our understanding of fairness.

Acknowledgments

Xiaobin Rui was supported by the NSFC through grant 62402496, Zhixiao Wang was supported by the Natural Science Foundation of Jiangsu Province through grant BK20242084, and Hao Peng was supported by the NSFC through grant 62322202, and the Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China through grant 246Z0102G.

References

[1] B. Peng and W. Chen, "Adaptive influence maximization with myopic feedback," in *Advances in Neural Information Processing Systems 32, NeurIPS*, 2019, pp. 5575–5584.

[2] L. Sun, W. Huang, P. S. Yu, and W. Chen, "Multi-round influence maximization," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 2249–2258.

[3] K. Ali, C. Wang, and Y. Chen, "Leveraging transfer learning in reinforcement learning to tackle competitive influence maximization," *Knowledge and Information Systems*, vol. 64, no. 8, pp. 2059–2090, 2022.

[4] P. Perrault, J. Healey, Z. Wen, and M. Valko, "Budgeted online influence maximization," in *Proceedings of the 37th International Conference on Machine Learning, ICML*, vol. 119, 2020, pp. 7620–7631.

[5] C. Feng, L. Fu, B. Jiang, H. Zhang, X. Wang, F. Tang, and G. Chen, "Neighborhood matters: Influence maximization in social networks with limited access," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2844–2859, 2022.

[6] J. Ali, M. Babaei, A. Chakraborty, B. Mirzasoleiman, K. P. Gummadi, and A. Singla, "On the fairness of time-critical influence maximization in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2875–2886, 2023.

[7] Y. Li, J. Fan, Y. Wang, and K. Tan, "Influence maximization on social graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852–1872, 2018.

[8] B. Wilder, L. Onasch-Vera, J. Hudson, J. Luna, N. Wilson, R. Petering, D. Woo, M. Tambe, and E. Rice, "End-to-end influence maximization in the field," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, 2018, pp. 1414–1422.

[9] A. Rahmattalabi, S. Jabbari, H. Lakkaraju, P. Vayanos, M. Izenberg, R. Brown, E. Rice, and M. Tambe, "Fair influence maximization: a welfare optimization approach," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021, pp. 11 630–11 638.

[10] A. Tsang, B. Wilder, E. Rice, M. Tambe, and Y. Zick, "Group-fairness in influence maximization," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI*, 2019, pp. 5997–6005.

[11] R. Becker, G. D'Angelo, S. Ghobadi, and H. Gilbert, "Fairness in influence maximization through randomization," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021, pp. 14 684–14 692.

[12] Y. Feng, A. Patel, B. Cautis, and H. Vahabi, "Influence maximization with fairness at scale," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4046–4055.

[13] G. Farnadi, B. Babaki, and M. Gendreau, "A unifying framework for fairness-aware influence maximization," in *Companion of The 2020 Web Conference 2020*, 2020, pp. 714–722.

[14] H. Moulin, *Fair division and collective welfare*. MIT Press, 2003.

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

[16] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2014, pp. 946–957.

[17] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: near-optimal time complexity meets practical efficiency," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014, pp. 75–86.

[18] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1539–1554.

[19] M. Lin, L. Sun, R. Yang, X. Liu, Y. Wang, D. Li, W. Li, and S. Lu, "Fair influence maximization in large-scale social networks based on attribute-aware reverse influence sampling," *Journal of Artificial Intelligence Research*, vol. 76, pp. 925–957, 2023.

[20] X. Rui, Z. Wang, J. Zhao, L. Sun, and W. Chen, "Scalable fair influence maximization," in *Advances in Neural Information Processing Systems 36, NeurIPS*, 2023.

[21] P. M. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 57–66.

[22] M. Richardson and P. M. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 61–70.

[23] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137–146.

[24] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 420–429.

[25] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011 (Companion Volume)*. ACM, 2011, pp. 47–48.

[26] C. Zhou, P. Zhang, W. Zang, and L. Guo, "On the upper bounds of spread for greedy algorithms in social network influence maximization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2770–2783, 2015.

[27] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 199–208.

[28] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *10th IEEE International Conference on Data Mining, ICDM*, 2010, pp. 88–97.

[29] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 545–576, 2012.

[30] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng, "Staticgreedy: solving the scalability-accuracy dilemma in influence maximization," in *22nd ACM International Conference on Information and Knowledge Management, CIKM*, 2013, pp. 509–518.

[31] N. Ohsaka, T. Akiba, Y. Yoshida, and K. Kawarabayashi, "Fast and accurate influence maximization on large networks with pruned monte-carlo simulations," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 138–144.

[32] J. Tang, X. Tang, X. Xiao, and J. Yuan, "Online processing algorithms for influence maximization," in *Proceedings of the 2018 SIGMOD International Conference on Management of Data*, 2018, pp. 991–1005.

[33] H. Zhang, L. Fu, J. Ding, F. Tang, Y. Xiao, X. Wang, G. Chen, and C. Zhou, "Maximizing the spread of effective information in social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4062–4076, 2023.

[34] A. Stoica and A. Chaintreau, "Fairness in social influence maximization," in *Companion of The 2019 World Wide Web Conference, WWW 2019*, 2019, pp. 569–574.

[35] A. Rahmattalabi, P. Vayanos, A. Fulginiti, E. Rice, B. Wilder, A. Yadav, and M. Tambe, "Exploring algorithmic fairness in robust graph covering problems," in *Advances in Neural Information Processing Systems 32, NeurIPS*, 2019, pp. 15 750–15 761.

[36] A. Stoica, J. X. Han, and A. Chaintreau, "Seeding network influence in biased networks and the benefits of diversity," in *WWW '20: The Web Conference 2020*, 2020, pp. 2089–2098.

[37] B. Fish, A. Bashardoust, danah boyd, S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "Gaps in information access in social networks?" in *The World Wide Web Conference, WWW 2019*, 2019, pp. 480–490.

[38] G. J. Glasser, "An unbiased estimator for powers of the arithmetic mean," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 23, no. 1, pp. 154–159, 1961.

[39] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 555–564.

[40] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012.

[41] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," in *12th IEEE International Conference on Data Mining, ICDM*, 2012, pp. 81–90.

[42] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *12th IEEE International Conference on Data Mining, ICDM*, 2012, pp. 745–754.

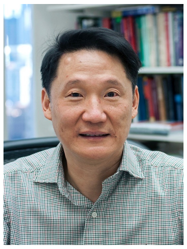
[43] F. Karimi, M. Génois, C. Wagner, P. Singer, and M. Strohmaier, "Homophily influences ranking of minorities in social networks," *Scientific reports*, vol. 8, no. 1, p. 11077, 2018.

Xiaobin Rui received his Ph.D. degree from China University of Mining and Technology. Currently, he serves as a Lecturer at the School of Computer Science and Technology at China University of Mining and Technology. He has published more than 20 papers in international conferences and journals. His research interests include social network analysis, social computing, and influence maximization.

Zhixiao Wang received his Ph.D. degree from Tongji University. Currently, he serves as a Professor at the School of Computer Science and Technology at China University of Mining and Technology. He has published more than 40 papers in international conferences and journals. His research interests include complex networks, social network analysis, and graph mining.

Hao Peng is currently a Professor at the School of Cyber Science and Technology in Beihang University. His current research interests include machine learning, deep learning, and reinforcement learning. He has published more than 60 research papers in top-tier journals and conferences, including the IEEE TKDE, TC, TPDS, TITS, ACM TOIS, TKDD, TIST, TSAS and Web Conference. He is the Associate Editor of the International Journal of Machine Learning and Cybernetics (IJMLC).

Wei Chen (Fellow, IEEE) received the bachelor's and master's degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, and the Ph.D. degree from the Department of Computer Science, Cornell University, Ithaca, NY, USA. He is a Principal Researcher at Microsoft Research Asia, Beijing. He is also an Adjunct Professor with the Institute of Interdisciplinary Information Sciences, Tsinghua University, and an Adjunct Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His main research interests include social and information networks, online learning, algorithmic game theory, Internet economics, distributed computing, and fault tolerance. Dr. Chen is a member of the Technical Committees of Big Data and Theoretical Computer Science of the Chinese Computer Federation.



Philip S. Yu (Life Fellow, IEEE) received the B.S. Degree in E.E. from National Taiwan University, the M.S. and Ph.D. degrees in E.E. from Stanford University. He is a Distinguished Professor in Computer Science at the University of Illinois at Chicago and also holds the Wexler Chair in Information Technology. Dr. Yu is a Fellow of the ACM and the IEEE. He is the recipient of ACM SIGKDD 2016 Innovation Award for his influential research and scientific contributions on mining, fusion and anonymization of big data, the IEEE Computer Society's 2013 Technical Achievement Award for "pioneering and fundamentally innovative contributions to the scalable indexing, querying, searching, mining and anonymization of big data", and the Research Contributions Award from IEEE Intl. Conference on Data Mining (ICDM) in 2003 for his pioneering contributions to the field of data mining. He was the Editor-in-Chiefs of ACM Transactions on Knowledge Discovery from Data (2011-2017) and IEEE Transactions on Knowledge and Data Engineering (2001-2004).

Appendix

6.1 Proofs

Fact 2. *(Chernoff bound) Let $X_1, X_2, \dots, X_R$ be R independent random variables with X_i having range [0, 1], and there exists $\mu \in [0,1]$ making $\mathbb{E}[X_i] = \mu$ for any $i \in [R]$. Let $Y = \sum_{i=1}^{R} X_i$, for any $\gamma > 0$,*

$$\Pr\{Y - t\mu \geq \gamma \cdot t\mu\} \leq \exp(-\frac{\gamma^2}{2 + \frac{2}{3}\gamma} t\mu).$$

For any $0 < \gamma < 1$,

$$\Pr\{Y - t\mu \leq -\gamma \cdot t\mu\} \leq \exp(-\frac{\gamma^2}{2} t\mu).$$

Lemma 3. *Let $\delta_1 \in (0,1)$, $\varepsilon_1 \in (0,1)$, and $\theta_1 = \frac{12Q^2 \ln(C/\delta_1)}{\varepsilon_1^2 (1-b)}$ where Q is the approximation parameter, $b = max(\boldsymbol{u}_c(S^*)), \forall c \in \mathcal{C}$, and $S^* = \operatorname{argmax}_{S:|S| \leq k} F_\alpha(S)$ denotes the optimal solution for the FIM problem based on $\mathcal{R}$, then $\hat{F}_\alpha(S^*, \mathcal{R}) \geq (1 - \varepsilon_1) \cdot OPT_\alpha$ holds with at least $1 - \delta_1$ probability if $\theta \geq C\theta_1$.*

Proof. Let X_c^i be the random variable for each $R_i \in \mathcal{R}$ (R_i rooted in c), such that $X_c^i = 1$ if $S^* \cap R_c(i) \neq \varnothing$, and $X_c^i = 0$ otherwise. Let $\pi_c = \theta_c - \sum_{i \in [\theta_c]} X_c^i$, then

$$\begin{aligned}
&Pr\left\{\hat{F}_\alpha(S^*, \mathcal{R}) < (1 - \varepsilon_1) \cdot OPT_\alpha\right\} \\
&= Pr\left\{\sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \frac{(\theta_c - n)!}{\theta_c!} \sum \prod_{d=1}^{n} \overline{X_c^{i_d}}\right) < (1 - \varepsilon_1) \cdot F_\alpha(S^*)\right\} \\
&= Pr\left\{\sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \frac{(\theta_c - n)!}{\theta_c!} \sum \prod_{d=1}^{n} \overline{X_c^{i_d}}\right) < (1 - \varepsilon_1) \sum_{c \in \mathcal{C}} n_c \left(\boldsymbol{u}_c(S^*)\right)^\alpha\right\} \\
&= Pr\left\{\sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i}\right) < (1 - \varepsilon_1) \sum_{c \in \mathcal{C}} n_c \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)\right\} \\
&\leq 1 - \prod_{c \in \mathcal{C}} \left(1 - Pr\left\{1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} < (1 - \varepsilon_1) \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)\right\}\right).
\end{aligned} \tag{15}$$

For each community c in Eq.(15), let $\varepsilon_1' = \frac{1 - \alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \left(1 - \boldsymbol{u}_c(S^*)\right)^n}{\alpha \sum_{n=1}^{\theta_c} \eta(n,\alpha) \left(1 - \boldsymbol{u}_c(S^*)\right)^n} \cdot \varepsilon_1$, thus $\varepsilon_1' \geq \varepsilon_1$ when $\alpha \leq 1/2$, and

$$\begin{aligned}
&Pr\left\{1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} < (1 - \varepsilon_1) \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)\right\} \\
&= Pr\left\{1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} < 1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n - \varepsilon_1 \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)\right\} \\
&= Pr\left\{\alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} > \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n + \varepsilon_1 \left(1 - \alpha \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right)\right\} \\
&= Pr\left\{\sum_{n=1}^{\theta_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} > (1 + \varepsilon_1') \sum_{n=1}^{\theta_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right\} \\
&\leq Pr\left\{\sum_{n=1}^{\pi_c} \eta(n, \alpha) \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} > (1 + \varepsilon_1') \sum_{n=1}^{\pi_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right\} \qquad (16) \\
&\leq Pr\left\{\sum_{n=1}^{\pi_c} \eta(n, \alpha) (\frac{\pi_c}{\theta})^n > (1 + \varepsilon_1') \sum_{n=1}^{\pi_c} \eta(n, \alpha) \big(1 - \boldsymbol{u}_c(S^*)\big)^n\right\} \qquad (17)
\end{aligned}$$

Then,

$$\begin{aligned}
Eq.(\,17) &\leq 1 - Pr\left\{(\frac{\pi_c}{\theta})^{\pi_c} < (1 + \varepsilon_1')\big(1 - \boldsymbol{u}_c(S^*)\big)^{\pi_c}\right\} \\
&= Pr\left\{(\frac{\pi_c}{\theta})^{\pi_c} \geq (1 + \varepsilon_1')\big(1 - \boldsymbol{u}_c(S^*)\big)^{\pi_c}\right\} \\
&\left(\text{Let } 1 + \varepsilon_0 = \sqrt[\pi_c]{1 + \varepsilon_1'}\right) \\
&= Pr\left\{\frac{\pi_c}{\theta} \geq (1 + \varepsilon_0)\big(1 - \boldsymbol{u}_c(S^*)\big)\right\} \\
&= \Pr\left\{\pi_c - \theta_c\big(1 - \boldsymbol{u}_c(S^*)\big) \geq \varepsilon_0 \theta_c\big(1 - \boldsymbol{u}_c(S^*)\big)\right\} \\
&\leq exp\left(-\frac{\varepsilon_0^2}{3}\theta_c\big(1 - \boldsymbol{u}_c(S^*)\big)\right)
\end{aligned} \tag{18}$$

Since $0 \leq \frac{\epsilon}{2x} \leq \sqrt[x]{1 + \epsilon} - 1 \leq \frac{\epsilon}{x}$ for $0 \leq \epsilon \leq 1$ and $x \geq 1$, it holds $\frac{2x}{\epsilon} \geq \frac{1}{\sqrt[x]{1+\epsilon} - 1}$. Let $\theta_c \geq \frac{12\pi_c^2 \ln(C/\delta_1)}{\varepsilon_1^2 \left(1 - \boldsymbol{u}_c(S^*)\right)} \geq \frac{3\ln(C/\delta_1)}{\left(\sqrt[\pi_c]{1 + \varepsilon_1'} - 1\right)^2 \left(1 - \boldsymbol{u}_c(S^*)\right)} = \frac{3\ln(C/\delta_1)}{\varepsilon_0^2 \left(1 - \boldsymbol{u}_c(S^*)\right)}$, then

$$\begin{aligned}
Eq.(18) &= exp\left(-\frac{\varepsilon_0^2}{3}\theta_c\big(1 - \boldsymbol{u}_c(S^*)\big)\right) \\
&\leq exp\left(-\frac{\varepsilon_0^2}{3} \frac{3\ln(C/\delta_1)}{\varepsilon_0^2\big(1 - \boldsymbol{u}_c(S^*)\big)} \big(1 - \boldsymbol{u}_c(S^*)\big)\right) \\
&= \delta_1/C
\end{aligned} \tag{19}$$

Therefore,

$$Eq.(15) \leq 1 - \prod_{c \in \mathcal{C}} (1 - \delta_1/C) \leq \delta_1 \tag{20}$$

To limit Eq.(16) to the first Q ($Q \geq 2$) terms, it becomes

$$Pr\left\{\sum_{n=1}^{Q}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} > (1+\varepsilon_1')\sum_{n=1}^{\pi_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}$$

$$\leq Pr\left\{\sum_{n=1}^{Q}\eta(n,\alpha)(\frac{\pi_c}{\theta})^n > (1+\varepsilon_1')\sum_{n=1}^{\pi_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}$$

$$\leq 1 - Pr\left\{(\frac{\pi_c}{\theta})^Q < (1+\varepsilon_1')\big(1-\boldsymbol{u}_c(S^*)\big)^Q\right\}$$

$$\leq Pr\left\{(\frac{\pi_c}{\theta})^Q \geq (1+\varepsilon_1')\big(1-\boldsymbol{u}_c(S^*)\big)^Q\right\}$$

$$\left(\text{Let } 1+\varepsilon_0 = \sqrt[Q]{1+\varepsilon_1'}\right)$$

$$= \Pr\left\{\pi_c - \theta_c\big(1-\boldsymbol{u}_c(S^*)\big) \geq \varepsilon_0\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right\}$$

$$\leq exp\left(-\frac{\varepsilon_0^2}{3}\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right) \tag{21}$$

$$\left(\text{Let } \theta_c \geq \frac{12Q^2\ln(C/\delta_1)}{\varepsilon_1^2\big(1-\boldsymbol{u}_c(S^*)\big)} \geq \frac{3\ln(C/\delta_1)}{\varepsilon_0^2\big(1-\boldsymbol{u}_c(S^*)\big)}\right)$$

$$\leq \delta_1/C \tag{22}$$

Therefore,

$$Eq.(15) \leq \delta_1 \tag{23}$$

It indicates $Pr\left\{\hat{F}_\alpha(S^*,\mathcal{R}) \geq (1-\varepsilon_1)\cdot OPT_\alpha\right\} \geq 1-\delta_1$, thus concludes the proof. □

Lemma 4. *Let $\delta_2 \in (0,1)$, $\varepsilon_2 = (\frac{e}{e+1})\varepsilon - \varepsilon_1$, and $\theta_2 = \frac{8Q^2\ln(C\binom{n_G}{k}/\delta_2)}{\varepsilon_2^2(1-b_0)}$ where Q is the approximation parameter, $b_0 = max(\boldsymbol{u}_c(S^{\#})), \forall c \in \mathcal{C}$ where $S^{\#}$ could be an arbitrary fair solution. For each bad S (which indicates $F_\alpha(S) < (1-1/e-\varepsilon)\cdot OPT_\alpha$), $\hat{F}_\alpha(S,\mathcal{R}) \geq (1-1/e)(1-\varepsilon_1)\cdot OPT_\alpha$ holds with at most $\delta_2/\binom{n_G}{k}$ probability if $\theta \geq C\theta_2$.*

Proof. Let X_c^i be the random variable for each $R_i \in \mathcal{R}$ (R_i rooted in c), such that $X_c^i = 1$ if $S \cap R_c(i) \neq \varnothing$, and $X_c^i = 0$ otherwise. Let $\pi_c = \theta_c - \sum_{i\in[\theta_c]} X_c^i$, then

$$Pr\left\{\hat{F}_\alpha(S,\mathcal{R}) \geq (1-\frac{1}{e})(1-\varepsilon_1)\cdot OPT_\alpha\right\}$$

$$\leq Pr\left\{\hat{F}_\alpha(S,\mathcal{R}) \geq (1+\varepsilon_2)F_\alpha(S)\right\}$$

$$= Pr\left\{\sum_{c\in\mathcal{C}} n_c\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i}\right) \geq (1+\varepsilon_2)\sum_{c\in\mathcal{C}} n_c\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right)\right\}$$

$$\leq 1-\prod_{c\in\mathcal{C}}\left(1-Pr\left\{1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \geq (1+\varepsilon_2)\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right)\right\}\right) \tag{24}$$

For each community c in Eq.(24), let $\varepsilon_2' = \frac{1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n}{\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n}\cdot\varepsilon_2$, thus $\varepsilon_2' \geq \varepsilon_2$ when $\alpha \leq 1/2$, and

$$Pr\left\{1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \geq (1+\varepsilon_2)\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right)\right\}$$

$$= Pr\left\{1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \geq 1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n + \varepsilon_2\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right)\right\}$$

$$= Pr\left\{\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq \alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n - \varepsilon_2\left(1-\alpha\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right)\right\}$$

$$= Pr\left\{\sum_{n=1}^{\theta_c}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq (1-\varepsilon_2')\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\} \tag{25}$$

To limit Eq.(25) to the first Q ($Q \geq 2$) terms, let $y = \frac{(1-\boldsymbol{u}_c(S))^{Q+1}}{(Q+1)\boldsymbol{u}_c(S)}$, $x = \frac{y\theta_c^2}{\theta_c-\pi_c+y\theta_c}$, Eq.(25) becomes

$$Pr\left\{\sum_{n=1}^{Q}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq (1-\varepsilon_2')\sum_{n=1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\}$$

$$= Pr\left\{\frac{\pi_c}{\theta_c} - (1-\varepsilon_2')\sum_{n=Q+1}^{\theta_c}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n + \sum_{n=2}^{Q}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq (1-\varepsilon_2')\sum_{n=1}^{Q}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\}$$

$$\leq Pr\left\{\frac{\pi_c}{\theta_c} - y + \sum_{n=2}^{Q}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq (1-\varepsilon_2')\sum_{n=1}^{Q}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\}$$

$$\leq Pr\left\{\frac{\pi_c-x}{\theta_c-x} + \sum_{n=2}^{Q}\eta(n,\alpha)\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} \leq (1-\varepsilon_2')\sum_{n=1}^{Q}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\}$$

$$\leq Pr\left\{\sum_{n=1}^{Q}\eta(n,\alpha)(\frac{\pi_c-Q+1}{\theta_c-Q+1})^n \leq (1-\varepsilon_2')\sum_{n=1}^{Q}\eta(n,\alpha)\big(1-\boldsymbol{u}_c(S)\big)^n\right\} \text{(when } x \leq Q-1)$$

$$\begin{aligned}
&\leq 1 - Pr\left\{(\frac{\pi_c - Q + 1}{\theta_c - Q + 1})^Q > (1-\varepsilon_2')\big(1-\boldsymbol{u}_c(S)\big)^Q\right\}\\
&= Pr\left\{(\frac{\pi_c - Q + 1}{\theta_c - Q + 1})^Q \leq (1-\varepsilon_2')\big(1-\boldsymbol{u}_c(S)\big)^Q\right\}\\
&\left(\text{Let } 1-\varepsilon_0 = \sqrt[Q]{1-\varepsilon_2'}\right)\\
&= Pr\left\{\frac{\pi_c - Q + 1}{\theta_c - Q + 1} \leq (1-\varepsilon_0)\big(1-\boldsymbol{u}_c(S)\big)\right\} \qquad (26)
\end{aligned}$$

;

$$\begin{aligned}
&Pr\left\{\hat{F}_0(S^*,\mathcal{R}) < (1+\varepsilon_1)\cdot OPT_0\right\}\\
&= Pr\left\{-\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n}\frac{(\theta_c-n)!}{\theta_c!}\sum\prod_{d=1}^{n}\overline{X_c^{i_d}} < (1+\varepsilon_1)\sum_{c\in\mathcal{C}} n_c \ln\left(\boldsymbol{u}_c(S^*)\right)\right\}\\
&= Pr\left\{-\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c}\frac{1}{n}\prod_{i=0}^{n-1}\frac{\pi_c - i}{\theta_c - i} < -(1+\varepsilon_1)\sum_{c\in\mathcal{C}} n_c\sum_{n=1}^{\theta_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\\
&\leq 1 - \prod_{c\in\mathcal{C}}\left(1 - Pr\left\{\sum_{n=1}^{\theta_c}\frac{1}{n}\prod_{i=0}^{n-1}\frac{\pi_c - i}{\theta_c - i} > (1+\varepsilon_1)\sum_{n=1}^{\theta_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\right). \qquad (29)
\end{aligned}$$

For each community c in Eq.(29), it has

$$\begin{aligned}
&Pr\left\{\sum_{n=1}^{\theta_c}\frac{1}{n}\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} > (1+\varepsilon_1)\sum_{n=1}^{\theta_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\\
&\leq Pr\left\{\sum_{n=1}^{\pi_c}\frac{1}{n}\prod_{i=0}^{n-1}\frac{\pi_c-i}{\theta_c-i} > (1+\varepsilon_1)\sum_{n=1}^{\pi_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\} \qquad (30)\\
&\leq Pr\left\{\sum_{n=1}^{\pi_c}\frac{1}{n}(\frac{\pi_c}{\theta_c})^n > (1+\varepsilon_1)\sum_{n=1}^{\pi_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\\
&\leq Pr\left\{(\frac{\pi_c}{\theta})^{\pi_c} \geq (1+\varepsilon_1)\big(1-\boldsymbol{u}_c(S^*)\big)^{\pi_c}\right\}\\
&\big(\text{Let } 1+\varepsilon_0 = \sqrt[\pi_c]{1+\varepsilon_1}\big)\\
&= Pr\left\{\frac{\pi_c}{\theta} \geq (1+\varepsilon_0)\big(1-\boldsymbol{u}_c(S^*)\big)\right\}\\
&= \Pr\left\{\pi_c - \theta_c\big(1-\boldsymbol{u}_c(S^*)\big) \geq \varepsilon_0\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right\}\\
&\leq exp\left(-\frac{\varepsilon_0^2}{3}\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right) \qquad (31)
\end{aligned}$$

Following the previous proof of Lemma 3, we have

$$Eq.(31) \leq exp\left(-\frac{\varepsilon_0^2}{3}\frac{3\ln(C/\delta_1)}{\varepsilon_0^2\big(1-\boldsymbol{u}_c(S^*)\big)}\big(1-\boldsymbol{u}_c(S^*)\big)\right) = \delta_1/C$$

Therefore,

$$Eq.(29) \leq 1 - \prod_{c\in\mathcal{C}}(1-\delta_1/C) \leq \delta_1 \qquad (32)$$

To limit Eq.(30) to the first Q ($Q \geq 2$) terms, Eq.(30) becomes

$$\begin{aligned}
&Pr\left\{\sum_{n=1}^{Q}\frac{1}{n}\prod_{i=0}^{n-1}\frac{\pi_c - i}{\theta_c - i} > (1+\varepsilon_1)\sum_{n=1}^{\theta_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\\
&\leq Pr\left\{\sum_{n=1}^{Q}\frac{1}{n}(\frac{\pi_c}{\theta_c})^n > (1+\varepsilon_1)\sum_{n=1}^{\theta_c}\frac{1}{n}\big(1-\boldsymbol{u}_c(S^*)\big)^n\right\}\\
&\leq Pr\left\{\frac{\pi_c}{\theta} \geq (1+\varepsilon_1)\big(1-\boldsymbol{u}_c(S^*)\big)\right\}\\
&= Pr\left\{(\frac{\pi_c}{\theta})^Q \geq (1+\varepsilon_1)\big(1-\boldsymbol{u}_c(S^*)\big)^Q\right\}\\
&\big(\text{Let } 1+\varepsilon_0 = \sqrt[Q]{1+\varepsilon_1}\big)\\
&= \Pr\left\{\pi_c - \theta_c\big(1-\boldsymbol{u}_c(S^*)\big) \geq \varepsilon_0\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right\}\\
&\leq exp\left(-\frac{\varepsilon_0^2}{3}\theta_c\big(1-\boldsymbol{u}_c(S^*)\big)\right) \qquad (33)\\
&\left(\text{Let } \theta_c \geq \frac{12Q^2\ln(C/\delta_1)}{\varepsilon_1^2\big(1-\boldsymbol{u}_c(S^*)\big)} \geq \frac{3\ln(C/\delta_1)}{\varepsilon_0^2\big(1-\boldsymbol{u}_c(S^*)\big)}\right)\\
&\leq \delta_1/C \qquad (34)
\end{aligned}$$

Therefore,

$$Eq.(29) \leq \delta_1 \qquad (35)$$

It indicates $Pr\left\{\hat{F}_\alpha(S^*,\mathcal{R}) \geq (1+\varepsilon_1)\cdot OPT_\alpha\right\} \geq 1-\delta_1$, thus concludes the proof. □

Lemma 6. *Let $\delta_2 \in (0,1)$, $\varepsilon_2 = (\frac{e}{e-1})\varepsilon - \varepsilon_1$, and $\theta_2 = \frac{8Q^2\ln(C\binom{n_G}{k}/\delta_2)}{\varepsilon_2^2(1-b_0)}$ where Q is the approximation parameter, $b_0 = max(\boldsymbol{u}_c(S^{\#})), \forall c \in \mathcal{C}$ where $S^{\#}$ could be an arbitrary fair solution. For each bad S (which indicates $F_0(S) < (1+1/e+\varepsilon)\cdot OPT_0$), $\hat{F}_0(S,\mathcal{R}) \geq (1+1/e)(1+\varepsilon_1)\cdot OPT_0$ holds with at most $\delta_2/\binom{n_G}{k}$ probability if $\theta \geq C\theta_2$.*

Proof. Let X_c^i be the random variable for each $R_i \in \mathcal{R}$ (R_i rooted in c), such that $X_c^i = 1$ if $S \cap R_c(i) \neq \varnothing$, and $X_c^i = 0$ otherwise. Let $\pi_c = \theta_c - \sum_{i\in[\theta_c]} X_c^i$, then

$$\begin{aligned}
&Pr\left\{\hat{F}_0(S,\mathcal{R}) \geq (1+\frac{1}{e})(1+\varepsilon_1)\cdot OPT_0\right\}\\
&\leq Pr\left\{\hat{F}_0(S,\mathcal{R}) \geq (1-\varepsilon_2)F_0(S)\right\}
\end{aligned} \qquad (36)$$

$$
\begin{aligned}
&= Pr\left\{-\sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} \geq -(1-\varepsilon_2) \sum_{c\in\mathcal{C}} n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \big(1 - \boldsymbol{u}_c(S)\big)^n \right\} \\
&\leq 1 - \prod_{c\in\mathcal{C}} \left(1 - Pr\left\{n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} \leq (1-\varepsilon_2) n_c \sum_{n=1}^{\theta_c} \frac{1}{n} \big(1 - \boldsymbol{u}_c(S)\big)^n \right\}\right) \qquad (37)
\end{aligned}
$$

For each community c in Eq.(37), it equals to

$$
Pr\left\{\sum_{n=1}^{\theta_c} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} \leq (1-\varepsilon_2) \sum_{n=1}^{\theta_c} \frac{1}{n} \big(1 - \boldsymbol{u}_c(S)\big)^n \right\} \qquad (38)
$$

Similar to the proof of Lemma 4, we limit Eq.(38) to the first Q ($Q \geq 2$) terms as

$$
\begin{aligned}
&Pr\left\{\sum_{n=1}^{Q} \frac{1}{n} \prod_{i=0}^{n-1} \frac{\pi_c - i}{\theta_c - i} \leq (1-\varepsilon_2) \sum_{n=1}^{\theta_c} \frac{1}{n} \big(1 - \boldsymbol{u}_c(S)\big)^n \right\} \\
&(\text{when } x \leq Q-1) \\
&\leq Pr\left\{\sum_{n=1}^{Q} \frac{1}{n} (\frac{\pi_c - Q + 1}{\theta_c - Q + 1})^n \leq (1-\varepsilon_2) \sum_{n=1}^{Q} \frac{1}{n} \big(1 - \boldsymbol{u}_c(S)\big)^n \right\} \\
&\leq Pr\left\{(\frac{\pi_c - Q + 1}{\theta_c - Q + 1})^Q \leq (1-\varepsilon_2) \big(1 - \boldsymbol{u}_c(S)\big)^Q \right\} \\
&\big(\text{Let } 1 - \varepsilon_0 = \sqrt[Q]{1-\varepsilon_2}\big) \\
&= Pr\left\{\frac{\pi_c - Q + 1}{\theta_c - Q + 1} \leq (1-\varepsilon_0) \big(1 - \boldsymbol{u}_c(S)\big) \right\} \qquad (39)
\end{aligned}
$$

Then, similar to the proof of Lemma 4, let $\varepsilon_0' = \varepsilon_0 + \frac{\theta_c}{\pi_c} \frac{\pi_c - Q + 1}{\theta_c - Q + 1} - 1$, $\varepsilon_0 \geq \varepsilon_0' \geq 1 - \frac{\theta_c}{\pi_c} \frac{\pi_c - Q}{\theta_c - Q}$, Eq.(39) becomes

$$
\begin{aligned}
&Pr\left\{\frac{\pi_c - Q + 1}{\theta_c - Q + 1} \leq (1-\varepsilon_0) \big(1 - \boldsymbol{u}_c(S)\big) \right\} \\
&= Pr\left\{\frac{\pi_c}{\theta_c} \leq (1 - \frac{\pi_c}{\theta_c} \frac{\theta_c - Q + 1}{\pi_c - Q + 1} \varepsilon_0') \big(1 - \boldsymbol{u}_c(S)\big) \right\} \\
&\leq exp\left(-\frac{\varepsilon_0'^2}{2} \theta_c \big(1 - \boldsymbol{u}_c(S)\big)\right) \\
&\leq \frac{1}{C} \cdot \delta_2 / \binom{n_G}{k} \qquad (40)
\end{aligned}
$$

Therefore,

$$
Eq.(37) \leq 1 - \prod_{c\in\mathcal{C}} (1 - \frac{1}{C} \cdot \delta_2 / \binom{n_G}{k}) \leq \delta_2 / \binom{n_G}{k} \qquad (41)
$$

It indicates $Pr\left\{\hat{F}_0(S, \mathcal{R}) \geq (1 + 1/e)(1 + \varepsilon_1) \cdot OPT_0\right\} \leq \delta_2 / \binom{n_G}{k}$, thus concludes the proof.

□