Journal of Artificial Intelligence Research 76 (2023) 925-957 Submitted 11/2022; published 04/2023

Fair Influence Maximization in Large-scale Social Networks Based on Attribute-aware Reverse Influence Sampling

Mingkai Lin MINGKAI@SMAIL.NJU.EDU.CN
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, China

Lintan Sun LINTANSUN@163.COM
Rui Yang RUIYANGWCMC@163.COM
Xusheng Liu XUSHENGLIU_SGCSC@163.COM
Yajuan Wang YAJUANWANGWCMC@SINA.COM
State Grid Customer Service Center,
State Grid Corporation of China, Tianjin, China

Ding Li LIDING@SMAIL.NJU.EDU.CN
Wenzhong Li LWZ@NJU.EDU.CN
Sanglu Lu SANGLU@NJU.EDU.CN
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, China

Abstract

Influence maximization is the problem of finding a set of seed nodes in the network that maximizes the influence spread, which has become an important topic in social network analysis. Conventional influence maximization algorithms cause "unfair" influence spread among different groups in the population, which could lead to severe bias in public opinion dissemination and viral marketing. To address this issue, we formulate the fair influence maximization problem concerning the trade-off between influence maximization and group fairness. For the purpose of solving the fair influence maximization problem in large-scale social networks efficiently, we propose a novel attribute-based reverse influence sampling (ABRIS) framework. This framework intends to estimate influence in specific groups with guarantee through an attribute-based hypergraph so that we can select seed nodes strategically. Therefore, under the ABRIS framework, we design two different node selection algorithms, ABRIS-G and ABRIS-T. ABRIS-G selects nodes in a greedy scheduling way. ABRIS-T adopts a two-phase node selection method. These algorithms run efficiently and achieve a good trade-off between influence maximization and group fairness. Extensive experiments on six real-world social networks show that our algorithms significantly outperform the state-of-the-art approaches.

1. Introduction

With the rapid development of Internet technology, online social networks have become the mainstream platform for interpersonal communication, which play an important role in spreading information, innovation, and influence among its users. Influence maximization (IM) in social networks is the problem of finding a set of seed nodes in the network that maximizes the spread of influence under a certain information prorogation model. Due to its great potential in viral marketing and public opinion analysis, the influence maximization problem has received great attention from academia and industry.

Kempe et al. (2003) formulated the influence maximization problem as a combinational optimization problem and proposed a greedy algorithm that yields a $(1-1/e)$-approximation under the Independent Cascade (IC) and Linear Thresholds (LT) diffusion models. Since then, many algorithms (Goyal, Lu, & Lakshmanan, 2011; Liu, Xiang, Chen, Xiong, Tang, & Yu, 2014; Chen, 2009; Chen, Wang, & Wang, 2010) have been proposed to improve the efficiency and scalability of the IM algorithm.

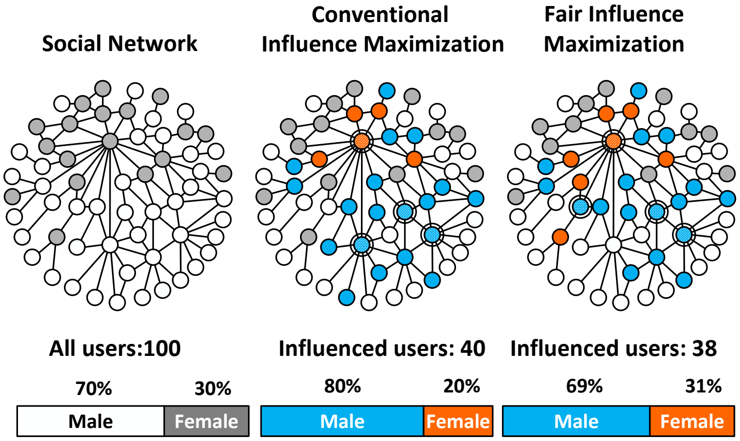


Figure 1: A toy example of fair influence maximization. The left figure shows the original social network, with two attributed groups: 70% males and 30% females. The middle figure shows that a conventional influence maximization approach activates 40 users, with 80% males and 20% females, resulting in more influence in the male group. The right figure shows that the fair influenced maximization approach activates 38 users with 69% males and 31% females, which is more consistent with the original male/female percentage.

However, the past research unilaterally sought the greatest overall influence in public networks without considering the attributes attached to the nodes, such as race, age, gender, and educational background. Generally, people in a social network or real life are meant to form different groups according to their attribute statuses. A group of nodes sharing the same attribute status in a social network can be called an attributed group. For instance, in a residential area, it may be the young or the old; in a school, it may be the faculties or the students; and in a society, it may be the sexual majorities or the sexual minorities. Many studies show that social groups exert huge impacts on the organization of social networks, ranging from user demographics to subjective preferences like political orientation and personal interests.

Conventional influence maximization approaches, while pursuing the overall influence without paying attention to alternative groups, will cause the problem of uneven influence spread among different groups. For a toy example, as illustrated in Figure 1, there are 100 users in the original network, where 70% are males and 30% are females. A conventional influence maximization algorithm activates/influences 40 users, where 80% are males and 20% are females. Clearly, the algorithm has more influence in the male group and less influence in the female group, causing an uneven spread of influence in the network. Uneven influence spread sometimes could lead to severe bias in the result, such as opinion polls for the election. An ideal influence maximization algorithm should approach the maximum influence spread while maintaining fairness among different groups. As suggested in Figure 1, it may sacrifice a few influenced nodes to trade for group fairness, e.g., activating 38 users, where 69% are males and 31% are females to keep it more consistent with the original gender percentage. For this purpose, this paper studies a novel problem referred to as the *fair influence maximization problem:* Given a network G with group information Ψ and a budget k, find k seed nodes to activate the other nodes in the network so that the numbers of influenced nodes among different groups are proportional to their original populations, and the total influence spread reaches the maximum. Solving the fair influence maximization problem is non-trivial and encounters several challenges.

The first challenge concerns the efficiency of the algorithm for large-scale social networks. Since most existing approaches rely on Monte Carlo simulation to estimate influence spread, repeatedly traversing the whole network for influence evaluation is highly inefficient. Making use of sampling techniques like Reverse Influence Sampling (RIS) can reduce complexity dramatically. But most sampling techniques emphasize the overall influence in the whole network without considering the influence in specific groups.

The second challenge refers to the fairness of influence spread among different groups. This requires an accurate estimation of influence in each individual attributed group and an efficient scheduling strategy or mechanism with a theoretical guarantee to select seed nodes for fair influence. The trade-off between influence maximization and group fairness is the key design of the solution.

To address the issues, this paper proposes an attribute-based reverse influence sampling (ABRIS) framework, including two novel node selection algorithms for the fair influence maximization problem in social networks. The ABRIS framework adopts the RIS technique to construct an attribute-based hypergraph, where influence spread in each individual group can be estimated with guaranteed accuracy. Under the ABRIS framework, we propose two different algorithms called ABRIS-G and ABRIS-T from different perspectives for seed nodes selection. To be specific, ABRIS-G is a basic greedy algorithm with a scheduling strategy that chooses seed nodes iteratively to achieve the objectives of influence maximization and influence fairness among groups. ABRIS-T is an extended algorithm with a two-phase node selection mechanism. The first phase seeks the solution to maximize the minimum fraction of influenced users within all influenced groups, where we provide a theoretical guarantee for the procedure in large-scale social networks. But it would probably select many low-influence seeds when some groups are not well-connected. Therefore in the second phase, we adopt a greedy approach to iteratively change the lowest influence node for the node with larger overall influence to chase better performance of fair influence maximization. We evaluate the performance based on six real-world datasets. Extensive experiments show that the proposed solutions perform very close to the conventional influence maximization algorithms in influence spread. They can also balance the influence among groups, which significantly outperforms the state-of-the-art approaches.

The proposed fair influence maximization algorithms have wide-range applications. They surely benefit the viral marketing and advertisement service by taking into account different finer-grained attributed groups. They are also acquired in public opinion dissemination, like candidates' political opinions and public policies for different groups of people, which is more important to achieve fair influence maximization instead of simply chasing the maximum influence spread. The contributions are summarized as follows.

- We identify the fair influence maximization problem in social networks and formulate the objective, simultaneously considering the influence maximization and the disparity among group influences. It is novel and fundamental for social network studies.
- We propose an attribute-based reverse influence sampling (ABRIS) framework. It estimates influence spread in specific groups with lower computation complexity and accuracy guarantee so that seed nodes can be selected strategically.
- Under the ABRIS framework, we design two node selection algorithms. They are ABRIS-G, based on a greedy scheduling strategy and ABRIS-T, through a two-phase node selection approach. These algorithms run efficiently and achieve a good trade-off between influence maximization and group fairness.
- We evaluate the performances of the proposed approaches based on six real-world social networks scaling from thousands of nodes to millions of nodes. Extensive experiments show that our solution significantly outperforms the state-of-the-art methods on fair influence maximization in large-scale social networks.

2. Related Work

In this section, we introduce the related works in detail through the basic influence maximization problem and the latest progress on fair influence maximization.

2.1 Influence Maximization Problem

Social networks play a fundamental role as a medium for the spread of information, innovation, and influence among its members. There had been a great number of models to describe the process of information/influence propagation in social networks (Shah & Zaman, 2010; Shah, 2011; Bailey et al., 1975; Easley & Kleinberg, 2010). In the last decades, the problem of identifying a set of influential users in social networks gradually received more and more attention by academia and industry.

Kempe et al. (2003) formulated the influence maximization (IM) problem as an optimization problem: given a budget k to find k seed nodes in the network, such that by activating them we can reach the maximum spread of influence in the network. They showed that the problem is NP-hard and proposed a greedy algorithm to get a $(1-1/e-\epsilon)$-approximate solution. Kempe et al.'s algorithm is simple and effective, but it suffers from high time complexity of $O(k(m+n)n)$ which is hard to apply in large-scale social networks. To address the inefficiency issue, Leskovec et al. (2007) improved the greedy method with the lazy-forward heuristic (CELF), which took advantage of the submodular property (Minoux, 1978) to reduce the number of evaluations on the influence spread of individuals. Goyal et al. further optimized the strategy and proposed CELF++ (Goyal et al., 2011) which reduced the times of Monte Carlo simulation for influence evaluation.

Some studies tried to solve the problem through a series of heuristic algorithms based on degree centrality or other structural properties (Kempe et al., 2003; Liu et al., 2014;

Chen, 2009; Chen et al., 2010; Jalili & Perc, 2017). Liu et al. (2014) developed a quantitative metric named Group-PageRank to quickly estimate the upper bound of the social influence. Chen et al. proposed an algorithm called PMIA (Chen et al., 2010), which used a tunable parameter to control the balance between the running time and the influence spread. However, all of these heuristic methods shared the same shortcoming of giving up the approximation guarantee to ensure the efficiency of algorithms.

Recently, Borgs et al. (2014) proposed a Reverse Influence Sampling (RIS) framework for influence maximization, which made a theoretical breakthrough on the problem by presenting a near-linear time algorithm under independent cascade (IC) model. Tang et al. proposed an enhanced algorithms called TIM that runs in $O((k+l)(n+m)\log n/\varepsilon^2)$ expected time and returns a $(1-1/e-\varepsilon)$-approximate solution with at least $1-n^{-l}$ probability (Tang, Xiao, & Shi, 2014). Later a few algorithms (Tang, Shi, & Xiao, 2015; Nguyen, Thai, & Dinh, 2016b) were proposed to improve the algorithm with better bounds.

2.2 Fair Influence Maximization

The research on influence spread in specific groups can trace back to the target influence maximization (TIM) problem, where the goal is to maximize the influence over a group of target users (Li, Zhang, & Tan, 2015; Li, Li, & Shan, 2011). However, this paper focuses on the fair influence maximization problem, which is a novel problem that tends to maximize the total social influence while keeping the influence fairness among users regarding different groups. In previous works, one closely related problem is the robust submodular observation selection (RSOS) problem (Krause, McMahan, Guestrin, & Gupta, 2008) which constrains influences in different monotone submodular functions. For the RSOS problem, Chekuri et al. (2010) proposed an optimal $(1-1/e)$-approximate algorithm with $O(n^8)$ time complexity. Recently, Udwani (2021) introduced a state-of-the-art three-stage method based on Multiplicative-Weight-Updates (MWU) with asymptotic $(1-1/e)^2-\lambda$-approximation and $\tilde{O}(n/\lambda^3)$ time with the assumption that the number of submodular functions is $o(k/\log^3 k)$.

The MaxMin problem of RSOS (Krause et al., 2008) maximizes the minimum fraction of users within each influenced group can be served as an effective objective to achieve group fairness in some studies (Tsang, Wilder, Rice, Tambe, & Zick, 2019; Becker, D'Angelo, Ghobadi, & Gilbert, 2022). However, this objective stems from the fact that the equality of outcomes may be undesirable and perform badly in overall influence, especially when some groups are much better connected than the others. Furthermore, in the past two years, many other researchers proposed alternative approaches to solve the fair influence maximization problem from different perspectives. Fish et al. (2019) proposed to use the social welfare function as an objective and gave empirical evidence that a simple greedy-based strategy worked well in practice. Some works (Ali, Babaei, Chakraborty, Mirzasoleiman, Gummadi, & Singla, 2022; Anwar, Saveski, & Roy, 2021; Lin, Li, & Lu, 2020) chose to define the influence disparity among groups and made it a penalty to the overall influence in the whole network. Later, a more recent fairness-aware IM framework was proposed by Farnadi et al. (2020), which was based on an integer programming formulation of the influence maximization problem. With the development of deep learning, the powerful tool could also be utilized to solve the fairness issue in the IM problem. Khajehnejad et al. (2021)

first introduced deep learning into the fair influence maximization problem. The authors proposed an adversarial network to obtain similarly distributed embeddings across sensitive attributes. The seed set was selected by clustering the embeddings. What's more, other variants of the fair influence problem (Ali et al., 2022; Anwar et al., 2021) were also studied. Recently Becker et al. (2022) studied the impact of randomization on fairness. The authors allowed randomized strategies for choosing the seed nodes rather than restricted to deterministic methods. And two probabilistic strategies based on node and set views were proposed to derive seeds.

Even though these works solved such problem to some extent, all of the proposed algorithms were only processed in small-size (less than 10k nodes) networks. While Gershtein et al. (2021) refined a Multi-Objective IM to solve the problem in large-scale networks, where all objectives except one are turned into constraints for groups, and the remaining objective is optimized subject to these constraints. Unlike all of these existing studies, in this work, we design a novel attribute-based reverse influence sampling (ABRIS) framework along with two node selection methods so that the fair influence maximization problem can be solved theoretically and efficiently in large-scale (more than 1000k nodes) social networks.

3. Problem Formulation

In this section, we propose the formulation of the fair influence problem and the solution framework. The notations used in the rest of the paper are summarized in table 1.

Table 1: Notations

Notation	Description
m, n	Nodes and edges number of graph $G = (V, E)$
ψ_i	The sign of group i.
Ψ	The attributed group set,$\psi_i \in \Psi$
$\mathbb{I}_{\psi_i}(S)$	Influence spread of seed set $S \subseteq V$ in ψ_i.
$\hat{\mathbb{I}}_{\psi_i}(S)$	The estimated influence of S in ψ_i through ABRIS.
OPT_i	The maximum $\mathbb{I}_{\psi_i}(S)$ for any size-k seed set S.
V_{ψ_i}	The set for vertices in group ψ_i.
n_i	The number of vertices in group ψ_i, $n_i = \lvert V_{\psi_i}\rvert$.
$D_\Psi(S)$	The disparity for influence among groups.
γ	The discount factor to penalty the disparity.
$\mathbb{F}(S)$	Objective function for fair influence maximization.
$\mathcal{R}(v)$	The Reverse Reachable (RR) set of v
$\mathcal{R}_{\psi_i}$	θ_i RR sets generated for group ψ_i
$F(S, \mathcal{R}_{\psi_i})$	The fraction of RR sets in $\mathcal{R}_{\psi_i}$ covered by S.

We model a social network as an undirected weighted graph $G(V, E, W)$ with $|V| = n$ nodes and $|E| = m$ edges. Each edge $(u, v) \in E$ is associated with a weight $W(u, v) \in [0, 1]$ representing the probability that u and v influence each other. Let $S \subseteq G$ be a set of seed nodes with size $|S| = k$.

Influence spread in a social network can be described as a diffusion process. As introduced by Kempe et al. (2003), there are two essential diffusion models called Linear Threshold (LT) and Independent Cascade (IC). In this paper, we particularly focus on the IC diffusion process to explain our algorithm. Note that the proposed approach can be extended to other influence diffusion models without difficulty.

The process of IC model runs round by round to activate/influence nodes in the social network as follows:

(1) In round 0, all nodes in S are activated, and all nodes in $V - S$ are not activated.

(2) In each subsequent round, the newly activated nodes will try to activate their neighbors. Each newly activated node u has a single chance to activate each inactive neighbor v with the probability proportional to the edge weight $W(u, v)$.

(3) Once a node becomes active, it will remain active in all subsequent rounds. The process ends when no more nodes get activated.

In the diffusion process, we refer to S as the *seed set*, and the size of $|S| = k$ as the *budget*. Let $\mathbb{I}(S)$ be the number of nodes that are activated when the diffusion process converges. We call $\mathbb{E}(\mathbb{I}(S))$ the *influence spread* of S under the IC diffusion model.

Further more, in a social network with attributed groups, each node is attached with one or more groups from a group set $\Psi = \{\psi_1, \psi_2, \cdots\}$, where Ψ can be overlapping or non-overlapping groups. For example, this is a group set $\Psi = \{male, female, student\}$. We further define the *influence spread in a group* ψ_i as the number of activated nodes in group ψ_i, which is denoted by $\mathbb{I}_{\psi_i}(S)(i = 1, 2, \cdots)$ accordingly.

Tsang et al. (2019) provided two representative fairness metrics as Maximin Fairness and Diversity Constraints Fairness from different perspectives and proposed solutions to them. Maximin Fairness maximizes the minimum influence received by any of the groups but may select many low-influence seeds. Diversity Constraints Fairness tries to maximize the overall influence on the premise of ensuring the least rational influence for each group. Also it may select influential nodes with severe unfair influence. Thereby we propose an objective function simultaneously considering both influence maximization and group fairness.

Let $V_{\psi_i}(i = 1, 2, \cdots)$ be the set of nodes with size n_i in the group ψ_i in network G. The percentage of influenced node for each group can be represented by $\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}(i = 1, 2, \cdots)$. To measure the fairness of influence spread in different group, we introduce the *disparity* metric $D_\Psi(S)$ (Lin et al., 2020; Anwar et al., 2021) as:

$$D_\Psi(S) = n \cdot (\max_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\} - \min_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\}), \tag{1}$$

where the right part measures the difference between the maximum and minimum influence percentage in attributed groups, and we multiply it by n to not only avoid the value being too small but also normalize the disparity by the size of the graph.

With the above definitions and notations, we formally describe the fair influence maximization problem and its objective in the following.

Definition 3.1 (Fair influence maximization problem)**.** *Given a graph G with attributed group set Ψ and a budget k, the influence maximization problem is to find a set S of at most k nodes maximizing*

$$\mathbb{F}(S) = \mathbb{E}(\mathbb{I}(S)) - \gamma D_\Psi(S) = \mathbb{E}(\mathbb{I}(S)) - n\gamma \max_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\} + n\gamma \min_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\} \tag{2}$$

where $\mathbb{E}(\mathbb{I}(S))$ is the total overall expected influence spread, $D_\Psi(S)$ is the disparity, and $\gamma \in [0, +\infty)$ is a discount factor.

According to the definition, the fair influence maximization problem aims to maximize the total expected influence spread and take disparity as a penalty term to achieve group fairness. With the value of γ getting larger, the emphasis on fair influence becomes larger. We generally adopt $\gamma = 1$ by default. Because at this time, the overall influence and disparity are both scaled by graph size n and can achieve good trade-off performances in practice. But γ can also be tuned according to practical needs. In particular, when $\gamma = 0$, it reduces to the conventional influence maximization problem. However, when $\gamma > 0$, the objective function $\mathbb{F}(S)$ is neither monotonic nor submodular, which is different from the conventional influence maximization problem. Therefore we have the following theorem.

Theorem 3.1. *When $\gamma > 0$, the objective function $\mathbb{F}(S) = \mathbb{E}(\mathbb{I}(S)) - \gamma D_\Psi(S)$ is neither monotonic nor submodular.*

Proof. We respectively prove the non-monotonicity and non-submodularity for $\mathbb{F}(S)$ when $\gamma > 0$ through offering counterexamples.

Non-monotonicity: When $\gamma \in (0, 1]$, we can construct two non-overlapping groups that one larger group has $2 \cdot \lfloor \frac{2}{\gamma} \rfloor$ nodes and the other smaller group has two nodes. Assume that a seed set S has an expected influence of $\lfloor \frac{2}{\gamma} \rfloor$ for the larger group and 1 for the smaller group, we can derive $\mathbb{F}(S) = \lfloor \frac{2}{\gamma} \rfloor + 1$. We now expand the seed set S with another node v which is the other uninfluenced node in the smaller group and it can only activate itself. In this way, we have $\mathbb{F}(S \cup \{v\}) = \lfloor \frac{2}{\gamma} \rfloor + 2 - \gamma(\lfloor \frac{2}{\gamma} \rfloor + 1) < \lfloor \frac{2}{\gamma} \rfloor + 2 - 1 = \lfloor \frac{2}{\gamma} \rfloor + 1 = \mathbb{F}(S)$. When $\gamma \in (1, +\infty)$, we can also assume two non-overlapping groups and each group has two nodes. The seed set S expectedly influences 1 node for each group and thus $\mathbb{F}(S \cup \{v\}) = 2$. If we expand the seed set S with the other uninfluenced node v in a certain group and v can only activate itself, we can derive $\mathbb{F}(S \cup \{v\}) = 3 - 2\gamma < 2 < \mathbb{F}(S)$. However, for the interval of $\gamma \in (0, +\infty)$ in the same construction of two non-overlapping groups with each group having one node, if S influences 1 node in a certain group and $(S \cup \{v\})$ influences both two nodes, we can derive $\mathbb{F}(S \cup \{v\}) = 2 > 1 - 2\gamma = \mathbb{F}(S)$ on the contrary. Hence, when $\gamma \in (0, +\infty)$, the function $\mathbb{F}(S)$ is non-monotonic.

Non-submodularity: Similarly, we construct two non-overlapping groups and each group has two nodes respectively, namely $G_1 = \{a, b\}, G_2 = \{c, d\}$. We have two seed sets as $S_1 = \{a, c\}$ and $S_2 = \{a, b, c\}$ together with a node $v = d$ to expand the seed sets. It's obvious that $S_1 \subset S_2$. Assuming that the nodes can only influence themselves respectively, we should get $\mathbb{F}(S_1 \cup \{v\}) - \mathbb{F}(S_1) = 3 - 2\gamma - 2 = 1 - 2\gamma$ and $\mathbb{F}(S_2 \cup \{v\}) - \mathbb{F}(S_2) = 4 - 3 + 2\gamma - 2 = 1 + 2\gamma$. There is no doubt that $\mathbb{F}(S_1 \cup \{v\}) - \mathbb{F}(S_1) < \mathbb{F}(S_2 \cup \{v\}) - \mathbb{F}(S_2)$ when $\gamma > 0$. In this way, for $\gamma > 0$, the function $\mathbb{F}(S)$ is non-submodular.

□

4. The ABRIS Framework

In this section, we propose the attribute-based reverse influence sampling (ABRIS) framework with two different node selection algorithms, ABRIS-G and ABRIS-T.

4.1 Preliminary

Reverse influence sampling (RIS) is an efficient algorithm framework introduced by Borgs et al. (2014) to find a set of seeds to maximize the influence spread with a guaranteed approximation ratio. To describe the RIS framework, we first introduce the following concepts.

Definition 4.1. *(Reverse Reachable (RR) Set (Tang et al., 2014)) Given a graph G and a node $v \in G$, let g be a sample graph from G obtained by removing each edge (v_x, v_y) in G with probability $1 - W(v_x, v_y)$. The Reverse Reachable (RR) set of v is the set of nodes that can reach v in g, which is denoted by $\mathcal{R}(v)$.*

Definition 4.2. *(RR Set Coverage) A node u is said to cover an RR set $\mathcal{R}(v)$ if and only if $u \in \mathcal{R}(v)$.*

By definition, if a node $u \in \mathcal{R}(v)$ is chosen as a seed, then it should have a chance to activate v along a certain path in G during information spread. Intuitively, if we generate a number of RR sets for random nodes in G, then the nodes appearing in more RR sets should have a higher probability to activate more nodes in G. This can be used to estimate the influence of a seed set: the set of nodes covering more RR sets yields higher probability to spread influence in G. Based on the rationale, Borgs et al. (2014) proposed RIS algorithm to solve influence maximization problem. The algorithm runs in two steps:

- Generate a collection of RR sets for random nodes from G.
- Use the greedy algorithm for the maximum coverage problem (Vazirani, 2001) to select k nodes to cover the maximum number of RR sets, and return the selected k nodes as the seed set.

Borgs et al. (2014) proved that if constructing RR sets for $144(m + n)\epsilon^{-3} \log(n)$, the RIS algorithm returns a $(1 - 1/e - \epsilon)$-approximate solution. Tang et al. (2014) further proved that if the number of generated RR sets is larger than $(8 + 2\epsilon)n(\log n + \log(\binom{n}{k}) + \log 2)/(OPT\epsilon^2)$, the RIS algorithm achieves a $(1 - 1/e - \epsilon)$-approximation ratio with at least $1 - 1/n$ probability.

4.2 Attribute-based Reverse Influence Sampling

Inspired by the RIS framework, we propose an attribute-based reverse influence sampling (ABRIS) framework. This framework includes three steps:

- *Sampling and Influence Estimation*: Firstly, we generate certain numbers of random RR sets for attributed groups in the network and use them to estimate the influence spread in different groups for the node set S . A basic theoretical lower bound is provided to derive the minimum number of RR sets to guarantee the estimation accuracy.
- *Attributed-based Hypergraph Construction*: Based on the generated RR sets, we construct an attribute-based hypergraph where the edges in the hypergraph represent the individual node's influence.

- *Node Selection*: Based on the attribute-based hypergraph, we can design algorithms to select nodes as seed set which can cause fair influence maximization in groups, and output the seed set with the given budget.

The details of these steps are explained below.

4.2.1 Sampling and Influence Estimation

In this step, we generate random RR sets for influence estimation under different groups. For each group ψ_i, we randomly sample θ_i nodes with replacement in group ψ_i to generate θ_i RR sets, denoted by $\mathcal{R}_{\psi_i}$, which are used for influence estimation. Theoretically, the more RR sets are generated, the more accurate the estimation is. However, the larger number of RR sets means a higher computational cost. Therefore it is important to determine the number of RR sets θ_i to guarantee some level of estimation accuracy.

Next, we provide theoretical analysis to derive the number of θ_i $(\forall \psi_i \in \Psi)$. The following theorems will be used in our derivation.

Lemma 4.1. *Given a graph G with groups, a RR set $\mathcal{R}(v)$ generated from a random node v in group $\psi_i \in \Psi$. For a seed set S, the expected influence of S in group ψ_i, denoted by $\mathbb{I}_{\psi_i}(S)$, can be estimated by $\mathbb{E}[\mathbb{I}_{\psi_i}(S)] = n_i \Pr[S \cap \mathcal{R}(v) \neq \emptyset]$.*

Lemma 4.1 is the direct extension of the work of Borgs et al. (2014), and the proof is omitted. According to Lemma 4.1, the expected influence spread of a seed set can be derived by the probability that S intersects $\mathcal{R}(v)$. According to definition 4.2 , $S \cap \mathcal{R}(v) \neq \emptyset$ is equal to S cover $\mathcal{R}(v)$, which probability can be estimated by the fraction of RR sets in $\mathcal{R}_{\psi_i}$ covered by S. This directly yields the following Lemma.

Lemma 4.2. *Let $F(S, \mathcal{R}_{\psi_i})$ be the fraction of RR sets in $\mathcal{R}_{\psi_i}$ covered by S. The equation $\mathbb{E}[\mathbb{I}_{\psi_i}(S)] = n_i \mathbb{E}[F(S, \mathcal{R}_{\psi_i})]$ always holds.*

Lemma 4.2 says that the influence for a group of a seed set S can be estimated by the ratio that S covers the RR sets in this group.

Lemma 4.3. *Let X be the sum of c i.i.d. random variables sampled from a distribution on $[0,1]$ with a mean μ. For any $\delta > 0$, $\Pr[X - c\mu \geq \delta \cdot c\mu] \leq \exp(-\frac{\delta^2}{2+\delta} c\mu)$.*

Lemma 4.3 is the classical Chernoff bound (Mitzenmacher & Upfal, 2005), which proof is omitted. This lemma will be used to derive the minimum number of RR sets for a group to guarantee the estimation accuracy of influence spread.

Theorem 4.4. *Consider an attribute graph G where the number of nodes attached with attribute ψ_i is n_i. Let OPT_i be the expected maximum influence in attribute ψ_i of any size-k seed set. Let θ_i be the number of RR sets generated for attribute ψ_i. For every given precision parameter $\epsilon \in (0,1)$, if θ_i satisfies:*

$$\theta_i \geq \frac{n_i(\epsilon + 2)(\log 2n_i|\Psi| + \log \binom{n}{k})}{OPT_i \cdot \epsilon^2} \tag{3}$$

Then, for any set S with k nodes, the following inequality holds with larger than $1 - 1/(n_i|\Psi|\binom{n}{k})$ probability:

$$|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| < \epsilon \cdot OPT_i. \tag{4}$$

Proof. By Lemma 4.1 , we have

$$\rho_i = \mathbb{E}[F(S, \mathcal{R}_{\psi_i})] = \mathbb{E}[\mathbb{I}_{\psi_i}(S)]/n_i$$

where ρ_i is the expected percentage of nodes with attribute ψ_i that are influenced by S. We have:

$$\begin{aligned}
&\Pr[|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| \geq \epsilon \cdot OPT_i] \\
&= \Pr[|\theta_i F(S, \mathcal{R}_{\psi_i}) - \rho_i \theta_i| \geq \frac{\epsilon \theta_i}{n_i} \cdot OPT_i] \\
&= \Pr[|\theta_i F(S, \mathcal{R}_{\psi_i}) - \rho_i \theta_i| \geq \frac{\epsilon \cdot OPT_i}{n_i \rho_i} \cdot \rho_i \theta_i]
\end{aligned}$$

Let $\delta = \epsilon \cdot OPT_i/(n_i \rho_i)$. Appling Lemma 4.3, and the fact that $\rho_i = \mathbb{E}[\mathbb{I}_{\psi_i}(S)]/n_i \leq OPT_i/n_i$, we have:

$$\begin{aligned}
&\Pr[|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| \geq \epsilon \cdot OPT_i] \\
&< 2\exp(-\frac{\delta^2}{2+\delta} \cdot \rho_i \theta_i) \\
&= 2\exp(-\frac{\epsilon^2 \cdot OPT_i^2}{2{n_i}^2 \rho_i + \epsilon n_i \cdot OPT_i} \cdot \theta_i) \\
&\leq 2\exp(-\frac{\epsilon^2 \cdot OPT_i^2}{2 n_i OPT_i + \epsilon n_i \cdot OPT_i} \cdot \theta_i) \\
&= 2\exp(-\frac{\epsilon^2 \cdot OPT_i}{(2+\epsilon) \cdot n_i} \cdot \theta_i)
\end{aligned}$$

Let the right part of the above equation less than $1/(n_i|\Psi|\binom{n}{k})$, that is $2\exp(-\frac{\epsilon^2 \cdot OPT_i}{(2+\epsilon)\cdot n_i} \cdot \theta_i) \leq 1/(n_i|\Psi|\binom{n}{k})$, and thus we have

$$\theta_i \geq \frac{n_i(\epsilon+2)(\log 2n_i|\Psi| + \log\binom{n}{k})}{OPT_i \cdot \epsilon^2}$$

Therefore the theorem is proved. □

Theorem 4.4 gives a basic bound for the estimation guarantee that if the generated RR sets exceed a certain number, the estimation of influence spread based on the RR sets approaches its expectation with a high probability (larger than $1/(n_i|\Psi|\binom{n}{k})$). To determine the minimum number of θ_i in Eq.(3), we need to know the exact value of OPT_i which is hard to obtain. We can only calculate the lower bound of OPT_i , notated as OPT_i^-, to ensure enough RR sets. Fortunately, there have been plenty of works on obtaining the lower bound of the maximum influence of a set of nodes in social networks (Tang et al., 2014; Li et al., 2011, 2015; Nguyen, Dinh, & Thai, 2016a; Li, Chen, Feng, Tan, & Li, 2014), which can be applied to calculate a lower bound of the maximum influence in a particular group as well. Here we adopt the iterative estimation method (Tang et al., 2014) to get the lower bound OPT_i^-.

4.2.2 Attribute-based Hypergraph Construction

After generating enough RR sets for each group ψ_i, we construct an attribute-based hypergraph $\mathcal{H}$ based on the RR sets. Let V be the set of nodes in G, and $\mathcal{R} = \cup_i \mathcal{R}_{\psi_i}$ be the set of generated RR sets. The hypergraph comprises the nodes in V and all the RR sets in $\mathcal{R}$. For each node $v \in V$, if RR set contains node v, there is an edge from v to this RR set. To clearly represent the influence in different groups, the RR sets are grouped according to the group identification of the sampled nodes to simulate inverse influence. An example of the constructed hypergraph is illustrated in Figure 2. Note that the red nodes in RR sets are the randomly chosen nodes to simulate inverse influence spread. The detailed algorithm to construct the attribute-based hypergraph is described in algorithm 1.

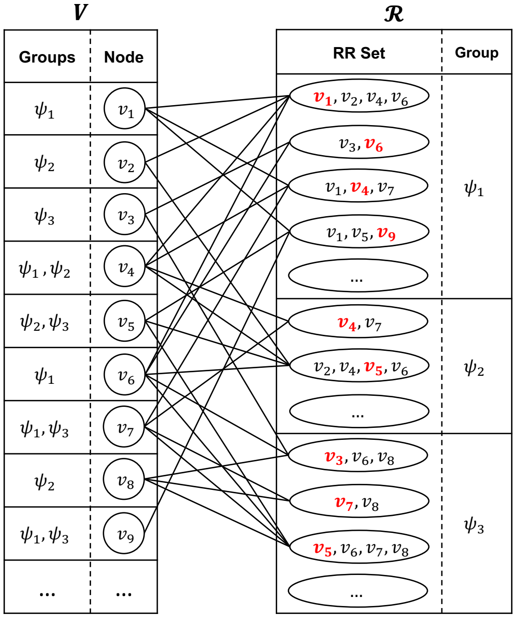


Figure 2: An example of hypergraph construction. On the left are graph nodes with their groups. On the right are the RR sets. Via a random node (the red one) sampled from a specific group with replacement, each RR set is generated by the inverse influence spread. And then each RR set links to the left according to its contained nodes.

According to the reverse influence sampling mechanism, the probability of the influence spread of a node is proportional to the number of RR sets it covers. In the hypergraph, an edge represents the relationship that a node covers a RR set. Therefore the influence of a node in V can be estimated by its degree (the number of RR sets linked) in the hypergraph. For a node $v \in V$, let $d_{\psi_i}(v)$ be the number of edges from v to the RR sets of group ψ_i. The total degree regarding group ψ_i for a seed set S can be represented as $d_{\psi_i}(S)$. According to Lemma 4.1 , the influence in group ψ_i of the seed set S can be estimated by

$$\hat{\mathbb{I}}_{\psi_i}(S) = n_i F(S, \mathcal{R}_{\psi_i}) = n_i \frac{d_{\psi_i}(S)}{\theta_i}. \tag{5}$$

```
Algorithm 1 BuildHypergraph ($(G, \Theta, \Psi)$)
Input
$\Theta : \{\theta_i\}$, the number of RR sets in $V_{\psi_i}$
$\Psi$ : The set of groups
Output  $\mathcal{H}$: The constructed hypergraph
Initialize $\mathcal{H}(\mathcal{R} = \emptyset, V = \emptyset, E = \emptyset)$
for attributed group $\psi_i$ in $\Psi$ do
    while $|\mathcal{R}_{\psi_i}| < \theta_i$ do
        Choose a node $u$ from $V_{\psi_i}$ uniformly at random and mark down its group $\psi_i$.
        Simulate inverse influence spread, starting from $u$.
        Let $Z$ be the discovered nodes set.
        Generate an $RR$ set based on $Z$ and add it into $\mathcal{R}_{\psi_i}$.
        Add the edges between $Z$ and $V$ to $\mathcal{H}$.
    end while
end for
return $\mathcal{H}$
```

4.2.3 Node Selection

On the foundation of previous two steps, we design two algorithms ABRIS-G and ABRIS-T to select seed nodes for fair influence maximization.

Algorithm ABRIS-G:

ABRIS-G is a basic greedy node selection algorithm, which runs round by round iteratively with a scheduling strategy according to the groups. In this algorithm, we first adopt Theorem 4.4 to construct the attribute-based hypergraph $\mathcal{H}$ since the number of required RR sets is easy to obtain. And then the procedure to select seed nodes is as follows:

(1) The seed set is initialized as an empty set, and its influence on each group is set to a random tiny real number larger than 0.

(2) In each round, with the constructed hypergraph, the algorithm estimates influence in each group as $n_i \frac{d_{\psi_i}(S)}{\theta_i}$ and chooses the group ψ_i with the least influence percentage. Then for each node in V, the algorithm counts the number of its linked RR sets in group ψ_i, namely $d_{\psi_i}(v)$. the node v with the maximum number is chosen to expand S, i.e., $S = S + \{v\}$.

(3) Removes v and all the RR sets linked by v from the hypergraph.

(4) Repeat (1)-(3) until the budget reaches, i.e., $|S| = k$.

The detailed algorithm is described in Algorithm 2. Since the algorithm identifies the group with the least influence percentage in each round and selects the node with the maximum influence in the group, it can achieve influence maximization while balancing the influence on different groups.

Complexity: we discuss the complexity of the whole process of ABRIS-G under the ABRIS framework. As illustrated, there are three steps in the whole process of ABRIS-G.

Algorithm 2 ABRIS-G (G, ϵ, k, Ψ)

```
Input:
$G$ : The graph with attributed groups; $\Psi$ : The attributed group set
$k$ : The number of budget; $\epsilon$ : The precision parameter for influence estimation
Output:
$S_k^*$: The set of selected vertices to spread influence
$\Theta = \emptyset$, $S_k^* = \emptyset$
for $\psi_i$ in $\Psi$ do
    Calculate $OPT_i^-$
    Compute $\theta_i$ by Eq.(3) and add $\theta_i$ to $\Theta$
end for
Construct $\mathcal{H} = \text{BuildHypergraph}(G, \Theta, \Psi)$
while $|S_k^*| < k$ do
    $\psi_i = \text{argmin}_{\psi_i}(\mathbb{E}[\mathbb{I}_{\psi_i}(S_k^*)]/n_i)$
    $v = \text{argmax}_v d_{\psi_i}(v)$
    Remove $v$ with all its connected RR sets and corresponding hyperedges
    $S_k^* = S_k^* \cup \{v\}$
end while
return $S_k^*$
```

The first step is sampling and influence estimation. In this step, we should calculate θ_i for each group. If we adopt the iterative estimation method (Tang et al., 2014) to estimate OPT_i, we can obtain the complexity of $O((m+n_i)\log n_i)$ for each group. Thus for all groups, the complexity is $O(\sum_i (m+n_i)\log n_i) = O(|\Psi|(m+n)\log n)$. In the second step, the main procedure is to generate RR sets for attribute-based hypergraph construction. Let $\bar{a}_i$ denote the expected number of random spread to generate an RR set for a randomly selected node in V_{ψ_i}. For a group ψ_i the complexity is $O(\theta_i \bar{a}_i)$. According to the work (Borgs et al., 2014), $\bar{a}_i \leq \frac{m}{n_i} OPT_i$. Together with Theorem 4.4, we have $O(\theta_i \bar{a}_i) = O(km \log n/\epsilon^2)$. Therefore for all groups, the complexity is $O(k|\Psi| m \log n/\epsilon^2)$. In the third step, the complexity of node selection can be easily derived as $O(kn)$ for we process k selection procedures and in each procedure, we traverse group influence for all nodes. Summarizing the above results, the complexity of ABRIS-G is $O(k|\Psi|(m+n)\log n/\epsilon^2)$.

Algorithm ABRIS-T:

Theorem 3.1 indicates that the objective function for the fair influence maximization problem is neither monotone nor submodular. Thus it's hard to achieve the objective of Eq.(2) directly. But on closer inspection, we can first consider maximizing the third term of $n \cdot \min_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\}$, which means the lowest proportional of influence over all attributed groups. Because when $\mathbb{E}(\mathbb{I}(S))$ is large enough, the continuous growth of the lowest group proportion influence would certainly lead to the decline of the highest group influence proportion until the two values are close. Thus continuously maximizing the third term of $n \cdot \min_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\}$ indicates minimizing the second term of $n \cdot \max_{\psi_i \in \Psi}\{\frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\}$, and the first term of $\mathbb{E}(\mathbb{I}(S))$ can also be relatively large. In this situation we can further utilize a simple greedy approach to seek for larger overall influence $\mathbb{E}(\mathbb{I}(S))$ so that the trade-off between influence maximization and group fairness can be achieved for better performance of the objective function $\mathbb{F}(\cdot)$.

Fortunately, many works (Krause et al., 2008; Tsang et al., 2019; Udwani, 2021) proposed algorithms for finding a set S with size k so that $S = \arg\max_{S \subseteq V} \min_{\psi_i \in \Psi}\{n \cdot \frac{\mathbb{E}(\mathbb{I}_{\psi_i}(S))}{n_i}\}$, which is called the MaxMin problem. The experiments of these studies show that the solution to the MaxMin problem would lead to group fairness to some degree. However, the algorithms of the MaxMin problem have two deficiencies according to our goal of large-scale fair influence maximization. One is that the algorithms are not suitable for large-scale social networks. All of them are run in small-size (less than 10k nodes) networks. While the other deficiency is that MaxMin does not take overall influence into consideration. When some groups are not well-connected, the algorithms would probably select many low-influence seeds into the result set. Even though great group fairness may be achieved, it does not satisfy the initial idea of influence maximization. Therefore, we develop a two-phase node selection algorithm ABRIS-T. The first phase is the larger-scale phase, where ABRIS-T maximizes the lowest proportional of influence among all attributed groups with a guarantee in large-scale social networks. The second phase is the larger-influence phase, where ABRIS-T uses a greedy method to pursue a larger influence spread for a better performance of fair influence maximization.

(1) ***The first phase: larger-scale phase.*** Generally, the MaxMin problem can be solved by finding the maximum constraint proportional factor α for a special *multi-objective optimization problem*: given a set of monotone submodular functions for attributed groups $F = \{\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]\}$ and a set of corresponding determined values $T = \{\alpha n_i\}$ which serve as constraints with the same proportional factor α, find a set S with size k so that $\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)] \geq \alpha n_i$ for every group ψ_i that has got to work. Thus in order to solve the MaxMin problem, we can continuously binary search for the largest proportional factor α with a feasible solution set S for the special multi-objective optimization problem. Once obtaining the maximum α with a feasible solution S, the MaxMin problem is finally settled.

Actually, the special multi-objective optimization problem is a special case of the robust submodular observation selection (RSOS) problem (Krause et al., 2008), which constrains arbitrarily for different monotone submodular functions. In work (Udwani, 2021), a state-of-the-art three-stage method based on Multiplicative-Weight-Updates (MWU) is proposed for the RSOS problem. MWU returns a solution with $\omega(1-1/e)^2 - \delta$ approximation and $O(\frac{n}{\lambda^3}\log|\Psi|\log\frac{n}{\lambda})$ time complexity for $|\Psi| = o(k/\log^3 k)$ submodular functions, where λ is a precision parameter and ω is a small factor approaching 1. Similarly, for the monotone submodular functions $F = \{\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]\}$ and feasible determined values $T = \{\alpha n_i\}$, with a precision parameter λ, the MWU method returns an $(\omega(1-1/e)^2 - \lambda)$-approximate solution for the special multi-objective optimization problem. However, calculating $\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]$ costs $O(m+n)$, and the joint complexity of adopting MWU is $O((m+n)\frac{n}{\lambda^3}\log|\Psi|\log\frac{n}{\lambda})$, which is extremely computationally expensive in large-scale social networks. But with the attribute-based hypergraph $\mathcal{H}$ in the ABRIS framework, $\hat{\mathbb{I}}_{\psi_i}(\cdot)$ can be directly utilized to estimate $\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]$. Such issue can be settled efficiently with a guarantee.

Under the ABRIS framework by exploiting $\hat{\mathbb{I}}_{\psi_i}(\cdot)$ to estimate $\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]$, the approximation guarantee to MWU also concerns the number of generated RR sets. But the number of RR sets derived beforehand by Theorem 4.4 cannot meet the requirement. Therefore, we prove the following Theorem 4.5 and Theorem 4.6, and derive the least amount of RR sets for each attributed group to guarantee the performance of MWU under ABRIS framework with a high probability (larger than $1 - 1/\min_i\{n_i\}$).

Theorem 4.5. *Consider a graph G with attributed groups where the number of nodes attached to group ψ_i is n_i. Let OPT_i be the expected maximum influence in group ψ_i of any size-k seed set. Assume that the constraint αn_i serves as the accuracy criterion where α is the proportionality factor. Let θ_i be the number of RR sets generated for group ψ_i. For every precision parameter $\epsilon \in (0, 1)$, if θ_i satisfies:*

$$\theta_i \geq (\log 2n_i|\Psi| + \log \tbinom{n}{k})\frac{2OPT_i + \epsilon\alpha n_i}{\epsilon^2\alpha^2 n_i} \tag{6}$$

Then, for any set S with k nodes, $|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| < \epsilon\alpha n_i$ holds with larger than $1 - 1/(n_i|\Psi|\binom{n}{k})$ probability.

Proof. Similar with the proof in Theorem 4.4, if $\rho_i = \mathbb{E}[F(S, \mathcal{R}_{\psi_i})] = \mathbb{E}[\mathbb{I}_{\psi_i}(S)]/n_i$ is the expected percentage of nodes in group ψ_i that are influenced by S. We have:

$$\begin{aligned}
&\Pr[|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| \geq \epsilon \cdot \alpha n_i] \\
=& \Pr[|\theta_i F(S, \mathcal{R}_{\psi_i}) - \rho_i\theta_i| \geq \frac{\epsilon\theta_i}{n_i} \cdot \alpha n_i] \\
=& \Pr[|\theta_i F(S, \mathcal{R}_{\psi_i}) - \rho_i\theta_i| \geq \frac{\epsilon \cdot \alpha n_i}{n_i\rho_i} \cdot \rho_i\theta_i]
\end{aligned}$$

Let $\delta = \epsilon\alpha/\rho_i$. Applying Lemma 4.3, and the fact that $\rho_i = \mathbb{E}[\mathbb{I}_{\psi_i}(S)]/n_i \leq OPT_i/n_i$, we have:

$$\begin{aligned}
&\Pr[|n_i F(S, \mathcal{R}_{\psi_i}) - \mathbb{E}[\mathbb{I}_{\psi_i}(S)]| \geq \epsilon \cdot \alpha n_i] \\
<& 2\exp(-\frac{\delta^2}{2+\delta} \cdot \rho_i\theta_i) \\
=& 2\exp(-\frac{\epsilon^2 \cdot \alpha^2}{2\rho_i + \epsilon\alpha} \cdot \theta_i) \\
\leq& 2\exp(-\frac{\epsilon^2\alpha^2 n_i}{2OPT_i + \epsilon\alpha n_i} \cdot \theta_i)
\end{aligned}$$

Let the right part of the above equation less than $1/(n_i|\Psi|\binom{n}{k})$, that is $2\exp(-\frac{\epsilon^2\alpha^2 n_i}{2OPT_i+\epsilon\alpha n_i} \cdot \theta_i) \leq \frac{1}{n_i|\Psi|\binom{n}{k}}$, and thus we have

$$\theta_i \geq (\log 2n_i|\Psi| + \log \tbinom{n}{k})\frac{2OPT_i + \epsilon\alpha n_i}{\epsilon^2\alpha^2 n_i}$$

Therefore the theorem is proved. □

With such number of RR sets for hypergraph $\mathcal{H}$, the following theorem of approximation guarantee in terms of MWU under ABRIS framework can be derived:

Theorem 4.6. *On the foundation of ABRIS framework, assume that the number of RR sets in the attribute-based hypergraph satisfies Eq.(6) with the accuracy criterion αn_i at the precision parameter $\epsilon \in (0, 1)$. For the special multi-objective optimization problem with $F = \{\mathbb{E}[\mathbb{I}_{\psi_i}(\cdot)]\}$ and $T = \{\alpha n_i\}$, where $|F'| = |F| = |\Psi| = o(k/\log^3 k)$, if using $F' = \{\hat{\mathbb{I}}_{\psi_i}(\cdot)\}$ to estimate F, the MWU method at precision parameter λ returns an $(\omega(1 - 1/e)^2 - \lambda - \epsilon)$-approximate solution with probability $1 - 1/\min_i\{n_i\}$.*

Proof. With parameters F', T, λ the MWU method returns a result S_k with the fact that

$$\hat{\mathbb{I}}_{\psi_i}(S_k) \geq [\omega(1 - 1/e)^2 - \lambda] \cdot \alpha n_i$$

In this case, by applying Theorem 4.5 we have:

$$\begin{aligned}
\mathbb{E}[\mathbb{I}_{\psi_i}(S_k)] &> n_i F(S_k, \mathcal{R}_{\psi_i}) - \epsilon \cdot \alpha n_i \\
&= \hat{\mathbb{I}}_{\psi_i}(S_k) - \epsilon \cdot \alpha n_i \\
&\geq [\omega(1 - 1/e)^2 - \lambda] \cdot \alpha n_i - \epsilon \cdot \alpha n_i \\
&= [\omega(1 - 1/e)^2 - \lambda - \epsilon]\alpha n_i
\end{aligned}$$

Obviously, the output of MWU under the ABRIS framework also achieves comparable approximation when ϵ is small. Specifically, the theorem here should hold simultaneously for all size-k seed sets in all groups and thus, the solution is obtained with probability $1 - 1/\min_i\{n_i\}$. □

Observing the required number of RR sets in Eq.(6), we can find that OPT_i is in the numerator of the equation. It is different from Eq.(3). Thereby, we should calculate an upper bound for OPT_i to ensure enough RR sets for the guarantee. We notate the upper bound of OPT_i as OPT_i^+ and give an efficient hypergraph-based method as follows.

The method adopts a hypergraph $\mathcal{H}$ constructed beforehand by Theorem 4.4. Besides, we introduce a greedy algorithm $\mathcal{A}_i(\mathcal{H}, k)$ to obtain an optimal size-k seed set S_i^k for the maximum influence in a specific group ψ_i. It is analogous to the vanilla IM problem solution based on RIS (Borgs et al., 2014; Tang et al., 2014, 2015) via a single modification. $\mathcal{A}_i(\mathcal{H}, k)$ runs for k rounds. In each round for every node in V, $\mathcal{A}_i(\mathcal{H}, k)$ counts the number of linked RR sets in group ψ_i. The node v with the maximum number is chosen to expand S_i^k, and then the node v together with all its linked RR sets is not considered in later rounds. It's with algorithm $\mathcal{A}_i(\mathcal{H}, k)$ that we can prove the following theorem for OPT_i^+ calculation.

Theorem 4.7. *When using $\Theta = \{\theta_i\}$ derived by Eq.(3) with a given precision parameter ϵ' in Theorem 4.4 to construct the hypergraph $\mathcal{H}$, and then adopting $\mathcal{A}_i(\mathcal{H}, k)$ to get a seed set S_i^k, we can have:*

$$OPT_i \leq \frac{\mathbb{E}[\mathbb{I}_{\psi_i}(S_i^k)]}{(1 - 1/e)(1 - \epsilon') - \epsilon'} \tag{7}$$

holds simultaneously for all groups Ψ with larger than $1 - 1/\min\{n_i\}$ probability.

Proof. Obviously, $\mathcal{A}_i(\mathcal{H}, k)$ is a standard greedy algorithm for set covering problem in order to find k nodes covering the maximum number of RR sets in group ψ_i. Assuming the optimal solution for the maximum RR sets coverage is $\tilde{S}_i^k$ and the optimal solution resulting in OPT_i is $\bar{S}_i^k$, we can derive:

$$d_{\psi_i}(S_i^k) \geq (1-1/e)d_{\psi_i}(\tilde{S}_i^k) \geq (1-1/e)d_{\psi_i}(\bar{S}_i^k)$$

The first inequality is obtained as the result of the ordinary bound through adopting the greedy approach to solve set covering problem. Since $\tilde{S}_i^k$ is the optimal solution for the set covering problem, there is no doubt that the second inequality holds. Thus applying Theorem 4.4, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{I}_{\psi_i}(S_i^k)] &> n_i F(S_i^k, \mathcal{R}_{\psi_i}) - \epsilon' OPT_i \\
&= n_i \frac{d_{\psi_i}(S_i^k)}{\theta_i} - \epsilon' OPT_i \\
&\geq n_i \frac{(1-1/e)d_{\psi_i}(\bar{S}_i^k)}{\theta_i} - \epsilon' OPT_i \\
&= (1-1/e)n_i F(\bar{S}_i^k, \mathcal{R}_{\psi_i}) - \epsilon' OPT_i \\
&\geq (1-1/e)(1-\epsilon')OPT_i - \epsilon' OPT_i
\end{aligned}$$

Therefore

$$OPT_i \leq \frac{\mathbb{E}[\mathbb{I}_{\psi_i}(S_i^k)]}{(1-1/e)(1-\epsilon')-\epsilon'}$$

For the reason that Theorem 4.4 holds with at least $1 - 1/(n_i|\Psi|\binom{n}{k})$ probability, by the union bound, the theorem here should hold simultaneously for all size-k node sets in all attributed groups with larger than $1 - \sum_i 1/(|\Psi|n_i) \geq 1 - 1/\min_i\{n_i\}$ probability. □

According to Theorem 4.7, we can first use a preliminary attribute-based hypergraph $\mathcal{H}$ constructed beforehand by Eq.(3) to calculate an upper bound OPT_i^+. With the upper bound OPT_i^+, we then obtain the required number of RR sets (Theorem 4.5) for the multi-objective optimization guarantee (Theorem 4.6). Only after generating enough RR sets into the preliminary hypergraph $\mathcal{H}$ (add more if not enough) can we continue to solve the MaxMin problem by MWU($\{\hat{\mathbb{I}}_{\psi_i}(\cdot)\}, \{\alpha n_i\}, \lambda$) under the ABRIS framework accordingly.

(2) ***The second phase: larger-influence phase.*** Once the optimal solution S_k^* for the MaxMin problem is obtained, we have finished the first phase of ABRIS-T. However, the solution in the first phase may perform badly in overall influence when some groups are worse connected than others since it probably expends many low-influence seeds to improve the influence fractions of the badly-connected groups. Thus in the second phase, we make up for such a disadvantage with a greedy approach. It can achieve the trade-off between influence maximization and group fairness for better performance of the objective function. For each time we find a node with the lowest influence in solution S_k^*. We exchange it for another node in $V \setminus S_k^*$ with a larger overall influence to chase the best performance of the objective function $\mathbb{F}(\cdot)$. Repeat the procedure until the objective function cannot get larger anymore. Even though the second greedy phase is not guaranteed, it can be proven effective in balancing the overall influence and disparity through experiments. In this way, the whole process of algorithm ABRIS-T under the ABRIS framework is shown in Algorithm 3.

Algorithm 3 ABRIS-T $(G, \epsilon, \epsilon', \lambda, k, \Psi)$

```
Input:
$G$ : The graph with attributed groups; $k$ : The number of budget
$\Psi$ : The group set; $\kappa$ : The stop condition for binary search
$\epsilon, \lambda$ : The precision parameters for influence estimation and MWU method
$\epsilon'$ : The parameter for $OPT^+$ calcualtion
Output:
$S_k^*$: the set of selected vertices to spread influence
 1: $\Theta = \emptyset$
 2: for $\psi_i$ in $\Psi$ do
 3:     Calculate $OPT_i^-$
 4:     With $OPT_i^-$ and $\epsilon'$, compute $\theta_i$ by Eq.(3); and then add $\theta_i$ to $\Theta$
 5: end for
 6: Construct $\mathcal{H} = \text{BuildHypergraph}(G, \Theta, \Psi)$
 7: for $\psi_i$ in $\Psi$ do
 8:     Calculate $OPT_i^+$ by Theorem 4.7
 9: end for
10: $a = 0, b = \min_i\{OPT_i^+/n_i\}, \alpha = (a+b)/2$
11: while $(b - a) \cdot n > \kappa$ do
12:     for $\psi_i$ in $\Psi$ do
13:         With $OPT_i^+, \epsilon$ and $\alpha$, compute $\theta'_i$ by Eq.(6) for the required guarantee
14:         if $\theta_i < \theta'_i$ then
15:             Generate $(\theta'_i - \theta_i)$ RR sets with $\psi_i$
16:             Add these RR sets to $\mathcal{H}$ and $\theta_i = \theta'_i$
17:         end if
18:     end for
19:     $S_k^* = \text{MWU}(\{\hat{\mathbb{I}}_{\psi_i}(\cdot)\}, \{\alpha n_i\}, \lambda)$ (Udwani, 2021)
20:     if $S_k^*$ is feasible then
21:         $a = \alpha, \alpha = (\alpha + b)/2$
22:     else $b = \alpha, \alpha = (\alpha + a)/2$
23:     end if
24: end while
25: $\mathbb{F}' = \mathbb{F}(S_k^*)$
26: while True do
27:     $u = \arg\min_{u \in S_k^*} \hat{\mathbb{I}}(u)$
28:     $v = \arg\max_{v \in V \setminus S_k^*,\, \hat{\mathbb{I}}(v) > \hat{\mathbb{I}}(u)} \mathbb{F}(S_k^* \cap \{v\} \setminus \{u\})$
29:     if $\mathbb{F}(S_k^* \cap \{v\} \setminus \{u\}) > \mathbb{F}'$ then
30:         $S_k^* = S_k^* \cap \{v\} \setminus \{u\}$ and $\mathbb{F}' = \mathbb{F}(S_k^*)$
31:     else return $S_k^*$
32:     end if
33: end while
```

Algorithm 3 can be divided into four parts. The first part from line 1 to line 6 is constructing the preliminary hypergraph by Eq.(3), which is the same as ABRIS-G. The second part is from line 7 to line 9. The process intends to obtain the upper bound OPT_i^+ by Theorem 4.7. With OPT_i^+, we can calculate the required number of RR sets through Eq.(6) and also give a relatively small value as the beginning for binary search. The third part from line 10 to line 24, is the process of adopting the MWU method in large-scale social networks based on the hypergraph $\mathcal{H}$. In the process, binary search is utilized to seek the largest feasible α with determined constraints $T = \{\alpha n_i\}$. In each α-search iteration, according to Theorem 4.6 with the searched α, we should generate enough RR sets into hypergraph $\mathcal{H}$ (add more if not enough) to ensure the approximation guarantee. And then we make use of $\text{MWU}(\{\hat{\mathbb{I}}_{\psi_i}(\cdot)\}, \{\alpha n_i\}, \lambda)$ to obtain a feasible solution S_k^*. The fourth part is from line 25 to line 33. In this part, we try to elevate overall influence through a greedy approach. Each time from the pending seed set S_k^*, we find the node with minimum influence and then node u is replaced by a node $v \in V \setminus S_k^*$ with a larger overall influence to bring about the maximum increase of the objective function. Note that instead of computing $\mathbb{E}(\mathbb{I}(\cdot))$, we exploit $\hat{\mathbb{I}}(\cdot)$ through the hypergraph to reduce complexity for the procedure. The objective of this part is to find a higher overall influence to obtain a better objective result.

Complexity: we also discuss the complexity of ABRIS-T. The complexity of the first part in the algorithm is similar with ABRIS-G, which is $O(k|\Psi|m \log n/\epsilon'^2)$. For the second part, the complexity of line 6 is $O(k|\Psi|n)$ because we process k selection procedures for $|\Psi|$ groups respectively, and in each procedure we traverse group influence for all nodes. The procedure of binary search consists of two major subprocedures: RR sets generation and MWU. RR sets generation can be combined with the complexity in the first part since they jointly construct hypergraph $\mathcal{H}$. Thus their joint complexity depends on the final number of RR sets. If we denote r as $\min_i\{\alpha n_i/OPT_i\}$, similar with the complexity analysis in ABRIS-G, we can derive the final complexity for hypergraph construction is $O(k|\Psi|(m + n) \log n/\epsilon^2 r^2)$. For the third part of MWU, the complexity is $O(\frac{n}{\lambda^3} \log |\Psi| \log \frac{n}{\lambda})$, and the total complexity including binary search procedure is $O(\log \min\{n_i\} \frac{n}{\lambda^3} \log |\Psi| \log \frac{n}{\lambda}) = O(n \log |\Psi| \log^2 n/\lambda^3)$. In the fourth part, assume that the iteration is executed for t times. In each time after identifying the node with minimum overall influence, we may search $O(n)$ times to find the node to exchange for. During each searching process, we need to estimate $\mathbb{F}(S_k^* \cap \{v\} \setminus \{u\})$ through hypergraph $\mathcal{H}$, the complexity is $O(k|\Psi|m \log n/\epsilon^2 r^2 n)$. Therefore, the complexity for the third part is $O(tk|\Psi|m \log n/\epsilon^2 r^2)$. To sum up, the final complexity for ABRIS-T is $O(k|\Psi|(m + n) \log n/\epsilon^2 r^2 + n \log |\Psi| \log^2 n/\lambda^3 + tk|\Psi|m \log n/\epsilon^2 r^2)$, which mostly depends on $O(n \log |\Psi| \log^2 n/\lambda^3)$. Therefore when the graph scale grows larger, the proportion of time consumption for the second phase will get smaller and smaller.

5. Performance Evaluation

In this section, we conduct experiments to evaluate the performances of our proposed algorithms based on six real-world datasets.

5.1 Experimental Settings

5.1.1 Default System Parameters

Here we present the setting of default system parameters. In the graph model, we follow the independent cascade model proposed by Kempe et al. (2003) and set the influence probability on each edge to 0.01 (Kempe et al., 2003; Chen, Wang, & Yang, 2009; Anwar et al., 2021). In the ABRIS-G algorithm, we set the precision parameter $\epsilon = 0.1$. For ABRIS-T algorithm, we make the parameters $\epsilon' = 0.1, \epsilon = 0.4, \lambda = 0.4$ and set $\kappa = 10$ as the stop condition of binary search. For each system settings, we repeat the experiment for 10 times to obtain the average value for performance evaluation and calculate the stand error as error bars to measure and compare the variability of different algorithms. The experiments are conducted on a personal computer with Intel Xeon E5-2620 v2 2.10GHz CPU and 128GB memory, running 64-bit CentOS Linux 7.2.

5.1.2 Datasets

The experiments are conducted on six real-world networks. They are three small-size graphs: UVM, UCSC and UPENN and three large-size graphs: DBLP, Pokec and AMiner. We consider overlapping groups for datasets UVM, UCSC, Pokec and AMiner and non-overlapping groups for UPENN and DBLP. The statistics of these networks are summarized in Table 2. And the details of the datasets are as below.

Table 2: Statistics of the datasets.

Name	**Nodes**	**Edges**	**Groups**	**Groups Description**
UVM	7,322	191,197	4	**Status:** Faculties (12%), Students (88%) **Grade:** Senior (40%), Junior (60%)
UCSC	8,990	224,545	4	**Status:** Faculties (10%), Students (90%) **Gender:** Males (45%), Females (55%)
UPENN	29,634	831,213	2	**Grade:** Senior (29%), Junior (71%)
DBLP	280,200	750,601	2	**Gender:** Males (77%), Females (23%)
Pokec	1,099,121	10,794,057	4	**Age:** The old (18%), The young (82%) **Gender:** Males (51%), Females (49%)
AMiner	1,560,640	4,258,946	5	**Nation:** Developing (42%), Developed (58%) **Study Interset:** Software (12%), Data (14%), Modeling (32%)

(1) **UVM (Traud, Mucha, & Porter, 2012)**: A facebook social network in UVM. We remove the nodes without user information in the network profile and choose four overlapping attributed groups according to users' status and grade. These groups are respectively the faculties (12%), the students (88%), the senior (40%) and the junior (60%).

(2) **UCSC (Traud et al., 2012)**: A facebook social network in UCSC. Based on the basic information of the users, we choose four overlapping groups in terms of status and gender. They are the faculties (10%), the students (90%), the males (45%) and the females (55%) respectively.

(3) **UPENN (Traud et al., 2012)**: A facebook social network of the users in UPENN. We remove the non-student users from the network and attach each node with the group of

senior and junior students in light of the year of enrollment. The majority are 71% junior students, and the minority are 29% senior students.

(4) **DBLP (Karimi, Génois, Wagner, Singer, & Strohmaier, 2018)**: A co-authorship network from DBLP, a website that provides comprehensive list of research papers in the area of computer science. The nodes represent scientists and the edges represent paper co-authorships. The groups are the male and the female scientists. The minority are 23% of the female, and the majority 77% of the male.

(5) **Pokec (Takac & Zabovsky, 2012)**: The most popular online social network in Slovakia, where the nodes represent users and the edges represent the friendship between them. We consider the groups as the old (19%), the young (81%), the males (51%) and the females (49%) in terms of the age and gender information.

(6) **AMiner (Zhang, Tang, Ma, Tong, Jing, & Li, 2015)**: A co-author graph extracted from AMiner.org. We consider 5 different overlapping groups. According to the affiliations, we divide the authors into two groups that they are from developing (42%) or developed (58%) nations. With respect to the descriptions of their study interests, we derive three groups as software (12%), data (14%) and modeling (32%) fields.

5.1.3 Metrics and Baselines

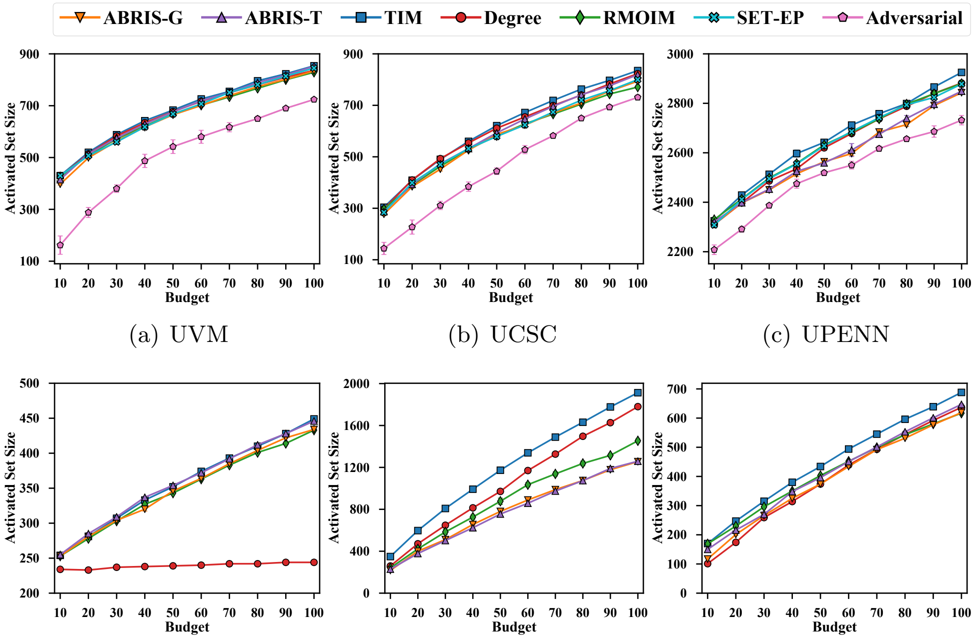
To evaluate the performance of ABRIS algorithm, we adopt the following three metrics.

- **Activated set size:** the total number of nodes activated by the seed set in the original graph, which measures the scale of influence spread.
- **Disparity:** the difference between the maximum and minimum influence percentage in the attributed groups defined in Eq. (1), which measures the fairness of influence spread in different attributed groups.
- **Objective function value:** the value to describe the fair influence result of a seed set. The calculation is defined in Eq.2. We use $\gamma = 1$ by default since in this case, overall influence and disparity penalty are both scaled by graph size n can achieve good trade-off performances in practice.

We implement two algorithms ABRIS-G and ABRIS-T. And we compare the proposed algorithms with other five algorithms: (1) Degree (Kempe et al., 2003): this algorithm selects the top-k nodes with highest degree centrality as seed set. (2) TIM (Tang et al., 2014): the state-of-the-art algorithm for influence maximization, which enhanced the RIS framework that runs efficiently with approximation guarantee. We make the parameters in this algorithm as $l = 1, \epsilon = 0.1$. (3) RMOIM (Gershtein et al., 2021): an algorithm based on multi-objective constraints. It adopts the vanilla RIS framework as well and can be used in large-scale social networks. We set the threshold parameter $t = (1 - 1/e)/|\Psi|$ as the authors suggest. (4) SET-EP (Becker et al., 2022): it proposed set-based and node-based randomized strategies for choosing the seed nodes. We choose the *set_based_ep* method because it shows greater performance with a determined seed set S. (5) Adversarial (Khajehnejad et al., 2021): By using an autoencoder coupled with a discriminator in an adversarial setting, the algorithm is the first to exploit embedding learning for fair influence maximization. Note that we display the results for SET-EP and Adversarial methods only in small-size graphs (i.e., UVM, UCSC, UPENN) because they respectively throw out of time (more than 24 hours) and out of memory errors when running in large-size graphs (i.e., DBLP, Pokec, AMiner).

5.2 Performance of ABRIS Algorithms

Activated Set Size Comparison. Figure 3 shows the total influence spread of different algorithms under different budgets. From the figure, we can see that TIM outperforms all other algorithms in all datasets. The reason is that TIM seeks maximum influence spread over the whole network. The standard error for all baselines is not significant. The overall influence for ABRIS-G and ABRIS-T are very close to TIM for all datasets except Pokec because the phenomenon of sacrificing overall influence for fairness in Pokec is visible. In addition, ABRIS-G and ABRIS-T perform very close even though they select seeds from different perspectives. In UVM, UCSC and UPENN datasets, the results of all methods are very close except Adversarial. Adversarial cannot perform well, probably because the feature space of the initial node representations is too large to enable DNN capturing key influence patterns. In DBLP, the method of degree heuristic performs much worse than the other methods, which implies that in academic collaboration networks the users with the most collaborators may not be the most influential ones.

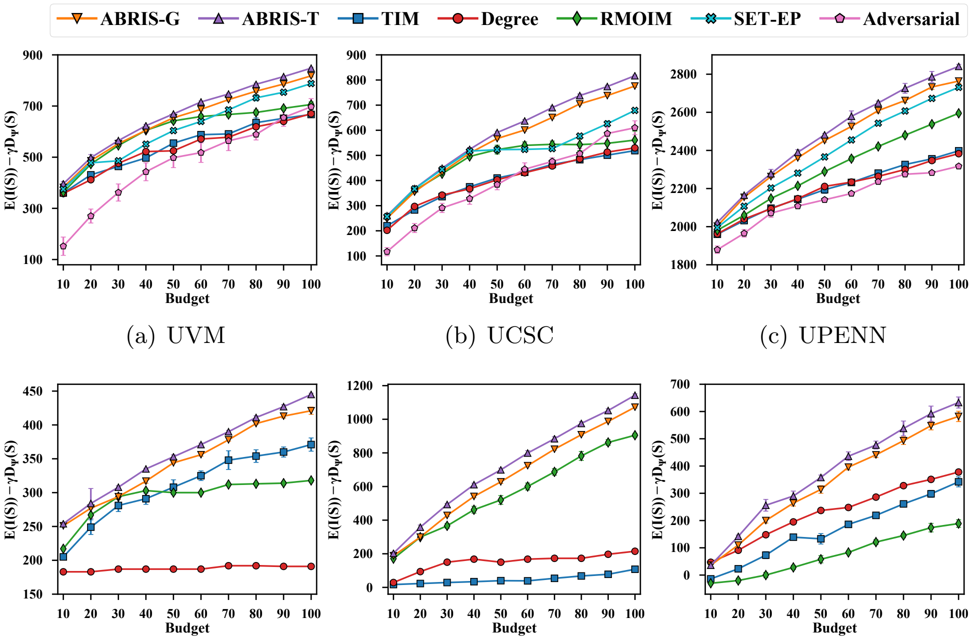


(a) UVM (b) UCSC (c) UPENN

(d) DBLP (e) Pokec (f) AMiner

Figure 3: Comparison of activated set size of different algorithms under different budgets. The x-axis stands for different sizes of seed sets. The y-axis is the overall influence caused by the seed sets from different algorithms.

Objective Function Comparison. Figure 4 compares the objective function Eq. (2) of different algorithms. When taking disparity as a penalty, the proposed ABRIS-G and ABRIS-T algorithms significantly outperform the other algorithms. Generally, ABRIS-T performs better than ABRIS-G for all datasets. In small-size graphs, the performance of the Adversarial method is not so good, mainly because it fails to maximize overall influence even though the disparities are relatively small. But SET-EP shows greater performance than the other baselines. In datasets UCSC and DBLP, when the budget size is small (less than 50), RMOIM shows comparable performance with ABRIS-G and ABRIS-T. Since TIM and Degree will cause a large disparity, they often exhibit low objective function values. For the variability of the algorithms, we can find that ABRIS-G and ABRIS-T have relatively small variabilities, and ABRIS-G showed better stability than ABRIS-T. The variability for Adversarial in UVM is visible, which is mainly due to the indeterminacy of deep learning.



(a) UVM

(b) UCSC

(c) UPENN

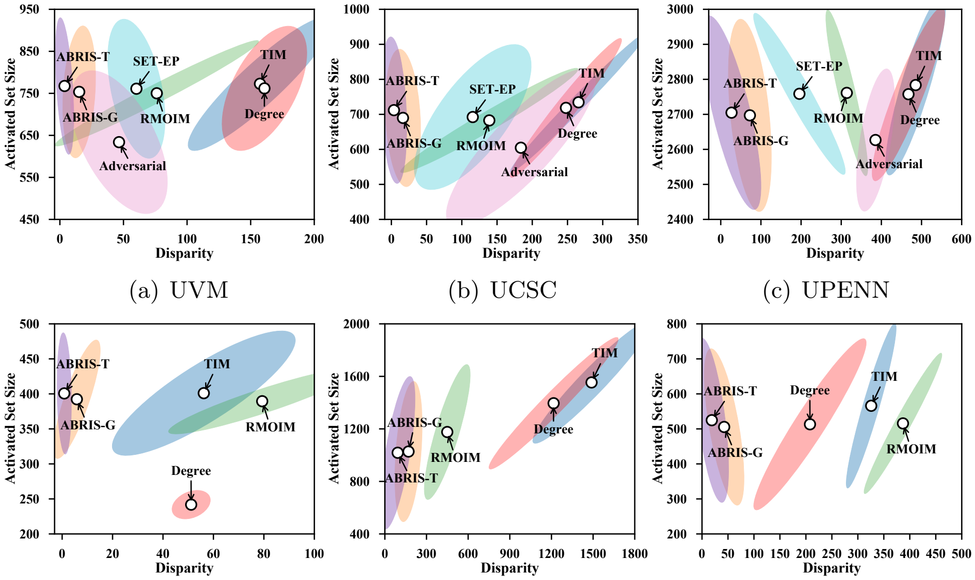
(d) DBLP

(e) Pokec

(f) AMiner

Figure 4: Comparison of objective function of different algorithms under different budgets. The x-axis stands for different sizes of seed sets. The y-axis is the objective function values caused by the seed sets from different algorithms.

Trade-off Results Comparison. We further study the trade-off between maximizing influence spread and minimizing disparity. The results of different algorithms are shown in Figure 5. We run simulations with a budget ranging from 50 to 100 and draw the $1 - \delta$ elliptic contour of the maximum-likelihood 2D Gaussian distribution. From the figure, TIM has the maximum activated set size, but its disparity is also very large. Therefore its influence spread among different attributed groups is highly unfair. In small-size graphs, RMOIM and SET-EP perform better than other baselines but worse than ABRIS-G and ABRIS-T. In DBLP, the performance of RMOIM is even worse than TIM. The ABRIS-G and ABRIS-T algorithms have much lower disparities than the other algorithms, and their activated set sizes are comparable to TIM. The performances for ABRIS-G and ABRIS-T are similar, but ABRIS-T is better than ABRIS-G in activated set size and disparity for all datasets. Therefore the proposed ABRIS algorithms can achieve a better trade-off between maximizing influence spread and minimizing disparity.

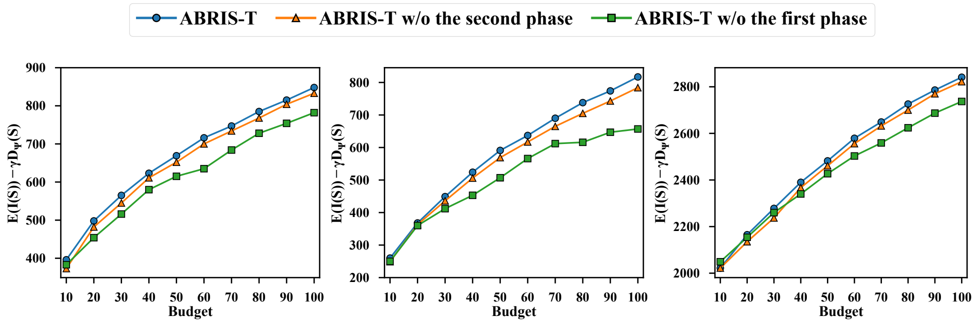


(a) UVM (b) UCSC (c) UPENN

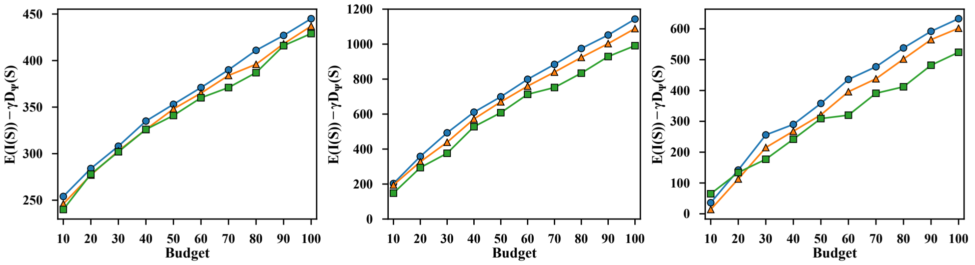
(d) DBLP (e) Pokec (f) AMiner

Figure 5: Comparison of the trade-off between influence maximization and disparity. The x-axis and y-axis display different levels of disparities and overall influences.

Ablation Experiment for ABRIS-T. ABRIS-T consists of two phases. One is the first larger-scale phase by adopting binary search. The other one is the second larger-influence phase to seek for larger overall influence spread. In order to study the influence of these two phases, we conduct an ablation experiment. The result is shown in Figure 6. From the figures, we can see that with either single phase, the resulting curve is lower than that of ABRIS-T. It means that these two phases are both effective in node selection to cause fair influence maximization. And generally speaking, the larger-scale phase with MWU is more important than the larger-influence phase since the results with a single larger-scale phase are better than those with a single larger-influence phase in all datasets.



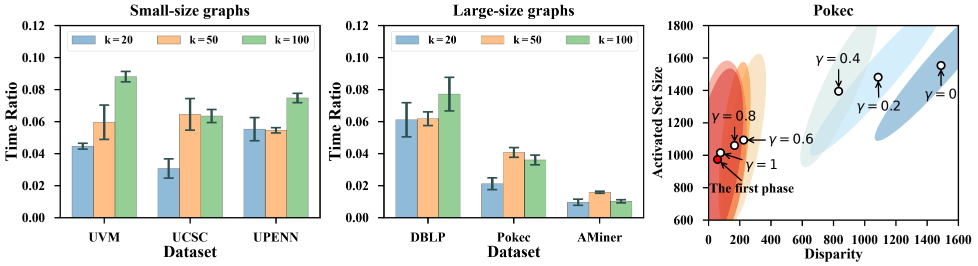
(a) UVM (b) UCSC (c) UPENN



(d) DBLP (e) Pokec (f) AMiner

Figure 6: Ablation experiments for ABRIS-T. The x-axis shows different sizes of seed sets. The y-axis is the objective function values regarding ABRIS-T, ABRIS-T without the first phase and ABRIS-T without the second phase.

Trade-offs for the second phase in ABRIS-T. The second phase for ABRIS-T is important in obtaining an effective seed set to balance the overall influence and disparity. Figures 7(a) and 7(b) display the ratios of time consumption for the second phase procedure. We can see that the time consuming for phase two takes up less than 10% in general, and when the graph grows larger in scale, the ratio will further decrease. The phenomenon conforms to the time complexity analysis of ABRIS-T. Thus running the second phase does not lead to considerable time consumption. In addition, we use Pokec to present the trade-off for the second phase between the overall influence and disparity in Figure 7(c). The red point is the result for only the first phase, where the phenomenon of sacrificing overall influence for less disparity is very visible. The white points are the results for further running the second phase under different γ values. The parameter γ is a discount factor to take disparity as a penalty to measure different importances of disparity. We can see that the second phase does have the ability to mitigate the phenomenon of selecting many low-influence nodes for less disparity, so that a better objective function performance is achieved. The extent of such mitigation can be tuned with parameter γ. The less importance of the disparity considered, the more extent for the mitigation in the second phase.



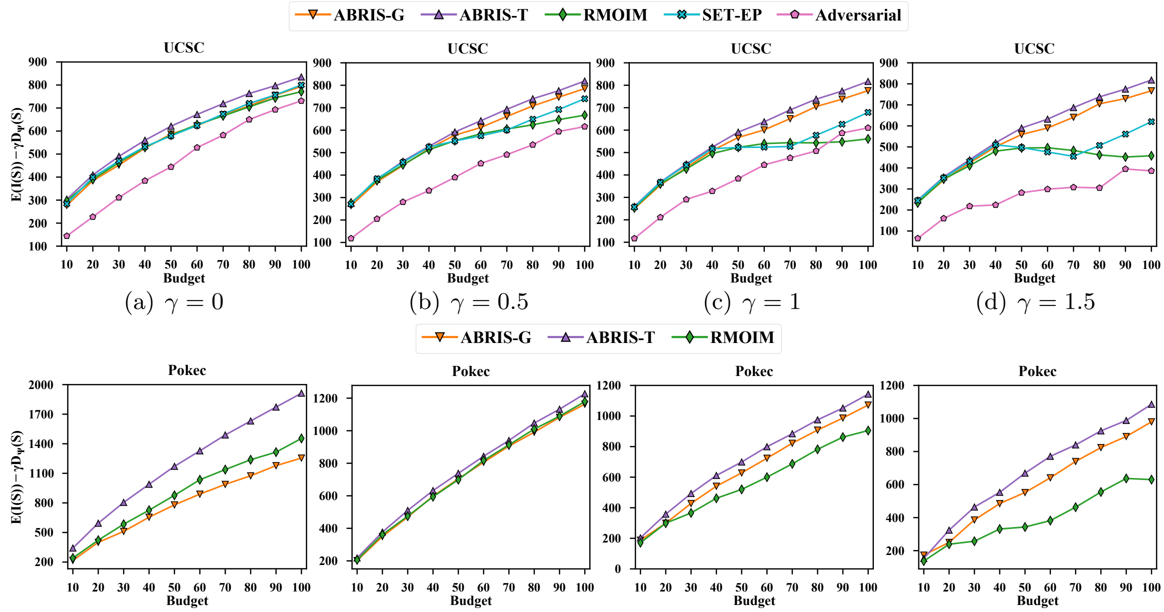
(a) (b) (c)

Figure 7: Trade-offs for the second phase in ABRIS-T. Figures 7(a) and 7(b) show the ratios of time consumption for the second phase. Figure 7(c) presents the trade-off between the overall influence and disparity for the second phase.

5.3 Parameter Analysis

In this subsection, we explore the sensitivity of hyper-parameters $\gamma, \epsilon, \lambda, \kappa$ which directly influence the performances of our proposed algorithms. We use datasets UCSC and Pokec with overlapping groups to conduct the experiments for ease of analysis.

Parameter γ. γ is a discount factor to take disparity as a penalty. We use $\gamma = 1$ by default because at this time the penalty of disparity is normalized by the graph size n, which is comparable with the overall influence in the whole network. But in real life, we do not exclude that γ can be assigned with another nonnegative real number. γ can be large when emphasizing absolute justice like race issue and γ can also be small when highlighting overall influence like marketing. Thus we vary γ in $\{0, 0.5, 1, 1.5\}$ to see the performances for the fair influence algorithms. The results can be seen in Figure 8. When $\gamma = 0$, the fair influence maximization problem is reduced to a vanilla IM problem. The result of ABRIS-T is much better than the other algorithms. Because in this case, the second phase of ABRIS-T reduces to the standard greedy approach for IM problem. When γ gets larger, we can find that the performance of ABRIS-T and ABRIS-G are both better than Adversarial, SET-EP and RMOIM. ABRIS-T can be adaptive to random γ when chasing for better objective function value with the second phase, while the other algorithms cannot.



(a) $\gamma = 0$ (b) $\gamma = 0.5$ (c) $\gamma = 1$ (d) $\gamma = 1.5$

(e) $\gamma = 0$ (f) $\gamma = 0.5$ (g) $\gamma = 1$ (h) $\gamma = 1.5$

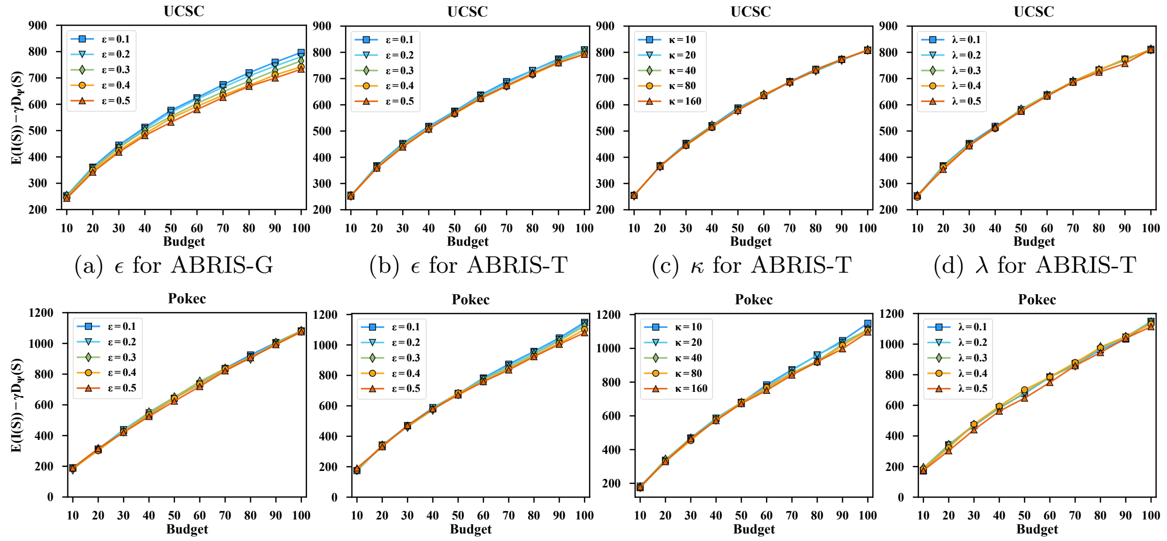
Figure 8: Comparison of the objective function of different algorithms under different budgets when parameter γ changes. γ is a parameter of the discount factor to take disparity as a penalty. The x-axis shows different sizes of seed sets. The y-axis is the objective function values under different γ settings.

Parameter ϵ. ϵ is the parameter to control the precision of influence estimation in each individual attributed group. It is adopted not only in ABRIS-G but also in ABRIS-T. Therefore in Figures 9(a), 9(e), 9(b) and 9(f), we vary the ϵs for both algorithms in

$\{0.1, 0.2, 0.3, 0.4, 0.5\}$ to show the influences on the objective function. From the figures, we can find that the influence of different ϵ for ABRIS-G in UCSC is more visible. It means that the change of parameter ϵ has a greater influence in ABRIS-G for small-size datasets. While for ABRIS-T, when the budget is small (no more than 50), the curves coincide. When the budget grows, the curve declines slightly with ϵ increasing. Though ϵ increases to 0.5, the performance is still satisfactory. Thus we can use larger ϵ to reduce computation complexity with fewer RR sets to obtain comparable results. Overall, our proposed algorithms can put up with considerable deviations for the influence estimation in most cases.

Parameter κ. κ is the parameter to control the stop condition of the ABRIS-T algorithm. Larger κ means earlier termination of the binary search procedure and less computation complexity. We vary parameter κ in range $\{10, 20, 40, 80, 160\}$ and the result is shown in Figures 9(c) and 9(g). We can see that the curve also reduces very slightly with κ getting larger. The reasons are two-fold. One is that MWU returns a similar output when the proportional factor α is within a certain interval. Therefore when ABRIS-T adopts larger κ, the curve rarely shows a sharp decline. While the other one is that the second phase can also help improve the performance of ABRIS-T to some extent.

Parameter λ. λ is the parameter for the MWU method and is also related to the precision for the first phase in the algorithm ABRIS-T. We vary λ in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and the corresponding influence can be seen in Figures 9(d) and 9(h). The curves of $\lambda = 0.1$ to $\lambda = 0.4$ nearly coincide, in which situation the influence of changing λ is small. The curve of $\lambda = 0.5$ declines slightly, which is not very visible. But as the key procedure in ABRIS-T, it's recommended to adopt relatively small λ to ensure the effectiveness of MWU.



(a) ϵ for ABRIS-G (b) ϵ for ABRIS-T (c) κ for ABRIS-T (d) λ for ABRIS-T

(e) ϵ for ABRIS-G (f) ϵ for ABRIS-T (g) κ for ABRIS-T (h) λ for ABRIS-T

Figure 9: Comparison of the objective function of different algorithms under different budgets when parameters $\epsilon, \kappa, \lambda$ change. ϵ is the parameter to control the precision of influence estimation for ABRIS-G and ABRIS-T. κ controls the stop condition of ABRIS-T. λ is the MWU parameter related to the precision of the first phase in ABRIS-T.

6. Conclusion and Discussion

Conventional influence maximization algorithms could cause severe bias for influence spread in social networks with attributed groups. To address this issue, we formulate the fair influence maximization problem and propose an attribute-based reversed influence sampling framework. Based on the solution framework, we design two novel seed node selection algorithms. One is called ABRIS-G through a basic greedy approach, and the other is called ABRIS-T by adopting a two-phase node selection method. Extensive experiments based on six real-world social network datasets show that our solutions significantly outperform the state-of-the-art approaches.

This paper does further work towards the design of efficient algorithms for the fair influence maximization problem in large-scale networks with a theoretical guarantee. Looking ahead, the following considerations concerning fair influence maximization may be helpful in future study.

- First, structural centralities would greatly help identify fair influential nodes. They are more practical and not subject to a specific diffusion model. And some metrics like degree and betweenness positively correlate with the cascade influence (Jalili & Perc, 2017; Ghanbari, Jalili, & Yu, 2018). Therefore, they can probably be reformed to measure the capability of fair influence. For example, the disparity derived from different groups can be added to degree centrality, Katz centrality (Katz, 1953) or truncated Katz centrality (Lin, Li, Song, Nguyen, Wang, & Lu, 2021) to measure the relative fair influence of a vertex.

- Second, recently graph neural networks (GNNs) have become more and more popular in dealing with NP-hard graph based problems (Ranjan, Grover, Medya, Chakravarthy, Sabharwal, & Ranu, 2022; Bai, Xu, Sun, & Wang, 2021). Even though many works have tried to propose deep learning methods (Li, Gao, Gao, Guo, & Wu, 2022; Khajehnejad et al., 2021) for influence maximization related problems, there is still great potential in solving fair influence maximization through GNN models.

- Third, besides the fair influence maximization problem discussed in this work, there are also fairness in budgets (Nguyen, Pham, Le, & Snášel, 2022), fairness of time (Ali et al., 2022) and fairness of content spread (Swift, Ebrahimi, Nova, & Asudeh, 2022) for the influence maximization problem. Therefore, more variants of fairness can be taken into consideration to meet the tangible needs.

Acknowledgments

This work was partially supported by the science and technology project from State Grid Corporation of China (No. 5700-202153172A-0-0-00). The corresponding authors are Wenzhong Li (lwz@nju.edu.cn) and Sanglu Lu (sanglu@nju.edu.cn).

References

Ali, J., Babaei, M., Chakraborty, A., Mirzasoleiman, B., Gummadi, K. P., & Singla, A. (2022). On the fairness of time-critical influence maximization in social networks. In *2022 IEEE 38th International Conference on Data Engineering (ICDE'22)*, pp. 1541–1542. IEEE.

Anwar, M. S., Saveski, M., & Roy, D. (2021). Balanced influence maximization in the presence of homophily. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM'21)*, pp. 175–183.

Bai, Y., Xu, D., Sun, Y., & Wang, W. (2021). Glsearch: Maximum common subgraph detection via learning to search. In *International Conference on Machine Learning (ICML'21)*, pp. 588–598. PMLR.

Bailey, N. T., et al. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Becker, R., D'Angelo, G., Ghobadi, S., & Gilbert, H. (2022). Fairness in influence maximization through randomization. *Journal of Artificial Intelligence Research*, *73*, 1251–1283.

Borgs, C., Brautbar, M., Chayes, J., & Lucier, B. (2014). Maximizing social influence in nearly optimal time. In *Proceedings of the 25th annual ACM-SIAM symposium on Discrete algorithms (SODA'14)*, pp. 946–957.

Chekuri, C., Vondrák, J., & Zenklusen, R. (2010). Dependent randomized rounding via exchange properties of combinatorial structures. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS'10)*, pp. 575–584. IEEE.

Chen, N. (2009). On the approximability of influence in social networks. *SIAM Journal on Discrete Mathematics*, *23*(3), 1400–1415.

Chen, W., Wang, C., & Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th international conference on Knowledge discovery and data mining (KDD'10)*, pp. 1029–1038. ACM.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. in: Kdd. *Proc of Acm Kdd*, *199-208*, 199–208.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Farnadi, G., Babaki, B., & Gendreau, M. (2020). A unifying framework for fairness-aware influence maximization. In *Companion Proceedings of the Web Conference (WWW'20)*, pp. 714–722.

Fish, B., Bashardoust, A., Boyd, D., Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2019). Gaps in information access in social networks?. In *The World Wide Web Conference (WWW'19)*, pp. 480–490.

Gershtein, S., Milo, T., & Youngmann, B. (2021). Multi-objective influence maximization. In *Proceedings of the 24th International Conference on Extending Database Technology (EDBT'21)*, pp. 145–156.

Ghanbari, R., Jalili, M., & Yu, X. (2018). Correlation of cascade failures and centrality measures in complex networks. *Future generation computer systems*, *83*, 390–400.

Goyal, A., Lu, W., & Lakshmanan, L. V. S. (2011). Celf++:optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th International Conference on World Wide Web (WWW'10)*, pp. 47–48.

Jalili, M., & Perc, M. (2017). Information cascades in complex networks. *Journal of Complex Networks*, *5*(5), 665–693.

Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific reports*, *8*(1), 1–12.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM international conference on Knowledge discovery and data mining (KDD'03)*, pp. 137–146. ACM.

Khajehnejad, M., Rezaei, A. A., Babaei, M., Hoffmann, J., Jalili, M., & Weller, A. (2021). Adversarial graph embeddings for fair influence maximization over social networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI'20)*, pp. 4306–4312.

Krause, A., McMahan, H. B., Guestrin, C., & Gupta, A. (2008). Robust submodular observation selection. *Journal of Machine Learning Research*, *9*(Dec), 2761–2801.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th international conference on Knowledge discovery and data mining (KDD'07)*, pp. 420–429. ACM.

Li, F.-H., Li, C.-T., & Shan, M.-K. (2011). Labeled influence maximization in social networks for target marketing. In *Proceedings of the 3rd Inernational Conference on Social Computing (SocialCom'11)*, pp. 560–563. IEEE.

Li, G., Chen, S., Feng, J., Tan, K.-l., & Li, W.-s. (2014). Efficient location-aware influence maximization. In *Proceedings of the 33rd international conference on Management of data (SIGMOD'14)*, pp. 87–98. ACM.

Li, Y., Gao, H., Gao, Y., Guo, J., & Wu, W. (2022). A survey on influence maximization: From an ml-based combinatorial optimization. *arXiv preprint arXiv:2211.03074*.

Li, Y., Zhang, D., & Tan, K.-L. (2015). Real-time targeted influence maximization for online advertisements. *Proceedings of the VLDB Endowment (PVLDB)*, *8*(10), 1070–1081.

Lin, M., Li, W., & Lu, S. (2020). Balanced influence maximization in attributed social network based on sampling. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM'20)*, pp. 375–383.

Lin, M., Li, W., Song, L. J., Nguyen, C.-T., Wang, X., & Lu, S. (2021). Sake: Estimating katz centrality based on sampling for large-scale social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(4), 1–21.

Liu, Q., Xiang, B., Chen, E., Xiong, H., Tang, F., & Yu, J. X. (2014). Influence maximization over large-scale social networks: A bounded linear approach. In *Proceedings of the 23rd International Conference on Information and Knowledge Management (CIKM'14)*, pp. 171–180. ACM.

Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*, pp. 234–243. Springer.

Mitzenmacher, M., & Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press.

Nguyen, B.-N. T., Pham, P. N., Le, V.-V., & Snášel, V. (2022). Influence maximization under fairness budget distribution in online social networks. *Mathematics*, *10*(22), 4185.

Nguyen, H. T., Dinh, T. N., & Thai, M. T. (2016a). Cost-aware targeted viral marketing in billion-scale networks. In *Proceedings the 35th International Conference on Computer Communications (INFOCOM'16)*, pp. 1–9. IEEE.

Nguyen, H. T., Thai, M. T., & Dinh, T. N. (2016b). Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *Proceedings of the 35th International Conference on Management of Data (SIGMOD'16)*, pp. 695–710. ACM.

Ranjan, R., Grover, S., Medya, S., Chakravarthy, V., Sabharwal, Y., & Ranu, S. (2022). Greed: A neural framework for learning graph distance functions. In *Annual Conference on Neural Information Processing Systems (NIPS'22)*.

Shah, D. (2011). Rumors in a network: Who's the culprit?. *IEEE Transactions on information theory*, *57*(8), 5163–5181.

Shah, D., & Zaman, T. (2010). Detecting sources of computer viruses in networks: theory and experiment. In *Proceedings of the ACM international conference on Measurement and modeling of computer systems (SIGMETRICS'10)*, pp. 203–214.

Swift, I. P., Ebrahimi, S., Nova, A., & Asudeh, A. (2022). Maximizing fair content spread via edge suggestion in social networks. *Proceedings of the VLDB Endowment (PVLDB)*, *15*(11), 2692–2705.

Takac, L., & Zabovsky, M. (2012). Data analysis in public social networks. In *International Scientific Conference and International Workshop Present Day Trends of Innovations*, Vol. 1.

Tang, Y., Shi, Y., & Xiao, X. (2015). Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 34th International Conference on Management of Data (SIGMOD'15)*, pp. 1539–1554. ACM.

Tang, Y., Xiao, X., & Shi, Y. (2014). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 33rd ACM SIGMOD international conference on Management of data (SIGMOD'14)*, pp. 75–86. ACM.

Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, *391*(16), 4165–4180.

Tsang, A., Wilder, B., Rice, E., Tambe, M., & Zick, Y. (2019). Group-fairness in influence maximization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*.

Udwani, R. (2021). Multiobjective maximization of monotone submodular functions with cardinality constraint. *Informs Journal on Optimization*, *3*(1), 74–88.

Vazirani, V. V. (2001). *Approximation algorithms*. Springer.

Zhang, J., Tang, J., Ma, C., Tong, H., Jing, Y., & Li, J. (2015). Panther: Fast top-k similarity search on large networks. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'15)*, pp. 1445–1454.