# A Multi-Feature Diffusion Model: Rumor Blocking in Social Networks

Jianxiong Guo, Tiantian Chen, and Weili Wu, *Senior Member, IEEE*

***Abstract*— Online social networks provide a convenient platform for the spread of rumors, which could lead to serious aftermaths such as economic losses and public panic. The classical rumor blocking problem aims to launch a set of nodes as a positive cascade to compete with misinformation in order to limit the spread of rumors. However, most of the related researches were based on a one-dimensional diffusion model. In reality, there is more than one feature associated with an object. A user's impression on this object is determined not just by one feature but by her overall evaluation of all features associated with it. Thus, the influence spread of this object can be decomposed into the spread of multiple features. Based on that, we design a multi-feature diffusion model (MF-model) in this paper and formulate a multi-feature rumor blocking (MFRB) problem on a multi-layer network structure according to this model. To solve the MFRB problem, we design a creative sampling method called Multi-Sampling, which can be applied to this multi-layer network structure. Then, we propose a Revised-IMM algorithm and obtain a satisfactory approximate solution to MFRB. Finally, we evaluate our proposed algorithm by conducting experiments on real datasets, which shows the effectiveness of our Revised-IMM and its advantage to their baseline algorithms.**

***Index Terms*— Multi-feature diffusion, rumor blocking, social networks, sampling, approximation algorithm, martingale.**

## I. Introduction

THE online social platforms, such as Facebook, Twitter, LinkedIn, and WeChat, have been growing rapidly over the last years and has been a major communication platform. There are more than 1.52 billion users active daily on Facebook and 321 million users active monthly on Twitter. Usually, these social platforms can be represented as online social networks (OSNs), which is a directed graph, including individuals and their relationships. Even that providing users with convenient information exchange, OSNs provide opportunities for rumor, namely false or negative information, to spread as well. It can cause something bad to happen and even panic. For example, in 2018, a video spread in Weibo that a bus fell down into a river from a bridge because of a car, leading to there are 15 people losing their lives. In Weibo, all the comments were unanimously pointed to that this tragic tragedy was caused by the driving mistakes of the car driver. However, after investigated by the police, this disaster was brought by a dispute between the bus driver and an unreasonable passenger. Thus, the car driver was acquitted immediately.

The influence in social networks is diffused from user to user, which can be initiated by a set of seed (initial) users. The notable study of influence diffusion can be traced back to Kempe *et al.*'s work [1] where influence maximization (IM) problem was formulated as a monotone submodular maximization problem: find a subset of users as the seed set that makes the follow-up adoptions (influence spread) maximized. They proposed two diffusion models that are accepted widely in subsequent researches, which called Independent Cascade model (IC-model) and Linear Threshold model (LT-model). Besides, they proved IM is NP-hard and implemented the Greedy algorithm [2] by Monte-Carlo (MC) simulations with $(1-1/e-\varepsilon)$-approximation. When opposite points of view, negative and positive information, from different cascades are spread at the same time on the same social network, users are more inclined to accept the information arriving on them first. Therefore, one solution of blocking rumor spread is to launch a positive cascade to compete with misinformation [3], [4]. Since the budget for positive seeds is limited, a classical rumor blocking (RB) problem is formulated, which spreads a positive cascade by selecting a positive seed set to prevent the spread of misinformation as much as possible.

The existing researches, regardless the problem about IM or RB, were based on the simple IC-model or LT-model. In other words, a piece of information that propagates through the network has only a boolean state, either good or bad. However, in the real world, the actual information diffusion is much more complicated. Let us look at an example first.

*Example 1: For a computer, the features associated with this computer are price, performance, appearance and brand. Whether a user will purchase this computer is determined by her overall evaluation of these features, for example, price is high or low, performance is good or bad and so on.*

Therefore, in this paper, we propose a multi-feature diffusion model (MF-model), which matches to the realistic scenario better. For a user, the quality of a product depends on her overall evaluation of all features associated with this product. The information diffusion is not simply one-dimensional, object by object, but multi-dimensional, feature by feature. Back to the above example, provided some company wants to promote its new computer, it will not tell others directly that this computer is very good, but tell others that its price is low, performance is satisfactory, apperance is beautiful, and so on. In our MF-model, we assume that each feature can be diffused independently. After the diffusion of each feature

Manuscript received November 8, 2019; revised June 7, 2020; accepted October 19, 2020; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Liu. Date of publication October 29, 2020; date of current version February 17, 2021. This work was supported in part by the National Science Foundation under Grant 1747818 and Grant 1907472. *(Corresponding author: Jianxiong Guo.)*

The authors are with the Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: jianxiong.guo@utdallas.edu; tiantian.chen@utdallas.edu; weiliwu@utdallas.edu).

Digital Object Identifier 10.1109/TNET.2020.3032893

terminates, users can determine whether this product is good or bad according to their own evaluation criteria. The importance of each feature is different for different users, but we assume they have the same weight vector to simplify our model.

Then, we propose a multi-feature rumor blocking (MFRB) problem, which selects a positive seed set to compete with the rumor cascade under the MF-model. The rumor from competitors is possible to spread wrong information on different features in order to lower down the reputation of the product. For example, somebody says the battery performance of iPhone is not good and its price is too expensive, or some presidential candidate's private life is extravagant. It is worth noting that although there is some negative news, this does not mean iPhone is not a good product or this presidential candidate is not qualified. The judgment for an object depends on the comprehensive evaluation of all features associated with it. Therefore, our MF-model is suitable to solve such problems. The influence spread under the MF-model can be constructed in a multi-layer network structure. We prove the objective function of MFRB problem is monotone non-decreasing and submodular. Unfortunately, computing the expected influence spread is #P-hard [5], thus the objective function is hard to compute despite the Greedy algorithm is simple and effective. To estimate the expected influence spread, they usually adopt the MC simulations, but its computational cost is not acceptable. In order to improve its efficiency, the randomized algorithms based on reverse influence sampling (RIS) popularized gradually [6]–[8]. Inspired by this idea, we propose a novel sampling method, called Multi-Sampling, which can be applied to the multi-layer network structure, and we show that this sampling method is effective to solve our MFRB problem. Then, based on Multi-Sampling and martingale analysis, the Revised-IMM (IM via martingales) is formulated, whose performance for the MFRB problem is as good as the Greedy but much more efficient than the Greedy. Besides, We can implement our Revised-IMM under the different parameter settings according to your requirements for error and running time. Finally, our proposed algorithms are evaluated on real-world datasets. The results show the Revised-IMM is much faster than the Greedy and almost get the same performance for MFRB problem.

*Organiztion:* In Section II, we survey the related works about RB and its algorithms. We then present MF-model and MFRB problem in Section III, introduce our sampling technique on multi-layer networks in Section IV, and design our randomized algorithms in Section V. Finally, we conduct experiments and conclude in Section VI and Section VII.

## II. Related Works

By spreading a positive cascade, the RB problem can be considered as a special case of competitive IM problem [9], [10]. Based on that, Budak *et al.* [3] proposed a multi-campaign IC-model and summarized RB problem as a monotone submodular maximization problem first. They proved that the objective function of RB is submodular and obtained a constant approximation ratio through Greedy algorithm. He *et al.* [11] considered the competitive LT-model for RB problem and designed a $(1-1/e)$-approximation algorithm. Fan *et al.* [12] proposed the least cost RB problem under the opportunistic one-active-one model and obtained a valid theoretical bound. Then, they [13] considered RB problem under the time constraint, constrained by a deadline $T$. Besides, In addition to spreading a positive cascade [14], there were two other technique that attempted to minimize the influence spread. One was protecting the most influential nodes from influenced by rumor cascade so that the influence of negative information can be reduced [12], [15], [16]. The other was removing some of relationships (edges) that play a central role in networks to limit the spread of misinformation [17]–[19]. Please read the Srijan's survey [20] about misinformation.

Since Kempe's seminal work [1], a large number of related researches have been done. They try to overcome the high time complexity of Greedy algorithm. It is #P-hard to compute the expected influence spread of a seed set under the IC-model [5] and LT-model [21]. The MC simulations was adopted by many researchers to estimate the expected influence spread, but the computational cost was unacceptable when applied to large networks. Subsequently, a lot of researchers attempted to overcome the low-efficiency of MC simulations [22]–[26]. For example, Leskovec *et al.* [22] proposed an CELF algorithm with a lazy-forward evaluation, which avoids unnecessary computation by estimating the upper bound of influence. CELF++ [27], an improved verson of CELF, reduced its time complexity. The effect was not satisfactory until the emergence of RIS, which was proposed firstly by Brogs *et al.* [6]. Based on that, a series of efficient randomized algorithm arised like TIM/TIM+ [7], IMM [8], and SSA/D-SSA [28]. They were scalable methods with $(1-1/e-\varepsilon)$-approximation guarantee for the IM problem. Recently, Li *et al.* [25] proposed TIPTOP based on RIS, an almost exact solutions for IM in in Billion-Scale Networks, which tried to reduces the number of samples as much as possible. Extended this RIS technique from IM to RM problem, Tong *et al.* [4] designed a unbiased estimator for the objective value of RB problem and presented an efficient randomized algorithm with $(1-1/e-\varepsilon)$-approximation. Guo *et al.* [29] created a targeted protection problem and designed a efficient heuristic algorithm by means of sampling reverse shortest path. Even though that, these problem and sampling techniques are based on one-dimensional diffusion model. Thus, how to construct a multi-dimensional diffusion model and design its sampling method is the main contribution of this paper.

## III. Problem Formulation

In this section, we introduce the MF-model, formulate the MFRB problem, and discuss its properties.

### A. Influence Model

A social network can be denoted by a directed graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of $n$ users and $E = \{e_1, e_2, \ldots, e_m\}$ is the set of $m$ directed edges. The node set (resp. edge set) of graph $G$ can be referred as $V(G)$ (resp. $E(G)$). For an edge $e = (u, v)$, $u$ (resp. $v$) is an incoming neighbor (resp. outgoing neighbor) of $v$ (resp. $u$). Then, we use $N^-(v)$ (resp. $N^+(v)$) to denote the set of incoming neighbors (resp. outgoing neighbors) of node $v$. Given a seed set $S \subseteq V(G)$, the influence diffusion from $S$ can be modelled by a discrete-time process. At time step $t_i$, we denote by $S_i$ the current active node set. We have $S_0 := S$ at $t_0$.

*Definition 1 (IC-model): Each edge $(u, v) \in E(G)$ is associated with a diffusion probability $p_{uv} \in (0, 1]$. At $t_i$ for $i \geq 1$, we have $S_i := S_{i-1}$ first; then, each new activated node $u \in (S_{t-1} \backslash S_{t-2})$ has one chance to activate its each inactive outgoing neighbor $v$ with the probability $p_{uv}$. We add $v$ into $S_i$ if $u$ activates $v$ successfully. The influence diffusion stops when no node can be activated later.*

### *B. Realization*

Given a dircted graph $G = (V, E)$, a realization $\mathrm{g} = (V, E(\mathrm{g}))$ is a subgraph of $G$ such that $E(\mathrm{g}) \subseteq E(G)$. The edges in $E(\mathrm{g})$ are referred as to live edges, otherwise, called blocked edges. Under the IC-model, for each edge $e = (u, v) \in E(G)$, it appears in realization g with probability $p_{uv}$. Let $\Pr[\mathrm{g}]$ be the probability of realization g generated from $G$ under the IC-model, we have

$$\Pr[\mathrm{g}] = \prod_{e \in E(\mathrm{g})} p_e \prod_{e \in E(G) \backslash E(\mathrm{g})} (1 - p_e) \tag{1}$$

Obviously, there are $2^m$ possible realizations in all. The diffusion process in a realization g is a deterministic process.

In classical IM problem, we usually denote by $\sigma(S)$ the expected number of active nodes (influence spread) given a seed set $S$. Under the IC-model, we have

$$\sigma(S) = \sum_{\mathrm{g} \in \mathcal{G}} \Pr[\mathrm{g}] \cdot \sigma_{\mathrm{g}}(S) \tag{2}$$

where $\mathcal{G}$ is the set of all possible realizations sampled from $G$ and $\sigma_{\mathrm{g}}(S)$ is the number of nodes that can be reached from a node in $S$ by the live edges in the realization g. Given a set function $h : 2^V \rightarrow \mathbb{R}$ and any two sets $S, T \subseteq V$, it is monotone if $h(S) \leq h(T)$ when $S \subseteq T \subseteq V$ and submodular if $h(S \cup \{u\}) - h(S) \geq h(T \cup \{u\}) - h(T)$ when $S \subseteq T \subseteq V$ and $u \notin T$. Based on that, we have

*Lemma 1 ( [1]): The objective function $\sigma(\cdot)$ is monotone non-decreasing and submodular under the IC-model.*

*Remark 1: The function $\sigma(\cdot)$ is a general notation to represent the expected influence spread, thus, every time we mention it, we need to emphasize what diffusion model we use.*

### *C. Problem Definition*

First, let us consider a scenario with composed influence under a single cascade. Considering a product with $r$ features and a directed social network $G = (V, E)$, the diffusion process can be regarded as discrete steps:

1) Each node represents a user, and there are two possible states associated with each user, active and inactive. The user is active when she is willing to purchase this product. Initially, all users are inactive.
2) Each edge $(u, v)$ is associated with a $r$-dimensional probability vector $(p^1_{uv}, p^2_{uv}, \ldots, p^r_{uv})$, where $p^i_{uv}$ represents the diffusion probability of feature $i$. When user $u$ is activated, she will attempt to motivate her inactive outgoing neighbor $v$ to accept feature $i$ with probability $p^i_{uv}$. In this activation attempt, maybe $v$ will accept one or many features.
3) If user $v$ receives influence from more than one active incoming neighbors simultaneously, $v$ will treat their features independently.
4) Each user $v$ has a threshold $\theta_v$, representing the threshold that $v$ will purchase this product, and a weight vector $(w^1_v, w^2_v, \ldots, w^r_v)$, where $w^i_v$ represents the weight of feature $i$ and $\sum_{i=1}^r w^i_v = 1$. User $v$ will be activated if and only if the total weight of accepted features is larger than or equal to $\theta_v$.
5) Initially, a seed set, containing initial users, is activated. At each step, every user checks whether the activated condition is satisfied. The process ends if no user becomes newly active at current step.

*Observation 1: According to above composed influence model, the expected influence spread $\sigma(\cdot)$ (active nodes) is not submodular.*

*Proof:* We take a counterexample to show that. Considering a product associated with five features, a user $v$ has five incoming neighbors $\{u_1, u_2, u_3, u_4, u_5\}$. For each edge $(u_i, v)$, we define $p^i_{u_i v} = 1$ and other $p^j_{u_i v} = 0$ for $j \neq i$. We assume that user $v$ has a threshold $\theta_v = 0.5$ and weight $w^i_v = 0.2$ on each feature $i$. Obviously, $\sigma(\{u_1, u_3\}) - \sigma(\{u_1\}) = 0 < \sigma(\{u_1, u_2, u_3\}) - \sigma(\{u_1, u_2\}) = 1$ and $\{u_1\} \subseteq \{u_1, u_2\}$, contradicting the property of diminishing marginal gain. Thus, $\sigma(\cdot)$ is not submodular under the composed influence model. □

Are there any techniques improving the composed influence to make the expected influence spread be submodular? We assume user $v$ will be influenced by the features of her incoming neighbor $u$ only when $u$ is activated. This condition can be relaxed. Here, each feature can be spread independently, in other words, $v$ can be influenced by the accepted features of $u$, but $u$ is inactive. Thus, we can treat this relaxed diffusion model as a multi-dimensional IC-model, which is still valid. For example, if someone tells me the appearance of iPhone is good, I will propagate this feature about appearance to my friends even though I do not know whether the iPhone's other features are good or bad. Thus, each feature diffuses in its own dimension like the diffusion of IC-model and consults with other dimensions only when making decision to purchase the product. In order to simulate the real scene better, a threshold $\theta_v$ should be distributed in interval $[0, 1]$ uniformly. In this paper, we assume that the weight for feature $i$ is equal for different users, $w^i \leftarrow w^i_u = w^i_v = \ldots = w^i_z$. So far, the multi-feature diffusion model (MF-model) is formulated:

*Definition 2 (MF-Model): Given a product with $r$ features and a directed social network $G = (V, E)$, there exists an equivalent directed multi-layer graph $G' = (V', E')$. For each feature $i \in \{1, 2, \ldots, r\}$, make a copy $G^i$ of $G$. Here, we define $u^i$ in $G^i$ is the copy of corresponding node $u$ in $G$. The new*

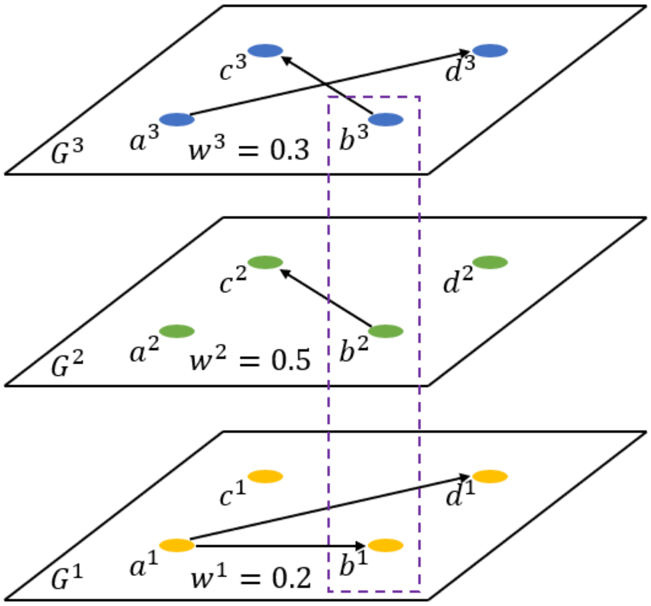


Fig. 1. The form of expression to multi-layer network structure in MF-model, where each layer represents one feature and the nodes in the same column correspond to one user.

*graph $G' = G^1 \cup G^2 \cup \ldots \cup G^r$. For each edge $(u^i, v^i)$, the diffusion probability $p_{u^i v^i}$ is equal to $p^i_{uv}$ defined in composed influence model. For each layer $G^i$, only feature $i$ is spread on it and the diffusion process in this layer is independent to other layers. After all diffusion terminate, we need to determine whether a user is activated. Here, we define $x(v^i) = 1$ if node $v^i$ accepts feature $i$, otherwise $x(v^i) = 0$. If user $v$ satisfies the following condition*:

$$\sum_{i=1}^{r} w^i \cdot x(v^i) \geq \theta_v \qquad (3)$$

*we say this user $v$ is activated. Other definition is similar to that in composed influence model.*

*Remark 2: The nodes in $V(G)$ are called user node, or user; but the nodes in $V(G')$ are called feature node, or feature. For example, a user node $u$ corresponds to feature node set $\{u^1, u^2, \ldots, u^r\}$. A user node can be activated when satisfying Equation (3). To avoid confusion, we say a feature node is accepted when it is activated in its layer.*

Then, we take an example, shown as Fig. 1, to demonstrate how MF-model works, where it is a realization of $G'$ and $\{G^1, G^2, G^3\}$ corresponds to three features. Initially, user $b$ is activated, namely features $\{b^1, b^2, b^3\}$ are accepted. After diffustion stops, for user $c$, feature $c^2$ and $c^3$ are accepted. Assuming $\theta_c = 0.5$, we have $\sum_{i=1}^{3} w^i \cdot x(c^i) = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 1 = 0.8 > \theta_c$, thus, user $c$ is activated.

Under the MF-model, there are multiple cascades diffusing on the same social network. A user is $C$-active if she is activated by cascade $C$. A feature node is $C$-accepted if it is activated by cascade $C$ in its layer. Initially, all users are $\emptyset$-active. Shown as Definition 2, each feature diffuses independently, and then we are able to determine whether the user is active after all feature diffusions have terminated. Let us consider the following scenario: there are two cascades spreading on network $G = (V, E)$, a positive cascade $C_p$ and a negative (rumor) cascade $C_r$, where rumor cascade will propagate false information on one or more features. Given the rumor seed set $S_r$, we want to launch a positive cascade to compete against the rumor cascade. Denote by $S_p$ the seed set of positive cascade, the information from $S_r$ and $S_p$ diffuses simultaneously under the MF-model. On the layer $G^i$, if two opposite cascades activate a node $v^i$ successfully at the same time, rumor cascade has a higher priority, thus $v^i$ will be $C_r$-accepted. After all feature diffusions stop, we are able to determine whether a user is $C_r$-active or $C_p$-active.

*Remark 3: For a user $u$, we define $\mathcal{F}(u) = \{u^1, u^2, \ldots, u^r\}$ as $u$'s corresponding feature nodes. Assuming that a seed set $S$ is served for cascade $C$, we say $S$ is partially $C$-active if there exists some user $u \in S$, only part of feature nodes in $\mathcal{F}(u)$ accept cascade $C$. For example, only $\{u^1, u^3\} \subset \mathcal{F}(u)$ accept cascade $C$. On the contrary, $S$ is fully $C$-active if all feature nodes of each user in $S$ accept cascade $C$. Then, we denote by $S^i$, $i \in \{1, 2, \ldots, r\}$, the set of corresponding feature nodes of $S$ in layer $i$ that accept cascade $C$. If $S$ is partially $C$-active, then $|S^i| \leq |S|$. If $S$ is fully $C$-active, then $|S^i| = |S|$.*

In the real world, a user can hardly be so stupid that she believes the rumor that announces all the features of a product are not good. Thus, we assume that rumor seed set $S_r$ is partially $C_r$-active, in other words, there exists some user $u \in S_r$ who does not believe all the features of this product are bad when rumor is this product is totally bad. And positive seed set $S_p$ is fully $C_p$-active. A user is $\bar{C}_r$-active if she is not $C_r$-active. For user $v$, we define $r(v^i) = 1$ if node $v^i$ accepts rumor cascade, otherwise $r(v^i) = 0$. If user $v$ satisfies the following condition:

$$\sum_{i=1}^{r} w^i \cdot (1 - r(v^i)) \geq \theta_v \qquad (4)$$

this user $v$ is not activated by rumor cascade. If Inequality (4) is satisfied, we say this user $v$ is $\bar{C}_r$-active. Besides, we denote by $f(S_p)$ the expected number of $\bar{C}_r$-active users given a positive seed set $S_p$. So far, the multi-feature rumor blocking (MFRB) problem is formulated.

*Problem 1 (MFRB): Given a social network $G = (V, E)$, a budget $k$ and a partially $C_r$-active rumor set $S_r$, MFRB selects an fully $C_p$-active positive seed set $S_p^\circ$, $|S_p^\circ| \leq k$, from $V(G) \backslash S_r$ to make the expected number of $\bar{C}_r$-active users $f(S_p)$ maximized under the MF-model. We have*

$$S_p^\circ = \arg \max_{S_p \subseteq V \backslash S_r, |S_p| \leq k} f(S_p) \qquad (5)$$

*Theorem 1: In MFRB problem, the expected number of $\bar{C}_r$-active users $f(S_p)$ is monotone non-decreasing and submodular with respect to $S_p$.*

*Proof:* First, we need to represent $f(S_p)$ mathmatically. The $f(S_p)$ under the MF-model can be defined as follows:

$$f(S_p) = \sum_{i=1}^{r} w^i \sum_{g^i \in \mathcal{G}^i} \Pr[g^i] \cdot f^i_{g^i}(S^i_p) \qquad (6)$$

**Algorithm 1** R-Sampling $(\mathrm{g}^i, v^i, S_r^i)$

```
Input: $\mathrm{g}^i = (V^i, E^i(\mathrm{g}))$, $v^i$ and $S_r^i$
Output: $V^*$ or $V^i$
Initialize $V_{cur} \leftarrow \{v^i\}$
Initialize $V^* \leftarrow \emptyset$
Initialize an empty queue $Q$
while true do
    if $V_{cur} = \emptyset$ then
        Return $V^i$
    end if
    if $V_{cur} \cap S_r^i \neq \emptyset$ then
        Return $V^*$
    end if
    $V^* \leftarrow V^* \cup V_{cur}$
    $Q \leftarrow Q \cup V_{cur}$
    $V_{cur} \leftarrow \emptyset$
    while $Q \neq \emptyset$ do
        $u^i = Q$.pop()
        $V_{cur} \leftarrow V_{cur} \cup \{t^i | t^i \in N^-(u^i)$ and $t^i \notin V^*\}$
    end while
end while
```

where $f^i_{\mathrm{g}^i}(S_p^i)$ is the number of feature nodes that cannot be reached by rumor cascade from $S_r^i$ in the realization $\mathrm{g}^i$ of graph $G^i$, and $S_p^i$ is the set of feature nodes in layer $i$ corresponding to users in $S_p$ according to fully active assumption of $S_p$.

Then, $\sum_{\mathrm{g}^i \in \mathcal{G}^i} \Pr[\mathrm{g}^i] \cdot f^i_{\mathrm{g}^i}(S_p^i)$ is the average number of feature nodes, which is $\bar{C}_r$-accepted in feature $i$. Becuase the threshold $\theta_v$ is uniformly distributed in $[0,1]$ and $\sum_{i=1}^r w^i = 1$, each $\bar{C}_r$-active node $u^i$ contributes $w^i$ to the expectation of $\bar{C}_r$-accepted users. In other words, the probability of user $u$ terminated as $\bar{C}_r$-active increases by $w^i$, so $f(S_p)$ increases by $w^i$. In addition, $f^i_{\mathrm{g}^i}(S_p^i)$ is monotone non-decreasing and submodular, which has been proven by Tong *et al.* [4]. $f(S_p)$ is a linear combination of $f^i_{\mathrm{g}^i}(S_p^i)$, thus, $f(S_p)$ is monotone and submodular with respect to $S_p$. □

## IV. Sampling Technique

In last section, we have proven that the MFRB problem is monotone non-decreasing and submodular, thereby the Greedy algorithm can get a $(1-1/e)$-approximation [2]. However, its computational cost is too high because computing the objective function of MFRB is #P-hard [5]. Thus, it is not advisable to compute $f(S_p)$ by MC simulations. In this section, we can find an estimator of $f(S_p)$ by some sampling techniques, and then make this estimator maximized. Here, we will get help from Random R-tuple sampling technique [4] to design our estimator. First, we define the expected $\bar{C}_r$-accepted feature nodes $f^i(S_p^i)$ in layer $i$ as

$$f^i(S_p^i) = \sum_{\mathrm{g}^i \in \mathcal{G}^i} \Pr[\mathrm{g}^i] \cdot f^i_{\mathrm{g}^i}(S_p^i) \tag{7}$$

**Algorithm 2** Single-Sampling $(G^i, S_r^i)$

```
Input: $G^i = (V^i, E^i)$ and $S_r^i$
Output: $R^i$
Select a node $v^i$ from $V^i$ uniformly.
Generate a realization $\mathrm{g}^i$ of $G^i$.
$R^i \leftarrow$ R-sampling $(\mathrm{g}^i, v^i, S_r^i)$
Return $R^i$
```

**Algorithm 3** Multi-Sampling $(G, S_r)$

```
Input: $G = (V, E)$ and $S_r$
Output: $R$
Select a node $v$ from $V^1 \cup V^2 \cup \ldots \cup V^r$ uniformly.
Confirm $v \in V^i$
Generate a realization $\mathrm{g}^i$ of $G^i$.
$R \leftarrow$ R-sampling $(\mathrm{g}^i, v, S_r^i)$
Return $R$
```

For any feature node $v^i \in V(G^i)$, we use R-tuple sampling technique [4] on graph $G^i = (V^i, E^i)$ given rumor accepted set $S_r^i$, here, we call it as R-sampling. Given $\mathrm{g}^i = (V^i, E^i(\mathrm{g}))$ as a realization of $G^i$, feature node $v^i$ and rumor accepted set $S_r^i$, the R-sampling is shown in Algorithm 1 which is a little different from the original version in [4]. The R-sampling starts from $v^i$ in $V^*$ and determine whether the incoming neighbors of the nodes in $V^*$ can be added to $V^*$ in a breadth-first searching until one of the rumor nodes in $S_r^i$ is reached or no node can be furthered reached. Then, the random R-sampling in graph $G^i$ can be generated by the following steps:

1) Select a node $v^i$ from $V(G^i)$ uniformly.
2) Generate a realization $\mathrm{g}^i$ of $G^i$.
3) Get an R-sampling $V^*$ returned by Algorithm 1, R-sampling $(\mathrm{g}^i, v^i, S_r^i)$.

This process is shown in Algorithm 2, called Single-Sampling. Intuitively, $R^i$ contains the feature nodes that could prevent $v^i$ in $\mathrm{g}^i$ from influenced by rumor set $S_r^i$ when one of them accepts positive cascade. For any positive seed set $S_p$, we define an indicator as

$$x(S_p^i, R^i) = \begin{cases} 1, & \text{if } S_p^i \cap R^i \neq \emptyset \\ 0, & otherwise \end{cases} \tag{8}$$

*Remark 4: For convenience, we can consider positive seed set as $S_p = S_p^1 \cup S_p^2 \cup \ldots \cup S_p^r$ and rumor seed set as $S_r = S_r^1 \cup S_r^2 \cup \ldots \cup S_r^r$.*

Here, it is easy to know that $x(S_p, R^i) = x(S_p^i, R^i)$ because $S_p^j \cap R^i = \emptyset$ when $i \neq j$. Under the set $S_r^i$, we generate a collection of Single-Sampling $\mathcal{R}^i = \{R_1^i, R_2^i, \ldots, R_\pi^i\}$ given the feature $i$. We define $F_{\mathcal{R}^i}(S_p^i)$, the fraction of Single-Sampling in $\mathcal{R}^i$ covered by $S_p^i$, as follows:

$$F_{\mathcal{R}^i}(S_p^i) = \frac{1}{\pi} \cdot \sum_{j=1}^{\pi} x(S_p^i, R_j^i) \tag{9}$$

*Lemma 2 ( [4]): Given $G^i = (V^i, E^i)$ and $S_r^i$ for feature $i$, we have $\mathbb{E}[n \cdot F_{\mathcal{R}^i}(S_p^i)] = f^i(S_p^i)$ for $S_p^i \subseteq V^i \backslash S_r^i$.*

So far, we have obtained an unbiased estimator for $f^i(S_p^i)$ but it cannot be applied to solve our FMRB problem directly because multiple features exist in our problem. We can consider this problem in another way. Given $G' = (V', E')$ and rumor seed set $S_r$, $V' = V^1 \cup V^2 \cup \ldots \cup V^r$, we select a feature node $v \in V'$ from these $nr$ nodes randomly. After confirming this feature node we select belongs to feature $i$, we generate a realization g$^i$ of $G^i$ and then get a R-sampling $R$ returned by Algorithm 1. This process is shown in Aglorithm 3, called Multi-Sampling. Let $\mathcal{R}$ be a collection of Multi-Samplings, $\mathcal{R} = \{R_1, R_2, \ldots, R_\theta\}$, that contains $\theta$ Multi-Samplings. We define $W_{\mathcal{R}}(S_p)$, the weighted average fraction of Multi-Samplings in $\mathcal{R}$ covered by $S_p$, as follows:

$$W_{\mathcal{R}}(S_p) = \frac{1}{\theta} \cdot \sum_{i=1}^{r} w^i \sum_{j=1}^{\theta} x(S_p^i, R_j) \tag{10}$$

*Theorem 2: Given $G = (V, E)$ and rumor seed set $S_r$, we have $\mathbb{E}[nr \cdot W_{\mathcal{R}}(S_p)] = f(S_p)$ for $S_p \subseteq V \backslash S_r$.*

*Proof:* In Algorithm 3, we select a node $v$ from $V'$ uniformly, which means that the average number of Multi-Samplings in $\mathcal{R}$ generated by a node in each feature $i$ is the same. We define the number of Multi-Samplings in $\mathcal{R}$ generated by a node in feature $i$ as $N_{\mathcal{R}}(i)$, thus, $\mathbb{E}[N_{\mathcal{R}}(i)] = \theta/r$ for $i \in \{1, 2, \cdots, r\}$. Therefore, $\mathbb{E}[F_{\mathcal{R}^i}(S_p^i)]$ can be expressed as

$$\mathbb{E}[F_{\mathcal{R}^i}(S_p^i)] = r \cdot \mathbb{E}[F_{\mathcal{R}}(S_p^i)] \tag{11}$$

According to Equation (10) and (11), for $\mathbb{E}[nr \cdot W_{\mathcal{R}}(S_p)]$, we have $\mathbb{E}[nr \cdot W_{\mathcal{R}}(S_p)] =$

$$\begin{aligned}
&= \sum_{i=1}^{r} w^i \cdot \left( nr \cdot \mathbb{E}\Big[ \frac{1}{\theta} \sum_{j=1}^{\theta} x(S_p^i, R_j) \Big] \right) \\
&= \sum_{i=1}^{r} w^i \cdot \big( nr \cdot \mathbb{E}[F_{\mathcal{R}}(S_p^i)] \big) = \sum_{i=1}^{r} w^i \cdot \big( n \cdot \mathbb{E}[F_{\mathcal{R}^i}(S_p^i)] \big) \\
&= \sum_{i=1}^{r} w^i \cdot f^i(S_p^i) = f(S_p)
\end{aligned}$$

From above, we know that $nr \cdot W_{\mathcal{R}}(S_p)$ is an unbiased estimator to $f(S_p)$. Then, the theorem is proved. □

## V. The Algorithm

Before designing our algorithm, we need to introduce martingale and its relative properties first, defined as follows:

*Definition 3 (Martingale [30]): A martingale is a sequence of random variables $Y_1, Y_2, Y_3, \cdots$, such that $\mathbb{E}[|Y_i|] < +\infty$ and $\mathbb{E}[Y_i | Y_1, Y_2, \ldots, Y_{i-1}] = Y_{i-1}$ for any $i$.*

Consider $\mathcal{R} = \{R_1, R_2, \cdots, R_\theta\}$ and $p = f(S_p)/nr$, we define $M_k$ as

$$M_k = \sum_{j=1}^{k} \left( \sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j) - p \right) \tag{12}$$

where $k = \{1, 2, \ldots, \theta\}$. Becasue of the linearity of expectation, $p = \mathbb{E}[W_{\mathcal{R}}(S_p)] = \mathbb{E}[\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j)]$, we have $\mathbb{E}[M_i] = 0$ and $\mathbb{E}[|M_i|] < +\infty$. The value of $x(S_p, R_j)$ is independent to the value from $x(S_p, R_1)$ to $x(S_p, R_{j-1})$, thus, $\mathbb{E}[M_i | M_1, M_2, \cdots, M_{i-1}] = M_{i-1}$. Therefore, $M_1, M_2, \ldots, M_\theta$ is formulated as a martingale.

*Lemma 3 ( [30]): Let $Y_1, Y_2, Y_3, \cdots$ be a martingale, such that $|Y_1| \leq a$, $|Y_j - Y_{j-1}| \leq a$ for each $j \in \{2, \cdots, i\}$ and $Var[Y_1] + \sum_{j=2}^{\theta} Var[Y_j | Y_1, Y_2, \cdots, Y_{j-1}] <= b$. Then for any $\gamma > 0$, we have*

$$\Pr[Y_i - \mathbb{E}[Y_i] \leq -\gamma] \leq \exp\left( -\frac{\gamma^2}{2b} \right) \tag{13}$$

$$\Pr[Y_i - \mathbb{E}[Y_i] \geq \gamma] \leq \exp\left( -\frac{\gamma^2}{(2/3)a\gamma + 2b} \right) \tag{14}$$

Considering the martingale $M_1, M_2, \cdots, M_\theta$, we can set $a = 1$ because $|M_1| \leq 1$ and $|M_j - M_{j-1}| \leq 1$ for each $j \in \{2, \cdots, \theta\}$. Here, we define the maximum weight $\bar{w}$ over all features as $\bar{w} = \max\{w_1, w_2, \ldots, w_r\}$. Obviously, we have $\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j) \leq \bar{w} \cdot x(S_p, R_j)$ because for each Multi-Sampling $R_j$, which can only be covered by one kind of feature nodes. If $x(S_p^y, R_j) = 1$, then we have $x(S_p^z, R_j) = 0$ for $z \in \{1, 2, \cdots, r\} \backslash \{y\}$. Based on the properties of variance and Equation (12), we can set $b = \bar{w} \cdot p\theta$ because $Var[M_1] + \sum_{j=2}^{\theta} Var[M_j | M_1, M_2, \ldots, M_{j-1}] =$

$$\begin{aligned}
&= \sum_{j=1}^{\theta} Var[\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j)] \\
&= \sum_{j=1}^{\theta} \{ \mathbb{E}[(\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j))^2] - (\mathbb{E}[\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j)])^2 \} \\
&= \sum_{j=1}^{\theta} \{ \mathbb{E}[\sum_{i=1}^{r} (w^i)^2 \cdot x(S_p^i, R_j)] - p^2 \} \qquad (15) \\
&\leq \sum_{j=1}^{\theta} \{ \mathbb{E}[\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j)] \} \cdot \bar{w} = p\theta \cdot \bar{w}
\end{aligned}$$

where the Inequality (15) holds because of the above analysis. If $x(S_p^y, R_j) = 1$, then we have $x(S_p^z, R_j) = 0$ for $z \in \{1, 2, \ldots, r\} \backslash \{y\}$. Thus, $\sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j) = w^y \cdot x(S_p^y, R_j)$, so $(w^y \cdot x(S_p^y, R_j))^2 = \sum_{i=1}^{r} (w^i)^2 \cdot x(S_p^i, R_j)$. From Equation (13) (14), we have the following for any $\varepsilon > 0$

$$\begin{aligned}
&\Pr\left[ \sum_{j=1}^{\theta} \sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j) - p\theta \leq -\varepsilon \cdot p\theta \right] \\
&\qquad \leq \exp\left( -\frac{\varepsilon^2}{2\bar{w}} \cdot p\theta \right) \qquad (16) \\
&\Pr\left[ \sum_{j=1}^{\theta} \sum_{i=1}^{r} w^i \cdot x(S_p^i, R_j) - p\theta \geq \varepsilon \cdot p\theta \right] \\
&\qquad \leq \exp\left( -\frac{\varepsilon^2}{2\bar{w} + (2/3)\varepsilon} \cdot p\theta \right) \qquad (17)
\end{aligned}$$

Borrowed from the idea of IMM algorithm [8], our solution of MFRB problem can be designed as two stages as follows: (1) Sampling Multi-Samplings: This stage generates Multi-Samplings iteratively and put them into $\mathcal{R}$ until satisfying a certain stopping condition; and (2) Node selection: This stage adopts greedy strategy to drive a size-k user set $S_p$ that covers sub-maximum weight of Multi-Samplings in $\mathcal{R}$.

### A. Node Selection

Let $\mathcal{R} = \{R_1, R_2, \ldots, R_\theta\}$ be a collection of Multi-Samplings and $W_{\mathcal{R}}(S_p)$ be the weighted average fraction of

**Algorithm 4** NodeSelection $(\mathcal{R}, k)$

```
Input: $\mathcal{R} = \{R_1, R_2, \ldots, R_\theta\}$ and $k$
Output: $\{S_p^*, W_{\mathcal{R}}(S_p^*)\}$
Initialize $S_p^* \leftarrow \emptyset$
for 1 to $k$ do
    $u = \arg\max_{u \in V \backslash S_r} (W_{\mathcal{R}}(S_p^* \cup \{u\}) - W_{\mathcal{R}}(S_p^*))$
    $S_p^* = S_p^* \cup \{u\}$
end for
Return $\{S_p^*, W_{\mathcal{R}}(S_p^*)\}$
```

**Algorithm 5** Sampling $(G, k, r, \varepsilon, \ell)$

```
Input: $G = (V, E)$, parameters $k$, $r$, $\varepsilon$ and $\ell$
Output: A collection $\mathcal{R}$
Initialize $\mathcal{R} = \emptyset$, $\mathrm{LB} = 1$, $\varepsilon' = \sqrt{2}\varepsilon$
Initialize $\mathcal{R}' = \emptyset$
$\lambda' = nr \left(2\bar{w} + \frac{2}{3}\varepsilon'\right) \left(\log \binom{n-n_r}{k} / \delta_3\right) \varepsilon'^{-2}$
$\lambda^* = 2nr\bar{w} \left(2 - \frac{1}{e}\right) \left(2 - \frac{1}{e} + \frac{\varepsilon}{3\bar{w}}\right) \left(\log \left(\binom{n-n_r}{k} \cdot 2n^\ell\right)\right) \varepsilon^{-2}$
for $i = 1$ to $\log_2(nr) - 1$ do
    $x_i = nr \cdot 2^{-i}$
    $\theta_i = \lambda'/x_i$
    while $|\mathcal{R}| \leq \theta_i$ do
        $R \leftarrow$ Multi-Sampling $(G, S_r)$
        $\mathcal{R} = \mathcal{R} \cup R$
    end while
    $\{S_i, W_{\mathcal{R}}(S_i)\} \leftarrow$ NodeSelection $(\mathcal{R}, k)$
    if $nr \cdot W_{\mathcal{R}}(S_i) \geq (1 + \varepsilon') \cdot x_i$ then
        $\mathrm{LB} = nr \cdot W_{\mathcal{R}}(S_i)/(1 + \varepsilon')$
        break
    end if
end for
$\theta \leftarrow \lambda^*/\mathrm{LB}$
while $|\mathcal{R}'| \leq \theta$ do
    $R \leftarrow$ Multi-Sampling $(G, S_r)$
    $\mathcal{R}' = \mathcal{R}' \cup R$
end while
Return $\mathcal{R}'$
```

Multi-Samplings in $\mathcal{R}$ covered by $S_p$. The node selection stage is shown in Algorithm 4. Here, we define the optimal solution as $S_p^\circ$ and optimal value as $\mathrm{OPT} = f(S_p^\circ)$. Because $W_{\mathcal{R}}(\cdot)$ is monotone non-decreasing and submodular, which guarantees that $W_{\mathcal{R}}(S_p^*)$ returned by Algorithm 4 satisfies $W_{\mathcal{R}}(S_p^*) \geq (1 - 1/e) \cdot W_{\mathcal{R}}(S_p^\circ)$.

*Lemma 4: Given rumor seed set $S_r$, $W_{\mathcal{R}}(S_p)$ is monotone non-decreasing and submodular with respect to $S_p$.*

*Proof:* First, we show $W_{\mathcal{R}}(\cdot)$ is monotone non-decreasing. For any positive seed set $S_p \subseteq V \backslash S_r$ and node $u \not\subseteq S_p \cup S_r$, we have $W_{\mathcal{R}}(S_p \cup \{u\}) - W_{\mathcal{R}}(S_p) =$

$$= \frac{1}{\theta} \cdot \sum_{i=1}^{r} w^i \sum_{j=1}^{\theta} (x(S_p^i \cup \{u^i\}, R_j) - x(S_p^i, R_j)) \tag{18}$$

It is monotone non-decreasing becuase $x(S_p^i, R_j) = 1$ implies $x(S_p^i \cup \{u^i\}, R_j) = 1$, $W_{\mathcal{R}}(S_p \cup \{u\}) - W_{\mathcal{R}}(S_p) \geq 0$. Next, we show $W_{\mathcal{R}}(\cdot)$ is submodular. Given any $S_{p1} \subseteq S_{p2} \subseteq V \backslash S_r$ and $u \not\subseteq S_{p2} \cup S_r$, it is equivalent to prove $x(S_{p1}^i \cup \{u^i\}, R_j) - x(S_{p1}^i, R_j) \geq x(S_{p2}^i \cup \{u^i\}, R_j) - x(S_{p2}^i, R_j)$ according to Equation (18). Here, we need to show that $x(S_{p1}^i \cup \{u^i\}, R_j) - x(S_{p1}^i, R_j) = 1$ whenever $x(S_{p2}^i \cup \{u^i\}, R_j) - x(S_{p2}^i, R_j) = 1$, which implies $x(S_{p2}^i \cup \{u^i\}, R_j) = 1$ and $x(S_{p2}^i, R_j) = 0$. $x(S_{p2}^i, R_j) = 0$ means that $S_{p2}^i \cup R_j = \emptyset$ and $S_{p1}^i \cup R_j = \emptyset$ because of $S_{p1} \subseteq S_{p2}$. Then, $x(S_{p2}^i \cup \{u^i\}, R_j) = 1$ means that $\{u^i\} \cup R_j \neq \emptyset$, so $x(S_{p1}^i \cup \{u^i\}, R_j) = 1$. Therefore, $x(S_{p1}^i \cup \{u^i\}, R_j) - x(S_{p1}^i, R_j) = 1$ and $W_{\mathcal{R}}(\cdot)$ is submodular, then the Lemma is proved. □

*Lemma 5: If the number of Multi-Samplings $\theta$ in $\mathcal{R}$ of Algorithm 4 satisfies that $\theta \geq \theta_1$,*

$$\theta_1 = \frac{2nr\bar{w} \cdot \log(1/\delta_1)}{\varepsilon_1^2 \cdot \mathrm{OPT}} \tag{19}$$

*then, $nr \cdot W_{\mathcal{R}}(S_p^*) \geq (1 - 1/e)(1 - \varepsilon_1) \cdot \mathrm{OPT}$ holds with at least $1 - \delta_1$ probability.*

*Lemma 6: If the number of Multi-Samplings $\theta$ in $\mathcal{R}$ of Algorithm 4 satisfies that $\theta \geq \theta_2$,*

$$\theta_2 = \frac{(2\bar{w} + \frac{2}{3}\varepsilon_2) nr \cdot \log\left(\binom{n-n_r}{k} / \delta_2\right)}{\varepsilon_2^2 \cdot \mathrm{OPT}} \tag{20}$$

*then, $nr \cdot W_{\mathcal{R}}(S_p^*) - f(S_p^*) \leq \varepsilon_2 \cdot \mathrm{OPT}$ holds with at least $1 - \delta_2$ probability, where $n_r = |S_r|$.*

*Theorem 3: Given any $\varepsilon_1 < \varepsilon$, $\varepsilon_2 = \varepsilon - (1 - 1/e) \cdot \varepsilon_1$ and $\delta_1, \delta_2 \in (0, 1)$ with $\delta_1 + \delta_2 \leq 1/n^\ell$, if the number of Multi-Samplings $\theta$ in $\mathcal{R}$ of Algorithm 4 satisfies $\theta \geq \max\{\theta_1, \theta_2\}$, it returns a $(1 - 1/e - \varepsilon)$-approximate solution of our MFRB problem with at least $1 - 1/n^\ell$ probability.*

*Proof:* By Lemma 4 and Lemma 5, they hold with $(1 - \delta_1)(1 - \delta_2) > 1 - (\delta_1 + \delta_2) \geq 1 - 1/n^\ell$ probability. Then, $f(S_p^*) \geq nr \cdot W_{\mathcal{R}}(S_p^*) - \varepsilon_2 \cdot \mathrm{OPT} \geq (1 - 1/e)(1 - \varepsilon_1) \cdot \mathrm{OPT} - \varepsilon_2 \cdot \mathrm{OPT} = (1 - 1/e - ((1 - 1/e) \cdot \varepsilon_1 + \varepsilon_2)) \cdot \mathrm{OPT} = (1 - 1/e - \varepsilon) \cdot \mathrm{OPT}$. Then the Theorem is proved. □

From Theorem 3, we need to compute $\theta \geq \max\{\theta_1, \theta_2\}$ and ensure $\mathcal{R}$ contains at least $\theta$ Multi-Samplings. In order to derive such a $\theta$, which is feasible to find the minimum $\theta$. Here, we set $\delta_1 = \delta_2 = 1/(2n^\ell)$ and $\varepsilon_1 = \varepsilon_2 = \varepsilon/(2 - 1/e)$ such that $\varepsilon_2 = \varepsilon - (1 - 1/e)\varepsilon_1$. We define $\lambda^*$ as

$$\lambda^* = \frac{2nr\bar{w}\left(2 - \frac{1}{e}\right)\left(2 - \frac{1}{e} + \frac{\varepsilon}{3\bar{w}}\right) \log\left(\binom{n-n_r}{k} \cdot 2n^\ell\right)}{\varepsilon^2} \tag{21}$$

and $\theta^* = \lambda^*/\mathrm{OPT}$. We can verify $\theta^* \geq \max\{\theta_1, \theta_2\}$ easily. However, it is difficult to compute the value of OPT in a direct manner. In the next subsection, we will find a lower bound LB of optimal value instead of OPT and determine the number of Multi-Samplings in $\mathcal{R}$ by $\lambda^*/\mathrm{LB}$.

### *B. Sampling Multi-Sampling*

In last subsection, we have obtained the approximate minimum value of $\theta$. Next, we aim to make the difference between LB and OPT as close as possible. The process of Sampling Multi-Sampling stage is shown in Algorithm 5. In iteration $i$, we generate a certain number of Multi-Samplings, put them into $\mathcal{R}$ and call Algorithm 4, then compare this result $W_{\mathcal{R}}(S_i)$ with statistical test $(1 + \varepsilon') \cdot x_i$. When the LB

```
Algorithm 6 Revised-IMM $(G, k, r, \varepsilon, \ell)$
Input: $G = (V, E)$, parameters $k$, $r$, $\varepsilon$ and $\ell$
Output: $\{S_p^*, W_{\mathcal{R}}(S_p^*)\}$
$\mathcal{R} \leftarrow$ Sampling($G$, $k$, $r$, $\varepsilon$, $\ell$)
$\{S_p^*, W_{\mathcal{R}}(S_p^*)\} \leftarrow$ NodeSelection($\mathcal{R}, k$)
Return $\{S_p^*, W_{\mathcal{R}}(S_p^*)\}$
```

is close to OPT enough, it terminates the for-loop with a high probability. Obviously, the Multi-Samplings generated by Algorithm 5 are not independent, because those Multi-Samplings generated in $i^{th}$ iteration are determined by whether the size of collection $\mathcal{R}$ in $(i-1)^{th}$ iteration is large enough to make the estimation accurate. It can be analyzed by use of martingale technique, which is shown as Lemma 7 and Lemma 8. Finally, we generate a new collection of Multi-Samplings, and we will explain why we need to do that later.

*Lemma 7: Consider the $i^{th}$ iteration in Algorithm 5, if the number of Multi-Samplings $\theta_i$ in $\mathcal{R}$ satisfies*

$$\theta_i \geq \frac{\left(2\bar{w} + \frac{2}{3}\varepsilon'\right) nr \cdot \left(\log \binom{n-n_r}{k} / \delta_3\right)}{\varepsilon'^2 \cdot x_i} \tag{22}$$

*If* $\mathrm{OPT} < x_i$*, then* $nr \cdot W_{\mathcal{R}}(S_i) < (1+\varepsilon') \cdot x_i$ *holds with at least* $1 - \delta_3$ *probability.*

*Lemma 8: Consider the $i^{th}$ iteration in Algorithm 5, if* $\mathrm{OPT} \geq x_i$*, then* $\mathrm{OPT} \geq nr \cdot W_{\mathcal{R}}(S_i)/(1+\varepsilon')$ *holds with at least* $1-\delta_3$ *probability.*

*Theorem 4: Given $\delta_3 = 1/(n^\ell \cdot \log_2(nr))$, the number of Multi-Samplings $|\mathcal{R}|$ returned by Algorithm 5 satisfies $|\mathcal{R}| \geq \theta^*$ with at least $1 - 1/n^\ell$ probability.*

*Proof:* Chen [31] pointed out this theorem cannot be obtained directly by combining Lemma 7 and Lemma 8. The Multi-Samplings generated in $i^{th}$ iteration are biased samples, because of the fact that it enters the $i^{th}$ iteration in Algorithm 5 means that the size of collection $\mathcal{R}$ in $(i-1)^{th}$ iteration cannot satisfy the termination condition with a high probability. The explanation and complete proof is in the appendix of [31], then this Theorem can be inferred from it. □

### C. Time Complexity

We can observe that the computational cost of Algorithm 5 mainly concentrates on the generation of Multi-Samplings. First, we need to analyze the time of generating a Multi-Samplings. At the high level, we use breath-first search from a feature node to visit each of its incoming neighbors until reaching a rumor node. Thus, the expected time to generate a Multi-Sampling is $\mathbb{E}[w(R)]$, where $w(R)$ denotes the number of edges in $G$ that are incoming edges to the nodes in $R$.

*Lemma 9: Consider the objective function $f^i(\cdot)$ defined by the Equation (7), we have*

$$\mathbb{E}[w(R)] = \frac{m \cdot \sum_{i=1}^r \mathrm{OPT}^i}{nr} \tag{23}$$

*where* $\mathrm{OPT}^i$ *is the optimal value of objective function* $f^i(S_p^i)$ *with* $|S_p^i| \leq k$ *and* $r$ *is the number of features.*

*Proof:* We denote by $\mathcal{H}(v^i)$ the collection of all possible Multi-Samplings for a feature node $v^i$. For any Multi-Sampling $R \in \mathcal{H}(v^i)$, we have $\mathbb{E}[w(R)] =$

$$\begin{aligned}
&= \frac{\sum_{i=1}^r \sum_{v^i \in V^i} \sum_{R \in \mathcal{H}(v^i)} \Pr[R] \cdot w(R)}{nr} \\
&= \frac{\sum_{i=1}^r \sum_{v^i \in V^i} \sum_{R \in \mathcal{H}(v^i)} \Pr[R] \cdot \sum_{(y^i, z^i) \in E^i} x(\{z^i\}, R)}{nr} \\
&= \frac{\sum_{(y^i, z^i) \in E^i} \sum_{i=1}^r \sum_{v^i \in V^i} \sum_{R \in \mathcal{H}(v^i)} \Pr[R] \cdot x(\{z^i\}, R)}{nr} \\
&= \frac{\sum_{(y^i, z^i) \in E^i} \sum_{i=1}^r f^i(\{z^i\})}{nr} \\
&\leq \frac{m \cdot \sum_{i=1}^r \mathrm{OPT}^i}{nr}
\end{aligned}$$

Then the Lemma is proved. □

*Lemma 10: Algorithm 4 runs in $O(r \cdot \sum_{R \in \mathcal{R}} |R|)$ time.*

*Proof:* The running time of Algorithm 4 can be derived directly from the Eqaution (10). □

Shown as above, the total number of Multi-Samplings generated in Algorithm 5 is $(|\mathcal{R}| + |\mathcal{R}'|)$. We denote by $i'$ the ending iteration of the for-loop, we have $|\mathcal{R}| = \lambda'/x_{i'}$ and $|\mathcal{R}'| = \lambda^*/\mathrm{LB}$ where $x_{i'} \leq \mathrm{LB} \leq \mathrm{OPT}$. The expected number of Multi-Samplings generated in Algorithm 5 can be expressed as $\mathbb{E}[|\mathcal{R}|] = O((\lambda' + \lambda^*)/\mathrm{OPT})$, thus

$$\mathbb{E}[|\mathcal{R}|] = O\left(\frac{(nr)(k+\ell)\log n}{\mathrm{OPT} \cdot \varepsilon^2}\right) \tag{24}$$

From above, we can know that the expected time of generating all Multi-Samplings in Algorithm 5 is $\mathbb{E}[\sum_{R \in \mathcal{R}} w(R)]$. Based on Theorem 3 in [8], another property of martingale [32], we have $\mathbb{E}[\sum_{R \in \mathcal{R}} w(R)] = \mathbb{E}[|\mathcal{R}|] \cdot \mathbb{E}[w(R)]$. Thus,

$$\mathbb{E}[\sum_{R \in \mathcal{R}} w(R)] = O((k+\ell) m \log n / \varepsilon^2) \tag{25}$$

due to the fact that $\sum_{i=1}^r \mathrm{OPT}^i = O(\mathrm{OPT})$. Besides, $\mathbb{E}[\sum_{R \in \mathcal{R}} |R|] \leq \mathbb{E}[\sum_{R \in \mathcal{R}} w(R)]$ because $|R| \leq w(R)$ for any $R \in \mathcal{R}$. Thus, the total running time is $O((k+\ell) mr \log n / \varepsilon^2)$. Then, we have the following theorem:

*Theorem 5: Algorithm 6 can be ganranteed to return a $(1 - 1/e - \varepsilon)$-approximate solution of our MFRB problem with at least $1 - 1/n^\ell$ probability, and runs in $O((k+\ell) mr \log n / \varepsilon^2)$ expected time.*

*Proof:* Chen [31] pointed out a direct combination of Theorem 3 and Theorem 4 is problematic. For Theorem 3, it is correct given a fixed value of $\theta$, which means that these $\theta$ Multi-Samplings are sampled from the same sample space. Theorem 4 is based on the satisfaction of Theorem 3, and it uses the same base sample from the probability space. Therefore, they gave us two workarounds to fix it. We adopt the first workaround in line 19 of Algorithm 5 that regenerates a new collection of Multi-Samplings. In line 18 of Algorithm 5, after determining the size of $\theta$, we regenerate a new collection of Multi-Samplings with the length of $\theta$ from line 19 to line 22. Then, we feed this new collections into Algorithm 4 to get the final result. It answers the question mentioned above why we need to generate a new collection of Multi-Samplings. □

TABLE I

THE STATISTICS OF THREE DATASETS

| Dataset | n | m | Type | Avg. degree |
|---|---|---|---|---|
| Netsci | 0.4K | 1.01K | directed | 5.00 |
| Wikivt | 1.0K | 3.15K | directed | 6.20 |
| Hethpt | 12.0K | 118.5K | directed | 19.8 |
| Epions | 75.9K | 508.8K | directed | 13.4 |

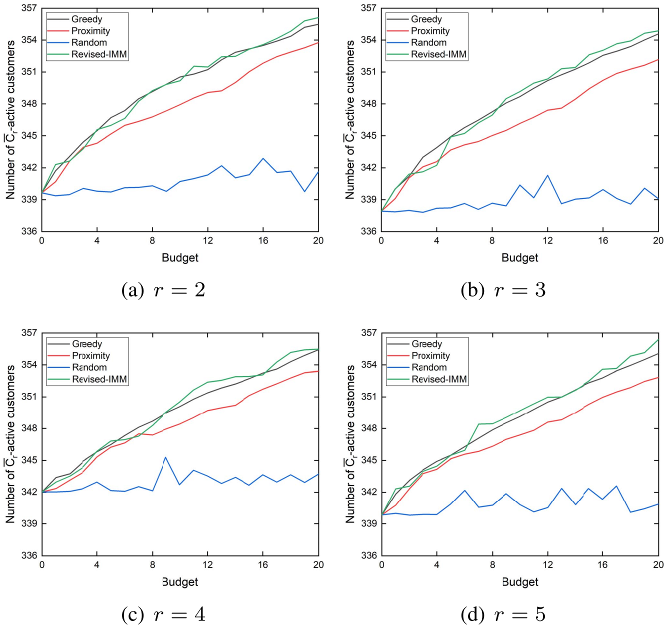
## VI. EXPERIMENT

In this section, we will show the effectiveness and efficiency of our proposed algorithms on four real social networks. There are four datasets: (1) Netsci [33]: a co-authorship network among scientists to publish papers about network science; (2) Wikivt [33]: a who-votes-on-whom network which come from the collection Wikipedia voting; (3) Hethpt [34]: an academic collaboration relationship on high energy physics area; and (4) Epions [34]: a who-trust-whom online social network on a general consumer review site Epinions.com. Basic statistics of these datasets are summarized in Table I. However, according to the multi-layer structure of MF-model, the number of feature nodes is different from this basic information, but determined by the number of features.

### A. Experimental Setup

The experiments are based on MF-model, thus for each edge $e = (u, v)$, we set $p_e = 1/|N^{-}(v)|$, which is is widely used in prior works [7], [27], [35]. The Revised-IMM algorithm is defaulted by $\varepsilon = 0.2$ and $1/n^{\ell} = 0.1$, then we compare it with some common baseline algorithms. They are shown as follows: (1) Greedy: it selects a node such that adding this node to current seed set can obtain the maximum marginal gain at each iteration, which is implemented by MC simulations with 200 times for each targeted set; (2) Proximity: it selects the outgoing neighbors of the nodes in rumor set according to the out-degree of these outgoing neighbors, where these neighbors with high out-degree are in priority; and (3) Random: this is a classical baseline algorithm, where the nodes in the positive set are selected randomly.

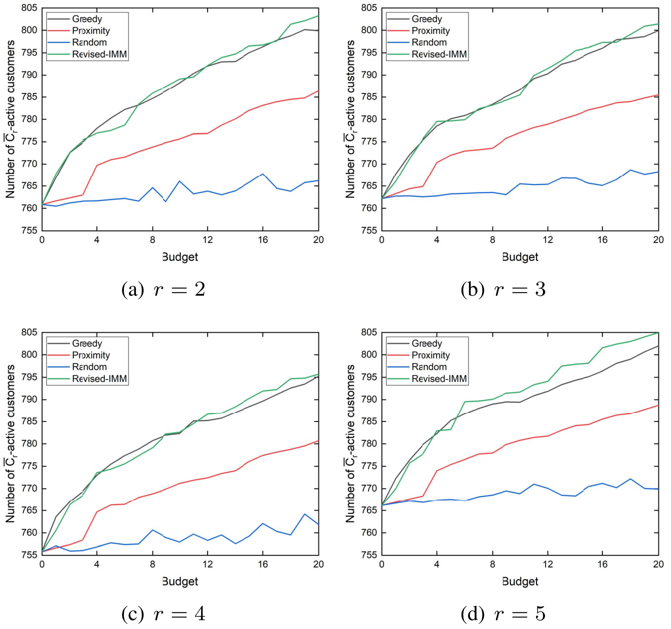
In our experiment, the users in rumor seed set $S_r$ are the nodes with the highest outgoing degree in original graph $G$ and the size $|S_r| = 20$. Because the $S_r$ is partially $C_r$-active, only part of features of those users in $S_r$ are $C_r$-accepted, thus, we set the probability that the corresponding feature nodes of $S_r$ accept rumor cascade is 80%. The number of users in positive set $S_p$ is from 1 to 20, and $S_p$ is fully $C_p$-active, so the corresponding feature nodes of $S_p$ are all $C_p$-accepted. Assume there are $r$ features given a product, we denote by $\boldsymbol{w} = (w^1, w^2, \cdots, w^r)$ its weight vector for features. Next, we evaluate the performance of Revised-IMM algorithm. It can be divided into four sub-cases: (1) there are two features where $w = (0.3, 0.7)$; (2) there are three features, where $w = (0.3, 0.3, 0.4)$; (3) there are four features $w = (0.2, 0.3, 0.4, 0.1)$; and (4) there are five features $w = (0.2, 0.1, 0.3, 0.1, 0.3)$. The actual number of nodes and edges in the corresponding graph $G'$ depends on the number of features. For example, they will be doubled when there are two features, and tripled when three features.



(a) $r = 2$ (b) $r = 3$

(c) $r = 4$ (d) $r = 5$

Fig. 2. The performance comparison achieved by different algorithms with the different number of features in Netsci dataset.



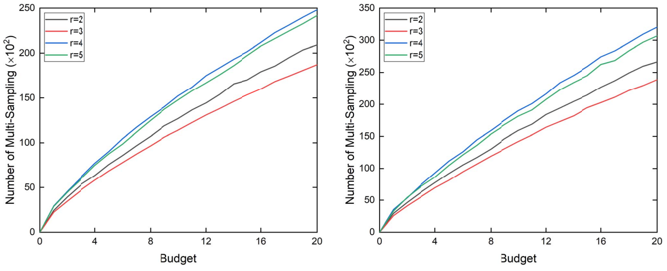
(a) $r = 2$ (b) $r = 3$

(c) $r = 4$ (d) $r = 5$

Fig. 3. The performance comparison achieved by different algorithms with the different number of features in Wikivt dataset.

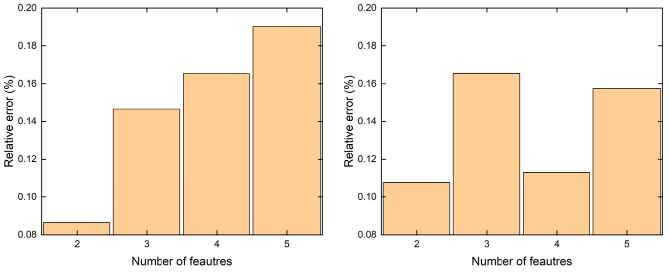
### B. Experimental Results

*1) Performance:* Fig. 2 and Fig. 3 draw the performance comparison achieved by different algorithms with the different number of features under the Netsci and Wikivt datasets. In this part, we only consider these two smaller datasets since the Greedy algorithm implemented by MC simulations is very inefficient which cannot be used in larger networks. Obviously, we can see that Revised-IMM and Greedy algorithm almost



(a) Netsci (b) Wikivt

Fig. 4. The number of Multi-Samplings generated by Algorithm 5 in Revised-IMM algorithm under the two datasets.



(a) Netsci (b) Wikivt

Fig. 5. The average relative error between the estimated value and objective value under the two datasets.

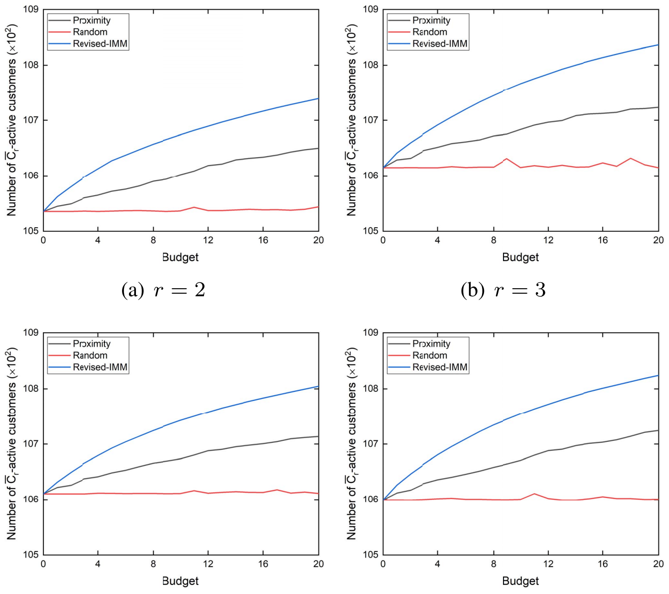
TABLE II

THE RUNNING TIME COMPARISON FOR $k = 20$, WHERE $s = second$, $m = minite$, AND $h = hour$

| Dataset | Method | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ |
|---|---|---|---|---|---|
| Netsci | Greedy | 7.56m | 11.3m | 14.4m | 17.6m |
| | R-IMM | 3.39s | 3.18h | 3.22s | 2.83s |
| Wikivt | Greedy | 1.16h | 1.68h | 3.01h | 3.61h |
| | R-IMM | 13.4s | 12.8h | 17.9s | 12.8s |
| Hethpt | R-IMM | 2.27m | 1.67m | 2.88m | 2.60m |
| Epions | R-IMM | 36.2m | 31.5m | 38.3m | 37.8m |

have the same performance, where the objective value of Revised-IMM is estimated by $nr \cdot W_{\mathcal{R}}(S_p)$. Besides, the performance of Revised-IMM is better than Proximity and Random algorithms, which proves its effectiveness.

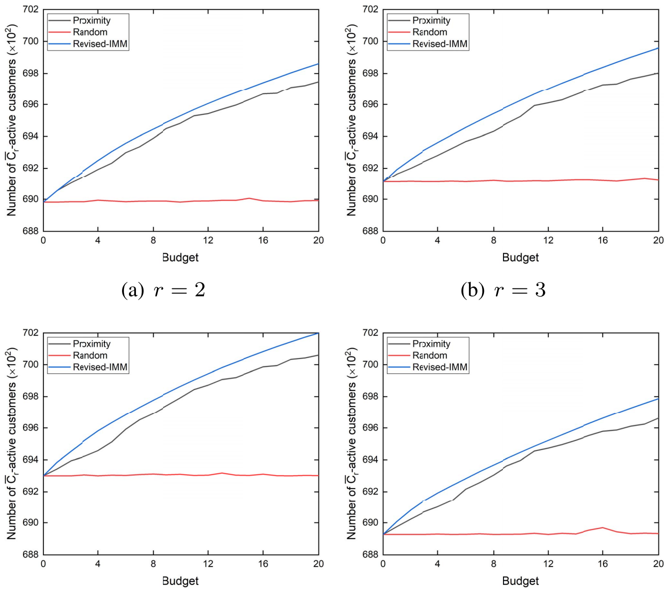
*2) Sampling and error analysis:* Fig. 4 draws the number of Multi-Samplings generated by Algorithm 5 in Revised-IMM algorithm. We can see that this is in line with our expectation, the number of Multi-Samplings increases as the budget increases. It is related to the weight vector we defined as well. Show as Fig. 4, the number of Multi-Sampling at $r = 2$ is larger than that at $r = 3$. This is because the $\bar{w} = 0.7$ at $r = 2$, which is larger than $\bar{w} = 0.4$ at $r = 3$. Fig. 5 draws the average relative error between the estimated value and objective value given a positive seed set. Consider a estimated value $nr \cdot W_{\mathcal{R}}(S_p^*)$ returned by Revised-IMM, the objective value is defined as $f(S_p^*)$ implemented by MC simulations with 2000 times. Thus, the relative error is $|f(S_p) - nr \cdot W_{\mathcal{R}}(S_p)|/f(S_p)$. Shown as Fig. 5, the average relative error is less than 0.2% generally, which proves the correctness of our estimator further.



(a) $r = 2$ (b) $r = 3$

(c) $r = 4$ (d) $r = 5$

Fig. 6. The performance comparison achieved by different algorithms with different number of features in Hethpt dataset.



(a) $r = 2$ (b) $r = 3$

(c) $r = 4$ (d) $r = 5$

Fig. 7. The performance comparison achieved by different algorithms with different number of features in Epions dataset.

*3) Scalability:* Fig. 6 and Fig. 7 draw the performance comparison under the Hethpt and Epions datasets. Here, we invoke Revised-IMM algorithm with $k = 20$ only once and print out the intermediate estimated values at the process of node selection, thus these curves are very smooth. The results from Fig. 2 and Fig. 3 are by invoking Revised-IMM with different $k$ one by one because we need to record the number of Multi-Samplings. Shown as Fig. 6 and Fig. 7, the Revised-IMM has a good performance as well in these two larger networks. Moreover, the running time comparison in these experiments is shown as Table II. Compared to Greedy algorithm, our Revised-IMM reduced the running time significantly without losing approximation, which proves the efficiency and scalability of our proposed algorithm.

## VII. Conclusion

In this paper, we propose a MF-model to simulate a real scenario in which multiple features can be propagated independently in social networks. Based on MF-model, the MFRB problem is formulated as a monotone non-decreasing submodular maximization problem. Then, we design a novel Multi-Sampling, which is an unbiased estimator to objective function of MFRB. Inspired by martingale analysis, the Revised-IMM algorithm is proposed, which returns a $(1-1/e-\varepsilon)$-approximation solution and runs in $O((k+\ell)mr\log n/\varepsilon^2)$ expected time. The experimental result verified the effectiveness and correctness of our Revised-IMM algorithm.

However, one of the shortcomings of this paper is that the weight for each feature is equal for different users, which is not entirely realistic. Actually, we now verify that the objectives still monotone submodular when the weight vector is different for different users. Our methods can be extended to such a more generalized case by simple mathematical treatments.

## Appendix

In this section, we provide the detialed proofs for the Lemmas which are neglected in the main paper.

### A. Proof of Lemma 4

*Proof:* For optimal solution $S_p^\circ$, we have defined $p=f(S_p^\circ)/nr$, thus, $p=\mathrm{OPT}/nr=\mathbb{E}[W_{\mathcal{R}}(S_p^\circ)]$. Then, by Equation (16), we have $\Pr[nr\cdot W_{\mathcal{R}}(S_p^\circ)\leq(1-\varepsilon_1)\cdot\mathrm{OPT}]=\Pr[nr\cdot W_{\mathcal{R}}(S_p^\circ)\leq(1-\varepsilon_1)\cdot pnr]=\Pr[\sum_{j=1}^{\theta}\sum_{i=1}^{r}w^i\cdot x(S_p^i,R_j)-p\theta\leq-\varepsilon_1\cdot p\theta]\leq\exp(-\frac{\varepsilon_1^2}{2\bar{w}}\cdot p\theta)\leq\exp(-\frac{\varepsilon_1^2}{2\bar{w}}\cdot p\theta_1)=\delta_1$.

Thus, $nr\cdot W_{\mathcal{R}}(S_p^\circ)\geq(1-\varepsilon_1)\cdot\mathrm{OPT}$ holds with at least $1-\delta_1$ probability. By Lemma 3 and greedy properties, $nr\cdot W_{\mathcal{R}}(S_p^*)\geq(1-1/e)\cdot nr\cdot W_{\mathcal{R}}(S_p^\circ)\geq(1-1/e)(1-\varepsilon_1)\cdot\mathrm{OPT}$. Then the Lemma is proved. □

### B. Proof of Lemma 5

*Proof:* For any $k$-size seed set $S_p$, we have defined $p=f(S_p)/nr$, thus, $p=\mathbb{E}[W_{\mathcal{R}}(S_p)]$. Then, by Equation (17) and $\zeta=\varepsilon_2\cdot\mathrm{OPT}/pnr$, we have $\Pr[nr\cdot W_{\mathcal{R}}(S_p^*)-f(S_p)\geq\varepsilon_2\cdot\mathrm{OPT}]=\Pr[nr\cdot W_{\mathcal{R}}(S_p)-pnr\geq\varepsilon_2\cdot\mathrm{OPT}]=\Pr[\theta\cdot W_{\mathcal{R}}(S_p)-p\theta\geq\frac{\varepsilon_2\cdot\mathrm{OPT}}{pnr}\cdot p\theta]=\Pr[\sum_{j=1}^{\theta}\sum_{i=1}^{r}w^i\cdot x(S_p^i,R_j)-p\theta\geq\frac{\varepsilon_2\cdot\mathrm{OPT}}{pnr}\cdot p\theta]\leq\exp(-\frac{\zeta^2}{2\bar{w}+(2/3)\zeta}\cdot p\theta)=\exp(-\frac{\varepsilon_2^2\cdot\mathrm{OPT}^2}{2\bar{w}pn^2r^2+(2/3)\varepsilon_2 nr\cdot\mathrm{OPT}}\cdot\theta)\leq\exp(-\frac{\varepsilon_2^2\cdot\mathrm{OPT}^2}{2\bar{w}nr\cdot\mathrm{OPT}+(2/3)\varepsilon_2 nr\cdot\mathrm{OPT}}\cdot\theta)\leq\exp(-\frac{\varepsilon_2^2\cdot\mathrm{OPT}}{(2\bar{w}+(2/3)\varepsilon_2)\cdot nr}\cdot\theta_2)=\delta_2/\binom{n-n_r}{k}$.

Because there exists at most $\binom{n-n_r}{k}$ positive size-$k$ seed sets and by union bound, there is at least $1-\delta_2$ probability that no such $S_p^*$ satisfies $nr\cdot W_{\mathcal{R}}(S_p^*)-f(S_p^*)\geq\varepsilon_2\cdot\mathrm{OPT}$. Then the Lemma is proved. □

### C. Proof of Lemma 6

*Proof:* For any $k$-size seed set $S_i$, we have defined $p=f(S_i)/nr$, thus, $p=\mathbb{E}[W_{\mathcal{R}}(S_i)]\leq\mathrm{OPT}/nr<x_i/nr$. Then, by Equation (17) and $\zeta=\frac{(1-\varepsilon')\cdot x_i}{pnr}-1$, we know that $\zeta>\varepsilon'\cdot x_i/(pnr)>\varepsilon'$, and we have $\Pr[nr\cdot W_{\mathcal{R}}(S_i)\geq(1+\varepsilon')\cdot x_i]=\Pr[\theta_i\cdot W_{\mathcal{R}}(S_i)-p\theta_i\geq(\frac{(1-\varepsilon')\cdot x_i}{pnr}-1)\cdot p\theta_i]\leq\exp(-\frac{\zeta^2}{2\bar{w}+(2/3)\zeta}\cdot p\theta_i)<\exp(-\frac{\varepsilon'^2\cdot x_i/(pnr)}{2\bar{w}+(2/3)\zeta}\cdot\frac{(2\bar{w}+(2/3)\varepsilon')pnr(\log\binom{n-n_r}{k}/\delta_3)}{\varepsilon'^2\cdot x_i})<\exp(-\log\binom{n-n_r}{k}/\delta_3)=\delta_3/\binom{n-n_r}{k}$.

Because there is at least $1-\delta_3$ probability by union bound that no such $S_i$ satisfies $nr\cdot W_{\mathcal{R}}(S_i)\geq(1+\varepsilon')\cdot x_i$. Then the Lemma is proved. □

### D. Proof of Lemma 7

*Proof:* For any $k$-size seed set $S_i$, we have defined $p=f(S_i)/nr$, thus, $p=\mathbb{E}[W_{\mathcal{R}}(S_i)]\leq\mathrm{OPT}/nr$. Then, by Equation (17) and $\zeta=\frac{\varepsilon'\cdot\mathrm{OPT}}{pnr}$, we have $\Pr[\mathrm{OPT}<nr\cdot W_{\mathcal{R}}(S_i)/(1+\varepsilon')]=\Pr[nr\cdot W_{\mathcal{R}}(S_i)-\mathrm{OPT}\geq\varepsilon'\cdot\mathrm{OPT}]<\Pr[\theta_i\cdot W_{\mathcal{R}}(S_i)-p\theta_i\geq\frac{\varepsilon'\cdot\mathrm{OPT}}{pnr}\cdot p\theta_i]\leq\exp(-\frac{\zeta^2}{2\bar{w}+(2/3)\zeta}\cdot p\theta_i)=\exp(-\frac{\varepsilon'^2\cdot\mathrm{OPT}^2}{2\bar{w}pn^2r^2+(2/3)\varepsilon' nr\cdot\mathrm{OPT}}\cdot\theta_i)=\exp(-\frac{\varepsilon'^2\cdot\mathrm{OPT}^2}{2\bar{w}pnr\cdot\mathrm{OPT}+(2/3)\varepsilon' nr\cdot\mathrm{OPT}}\cdot\theta_i)\leq\exp(-\frac{\varepsilon'^2\cdot\mathrm{OPT}}{(2\bar{w}+(2/3)\varepsilon')\cdot nr}\cdot\theta_i)=\delta_3/\binom{n-n_r}{k}$.

Because there is at least $1-\delta_3$ probability by union bound that no such $S_i$ satisfies $\mathrm{OPT}<nr\cdot W_{\mathcal{R}}(S_i)/(1+\varepsilon')$. Then the Lemma is proved. □

## References

[1] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2003, pp. 137–146.

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

[3] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the spread of misinformation in social networks," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, 2011, pp. 665–674.

[4] G. A. Tong *et al.*, "An efficient randomized algorithm for rumor blocking in online social networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[5] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 1029–1038.

[6] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. 25th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2014, pp. 946–957.

[7] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2014, pp. 75–86.

[8] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2015, pp. 1539–1554.

[9] S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in *Proc. Int. Workshop Web Internet Econ.* Berlin, Germany: Springer, 2007, pp. 306–311.

[10] W. Lu, W. Chen, and L. V. S. Lakshmanan, "From competition to complementarity: Comparative influence diffusion and maximization," *Proc. VLDB Endowment*, vol. 9, no. 2, pp. 60–71, Oct. 2015.

[11] X. He, G. Song, W. Chen, and Q. Jiang, "Influence blocking maximization in social networks under the competitive linear threshold model," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2012, pp. 463–474.

[12] L. Fan, Z. Lu, W. Wu, B. Thuraisingham, H. Ma, and Y. Bi, "Least cost rumor blocking in social networks," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, Jul. 2013, pp. 540–549.

[13] L. Fan, W. Wu, X. Zhai, K. Xing, W. Lee, and D.-Z. Du, "Maximizing rumor containment in social networks with constrained time," *Social Netw. Anal. Mining*, vol. 4, no. 1, p. 214, Dec. 2014.

[14] T. Chen, W. Liu, Q. Fang, J. Guo, and D.-Z. Du, "Minimizing misinformation profit in social networks," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 6, pp. 1206–1218, Dec. 2019.

[15] L.-L. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, "Identifying influential spreaders in complex networks based on gravity formula," *Phys. A, Stat. Mech. Appl.*, vol. 451, pp. 205–212, Jun. 2016.

[16] S. Wang, X. Zhao, Y. Chen, Z. Li, K. Zhang, and J. Xia, "Negative influence minimizing by blocking nodes in social networks," in *Proc. Workshops 27th AAAI Conf. Artif. Intell.*, 2013, pp. 134–136.

[17] E. B. Khalil, B. Dilkina, and L. Song, "Scalable diffusion-aware optimization of network topology," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 1226–1235.

[18] M. Kimura, K. Saito, and H. Motoda, "Minimizing the spread of contamination by blocking links in a network," in *Proc. AAAI*, vol. 8, 2008, pp. 1175–1180.

[19] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos, "Gelling, and melting, large graphs by edge manipulation," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 245–254.

[20] S. Kumar and N. Shah, "False information on Web and social media: A survey," 2018, *arXiv:1804.08559*. [Online]. Available: http://arxiv.org/abs/1804.08559

[21] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 88–97.

[22] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 420–429.

[23] J. Guo and W. Wu, "A novel scene of viral marketing for complementary products," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 4, pp. 797–808, Aug. 2019.

[24] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "A billion-scale approximation algorithm for maximizing benefit in viral marketing," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2419–2429, Aug. 2017.

[25] X. Li, J. D. Smith, T. N. Dinh, and M. T. Thai, "TipTop: (Almost) exact solutions for influence maximization in billion-scale networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 649–661, Apr. 2019.

[26] J. Guo and W. Wu, "Influence maximization: Seeding based on community structure," *ACM Trans. Knowl. Discovery From Data*, vol. 14, no. 6, pp. 66:1–66:22, 2020.

[27] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web (WWW)*, 2011, pp. 47–48.

[28] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2016, pp. 695–710.

[29] J. Guo, Y. Li, and W. Wu, "Targeted protection maximization in social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1645–1655, Jul. 2020.

[30] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Math.*, vol. 3, no. 1, pp. 79–127, Jan. 2006.

[31] W. Chen, "An issue in the martingale analysis of the influence maximization algorithm imm," in *Proc. Int. Conf. Comput. Social Netw.* Cham, Switzerland: Springer, 2018, pp. 286–297.

[32] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[33] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 4292–4293.

[34] J. Leskovec and A. Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. Accessed: Jun. 2014. [Online]. Available: http://snap.stanford.edu/data

[35] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 918–923.

**Jianxiong Guo** received the B.S. degree in energy engineering and automation from the South China University of Technology in 2015, and the M.S. degree in chemical engineering from the University of Pittsburgh in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas. His research interests include social networks, data mining, the IoT application, blockchain, and combinatorial optimization.

**Tiantian Chen** received the B.S. and M.S. degrees from the Ocean University of China in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas, under the supervision of D.-Z. Du and W. Wu. Her research interests include design and analysis of approximation algorithms and social networks.

**Weili Wu** (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the Department of Computer Science, University of Minnesota, Minneapolis, MN, USA, in 2002 and 1998, respectively. She is currently a Full Professor with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. Her research mainly deals with the general research areas of data communication and data management. Her research interest includes the design and analysis of algorithms for optimization problems that occur in wireless networking environments and various database systems.