Marginal Gains to Maximize Content Spread in Social Networks

Wenguo Yang, Jianmin Ma, Yi Li, Ruidong Yan, Jing Yuan, Weili Wu, and Deying Li

***Abstract*— The growing importance of social network for sharing and spreading various contents is leading to the changes in the way of information diffusion. To what extent can social content be diffused highly depends on the size of seed nodes and connectivity of the network. If the seed set is predetermined, then the best way to maximize the content spread is to add connectivities among the users. The existing work shows the content spread maximization problem to be NP-hard. One of the difficulties of designing an effective and efficient algorithm for the content spread maximization problem lies in that the objective function we aim to maximize lacks submodularity. In our work, we formulate the maximize content spread problem from an incremental marginal gain perspective. Although the objective function we derive is not submodular, both submodular lower and upper bounds are constructed and proved. Therefore, we apply the sandwich framework and devise a marginal increment-based algorithm (MIS) that guarantees a data-dependent factor. Furthermore, a novel scalable content spread maximization algorithm influence ranking and fast adjustment (IRFA), which is based on the influence ranking of a single node and fast adjustment with each boosting step in the network, is proposed. Through extensive experiments, we demonstrate that both MIS and IRFA algorithms are effective and outperform other edge selection strategies.**

***Index Terms*— Approximation factor, content spread, information diffusion, nonsubmodularity, social network.**

I. Introduction

With the rapid growth of social media and the rise in popularity of social networks, content sharing and spreading become the major activities for the social media users. The fact shows that there are 3.03 billion active social media users out of 3.5 billion Internet users [1]. The massive proportions of the Internet users are engaging in generating, searching, and spreading various social contents such as photos, videos, comments, reviews, news, advertisements, and so on so forth. Spreading these contents can help with recommendation system to amplify items' influence, viral marketing to broadcast products or services, and users to maximize their influence. In general, to what extent a social network spreads content is a key metric that impacts both user engagement and network revenue [2].

Ideally, users recursively share the contents with their neighbors will quickly reach and influence a large number of users on the network. However, sometimes, content spread efficiency is not what we expected. Cha *et al.* [3] observe the dynamics of photos spread on Flickr for 104 consecutive days and discover even most popular photos tend to influence users within two-hop neighbors on the network then burnout quickly. On the other hand, the breadth and depth of information dissemination are highly related to the initial seed sets. For example, in viral marketing, companies would like to spread their products and services to reach target users with the minimum startup costs. Most recent research studies focus on the influence maximization problem of selecting initial seeds [4], [9], [11], [32], [33]. However, there are some cases in which the seed users are predetermined due to the limit costs or companies' preferences. In this case, how we can help with content spread becomes a problem. Furthermore, even when the seed set is not fixed, since the seed selection problem is NP-hard [5] and the set of selected seeds is often suboptimal. Then, it comes the question of how to boost spread social content efficiently with fixed seed users. The classic influence maximization problem discusses how to select the appropriate seed set, making the ultimate influence spread as large as possible. The issue of influence boosting focuses on how to adopt other effective measures, such as increasing edge connections, increasing the propagation probability of some boosting nodes, and so on, to further promote the influence spread for the initial given social network and the given seed set. Both the influence maximization and the influence boosting are NP-hard. However, the objective function of the former is not submodular, while the latter is not, which makes it more difficult to solve the problem of influence boosting.

One of the effective measures is to increase the number of connected edges between the users. In some social network sites, such as Facebook and Twitter [6]–[8], they provide the friending services that recommend friends to you to make possible connections. These services based on the number of common friends, common interests, posted contents, similar communities, and other personal related features. However, due to some privacy issues, those information may not be

Manuscript received September 24, 2018; revised February 8, 2019 and March 29, 2019; accepted April 7, 2019. Date of publication May 6, 2019; date of current version June 10, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 11571015, Grant 11571091, and Grant 11671400 and in part by the National Science Foundation under Grant 1747818. *(Corresponding author: Yi Li.)*

W. Yang is with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangwg@ucas.ac.cn).

J. Ma is with the College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, China (e-mail: jianminma@yahoo.com).

Y. Li is with the Department of Computer Science, The University of Texas at Tyler, Tyler, TX 75799 USA (e-mail: yli@uttyler.edu).

R. Yan and D. Li are with the School of Information, Renmin University of China, Beijing 100872, China (e-mail: yanruidong@ruc.edu.cn; deyingli@ruc.edu.cn).

J. Yuan and W. Wu are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: jing.yuan@utdallas.edu; weiliwu@utdallas.edu).

Digital Object Identifier 10.1109/TCSS.2019.2911865

available to us. Also, as shown in [2], these methods cannot guarantee to maximize the content spread on social network.

Chaoji *et al.* [2] formulate the problem of boosting content spread on social network by seeking to add up to k connections per user such that the probabilistic propagation of content in the social network is maximized. Since the content maximizing problem is NP-hard and the content spread function is not submodular, they construct a more restricted variant that is submodular and devise an approximation algorithm that computes an edge set which satisfies constraints. However, their content spread function under independent cascade (IC) and restricted maximum probability path (RMPP) model has a few limitations. First, computing the spread of specific content C with any given seed set is #P-hard. It leads to substantial computation time for running expensive simulations. They derive a RMPP model from a heuristic method, which first proposed by Chen *et al.* [9]. This model restricts influence propagation to be only along a maximum probability path between a pair of nodes [9]. Based on the restriction of maximum probability paths, the RMPP model further restricts information propagation paths to contain at most one newly added edge from a candidate edge set. This setting helps the spread function under the RMPP model to be submodular but the content spread problem is still NP-hard. Second, with these restrictions, the information propagation may not reflect the real information flow on the network and the calculation of content spread for each node under the RMPP model will have a large deviation from the actual spread value. In addition, their model assumes that a predefined number of new links should be added for each user in the network, thus leading to all the users in the network to accept the same number of recommended connections. This assumption does not necessarily reflect the power law property of real-world social network.

In this paper, we formulate the content spread problem from a marginal incremental perspective to calculate the diffusion process on social network and describe the content spread function as accurate as possible. We propose a generalized content spread maximization problem that selects at most K edges from a candidate edge set and add these selected edges on network such that the information propagation on the network will be maximized. We obtain the marginal gain of content spread at each node by adding one edge from candidate edge set at a time. Our problem setting does not have the restriction on the propagation paths and can compute the spread value accurately. Although the content spread function obtained is not submodular, both submodular lower bound and upper bound are constructed, thus make the sandwich framework [23] lend itself well to the problem. A marginal increment-based algorithm (MIS) that guarantees a data-dependent approximation factor is proposed. We summarize our major contribution as follows.

1) Content spread maximization problem is formulated in a marginal gain incremental way with nearly no loss of the content spread.
2) The non-submodularity of the content spread function is given and both submodular lower bound and upper bound of the original content spread function are presented. A marginal increment-based sandwich algorithm (MIS) that guarantees a data-dependent approximation factor is devised.
3) A novel heuristic scalable algorithm of boosting content spread in social networks influence ranking and fast adjustment (IRFA) is proposed.
4) We conduct experiments on four data sets varying different parameters and show the effectiveness of our algorithm.

The rest of this paper is organized as follows. In Section II, we list some of the related works in maximizing content spread and influence propagation. In Section III, we propose our formulation of the content spread maximization problem and illustrate the nonsubmodularity. Then, we derive the submodular lower and upper bounds and a marginal increment-based algorithm (MIS) in Section IV. In Section V, we present a novel heuristic scalable algorithm (IRFA). Our experiment settings and results confirm the effectiveness of the proposed MIS and IRFA algorithms in Section VI and concluding remark and potential future works are given in Section VII.

II. Related Work

In this section, backgrounds and related works about content spread maximization problems are provided.

A. Classical Influence Maximization Problems

There have been abundant studies on various models and computational methods for influence maximization. Kempe *et al.* [11] first formulate the influence maximization problem that asks to find a set S of k nodes so that the expected influence spread is maximized under a predetermined influence propagation model. The problem is NP-hard under both IC and linear threshold (LT) models. Chen *et al.* [9], [10] show that to compute the expected influence spread for a given set is #P-hard. However, it can be formulated as a submodular and monotone function of S for both IC and LT models that can use a simple greedy algorithm [11]. Besides studying the influence maximization problem purely online, Shi *et al.* [32] considers the cyber-physical interactions and studies a location-driven influence maximization problem. There have been several studies on approximating the influence maximization problem under different diffusion models. However, in this paper, we think of the influence maximization problem in a different perspective. Instead of choosing the initial seed set, we aim to add edges to maximize the content spread.

B. Content Spread Optimization

There are some works that attempt to solve content spread optimization problem by increasing the connectivity of network through adding/deleting new edges [2], [16]–[20], [34]. Yan *et al.* [34] study the problem of minimizing the rumors spread via link deletion. While our work focus on maximizing the content spread. For maximization problems, Chaoji *et al.* [2] first attempt to integrate boosting content spread with friend recommendation. They define the content maximization problem to find at most k edges in candidate set

for each user that can maximize the content spread function on the restricted maximum probability path model. Their model differs from ours because we formulate the problem in an incremental marginal gain approach without the spread path restricts. Tong *et al.* [12] raise the question of which edges to add or remove from a network to speed-up a dissemination. They propose an algorithm to optimize the leading eigenvalue of the graph adjacency matrix that control the information dissemination process in their models. Antaris *et al.* [13], Rafailidis *et al.* [14], Rafailidis and Nanopoulos [15] define the link injection problem that is aiming at boosting information cascades. The injected links are being predicted in a collaborative-filtering fashion, based on factorizing the adjacency matrix that represents the structure of the social network. Li *et al.* [16] add the edges to the target users. Some alternative ways to maximize content spread are discussed in the following studies. When seed set is predetermined, Lu *et al.* [17] study how to maximize the expected number of adoptions by providing initial seed users with complementing products. Liontis and Pitoura [18] assume that it is possible to improve the reaction to diffusion process of a small number of nodes by investing extra resources. Then define boosting a node as improving its probability of influencing others, making the node react to an activation faster, or both. In [19], a k-boosting problem is defined which aims to find k users who are initially uninfluenced and increase their probability to be influenced. These studies differ from ours since they are boosting the seed nodes not the content spread. Yang *et al.* [20] first propose active friending problem where a user actively specifies a friending target and maximizes the probability that the friending target would accept the invitation. Their work differs from ours because they know which edges the user would like to connect. But our setting is to find these edges.

III. Formulation of Content Maximization Problem

For a given potential candidate connections set $\overline{X}$, the content maximization problem is to find a subset $X \subseteq \overline{X}$ that maximizes the content spread function $f(X)$ and satisfies some constraints. Now, we give our generalized content maximization problem (GCMP) as follows.

Definition 1 (GCMP): Given a directed acyclic graph $G = (V, E, P)$, a constant $\hat{k}$ and a particular content c with given initial seed set S, find an edge set $X \subseteq \overline{X} = \{e_{ij} : i, j \in V, i \in N_j, j \in N_i\}$ where N_i is the candidate node set of i to be connected such that: 1) at most $\hat{k}$ edges from X and 2)$f(X)$ maximize the content spread under $(V, E \bigcup X, P)$.

In definition 1 of GCMP, we just constraint the total cardinality of the edge set which corresponding to the most classical settings of influence maximization problem. To add average k edges to each seed in the work of Chaoji *et al.* [2] cannot be consistent with power law property of realistic social network. It is also validated in detail through the simulation experiments in Section VI.

In Section I, we have shown that in order to guarantee the submodularity of content spread function $f(X)$, the restricted variant RMPP model of Chaoji *et al.* [2] model is tight due to their restricted influence propagation between a pair of nodes only along the maximum probability path and only RMPP is allowed to calculate the content spread which may cause great deviation from the actual value. In this section, we formulate the content spread function from a marginal increment perspective and describe the content spread function value as accurate as possible. Our formulation is based on the classical discrete-time-independent cascade (IC) model [11] and its topic-aware version, i.e., topic-aware independent cascade (TIC), [31]. When a node v receives or generates a new piece of content c at time t, it has only one chance to share and active each of its inactive out-neighbors with an independent probability.

In the TIC model, the user-to-user influence probabilities depend on the topic. Therefore, for each edge $(v, u) \in E$ and each topic $z \in [1, K]$, we are given a probability p_{vu}^z, representing the strength of the influence exerted by user v on user u on topic z. For each content c that propagates in the network, we have a distribution over the topics $(\gamma_c^1, \gamma_c^2, \ldots, \gamma_c^K,)$ with $\gamma_c^z = P(Z = z|c)$ and $\Sigma_{z=1}^K \gamma_c^z = 1$. The tentative succeeds with a probability that is the weighted average of the link probability with respect to the topic distribution of the content c, that is $p_{vu}^c = \Sigma_{z=1}^K \gamma_c^z p_{vu}^z$. (We use p_{vu} instead of p_{vu}^c for the sake of simplicity of symbolic expression when content c is fixed in the rest of this paper.) For the potential edge $(i, j) \in \overline{X}$, the link probability p_{ij}^c (and p_{ij}) can be determined in the similar way.

For the given acyclic directed social network $G(V, E, P)$, whose nodes indicate users and edges represent the social relations among users. Given content c with distribution over the topics $(\gamma_c^1, \gamma_c^2, \ldots, \gamma_c^K,)$ and seed set S, denote $q_v^E := q_{v,S}^{cE} = \Sigma_{z=1}^K \gamma_c^z q_{v,S}^{zE}$ is the spread of a content c contained at $v \in V$ under the topology of E (which means only the edges in E can be used in the propagation of content c) with seed set S (that is, every node in S contain content c) and $q^E = (q_1^E, \ldots, q_v^E, \ldots, q_{|V|}^E)^T$ is the content spread $|V|-$ dimension vector correspondingly. Denote $\Delta q_v^E(s,t) = q_v^{E \bigcup (st)} - q_v^E$ the marginal gain of the spread on node v when an edge $e = (s, t)$ is merged into the edge set E. Then, we have the following formula to calculate the marginal gain $\Delta q_v^E(s,t)$ for content spread of c at node v.

Theorem 1: The marginal gain $\Delta q_v^E(s,t)$ of content spread of c at node v when an edge $(s,t) \in X$ from a candidate set is added to current topology of E is calculated recursively as follows:

$$\Delta q_t^E(s,t) = \left(1 - q_t^E\right) p_{st} q_s^E .$$

and for any $v \in N^{\text{out}}(t)$, where $N^{\text{out}}(t)$ is the out-neighbor set of vertex t, we have

$$\Delta q_v^E(s,t) = \frac{1 - q_v^E}{1 - p_{tv} q_t^E} p_{tv} \Delta q_t^E(s,t) . \tag{1}$$

Furthermore, for other vertex $v \in V$ that can be reachable from vertex t, we can update the marginal gain similarly according to the topology order in recursive manner. We have $\Delta q_v^E(s,t) = 0$, for the vertex which is unreachable from vertex t during this process.

Proof: For a new edge $(s,t) \in X$ is added to the network, it is easy to see the marginal gain of content spread at t is $\Delta q_t^E(s,t) = (1-q_t^E)p_{st}q_s^E$. For each vertex $v \in N^{\text{out}}(t)$, the content spread of c contained at v can be expressed as $q_v^E = q_v^{E\setminus\{(t,v)\}} + (1-q_v^{E\setminus\{(t,v)\}})p_{tv}q_t^E$. Therefore, we have $q_v^{E\setminus\{(t,v)\}} = ((q_v^E - p_{tv}q_t^E)/(1-p_{tv}q_t^E))$. After updating the spread at t, the content spread of c contained at v increase to

$$\begin{aligned}
&q_v^{(E\bigcup(s,t))}\\
&\quad= q_v^{E\setminus\{(t,v)\}} + \left(1-q_v^{E\setminus\{(t,v)\}}\right)p_{tv}\left(q_t^E + \Delta q_t^E(s,t)\right)\\
&\quad= q_v^{E\setminus\{(t,v)\}} + \left(1-q_v^{E\setminus\{(t,v)\}}\right)p_{tv}q_t^E\\
&\qquad+\left(1-q_v^{E\setminus\{(t,v)\}}\right)p_{tv}\Delta q_t^E(s,t)\\
&\quad= q_v^E + \frac{1-q_v^E}{1-p_{tv}q_t^E}p_{tv}\Delta q_t^E(s,t)\\
&\quad= q_v^E + \Delta q_v^E(s,t).
\end{aligned}$$

This update procedure can be processed recursively according to the topology order of the network until no more nodes can be updated. For those nodes to which there is no path from node t, then $\Delta q_v^E(s,t) = 0$. ∎

Note that during the process of updating marginal spread, if there are paths from vertex t reaching to different in-neighbor nodes of node w, the marginal gain of spread of w should be updated according to (1) multitimes. However, the overall marginal gain of content spread for w is independent of the update orders.

In fact, suppose there exist two paths from t to w via nodes u and v where $w \in N_E^{\text{out}}(u) \bigcap N_E^{\text{out}}(v)$. We first consider update from u to w, a marginal gain of spread $\Delta^u q_w^E(s,t) = ((1-q_w^E)/(1-p_{uw}q_u^E))p_{uw}\Delta q_u^E(s,t)$ is obtained. Then considering update from v to w, another marginal gain of spread $\Delta^{u+v}q_w^E(s,t) = ((1-(q_w^E+\Delta^u q_w^E(s,t)))/(1-p_{vw}q_v^E))p_{vw}\Delta q_v^E(s,t)$. Thus, the overall marginal gain of spread of w is

$$\begin{aligned}
&\Delta q_w^E\\
&= \Delta^u q_w^E(s,t) + \Delta^{u+v}q_w^E(s,t)\\
&= \Delta^u q_w^E(s,t) + \frac{1-\left(q_w^E+\Delta^u q_w^E(s,t)\right)}{1-p_{vw}q_v^E}p_{vw}\Delta q_v^E(s,t)\\
&\quad-\frac{\Delta^u q_w^E(s,t)}{1-p_{vw}q_v^E}p_{vw}\Delta q_v^E(s,t)\\
&= \Delta^u q_w^E(s,t) + \Delta^v q_w^E(s,t)\\
&\quad-\frac{1-q_w^E}{1-p_{vw}q_v^E}p_{vw}\Delta q_v^E(s,t)\Delta^u q_w^E(s,t)\frac{1}{1-q_w^E}\\
&= \Delta^v q_w^E(s,t) + \Delta^u q_w^E(s,t)\left(1-\Delta^v q_w^E(s,t)\frac{1}{1-q_w^E}\right)\\
&= \Delta^v q_w^E(s,t) + \frac{1-q_w^E-\Delta^v q_w^E(s,t)}{1-q_w^E}\frac{1-q_w^E}{1-p_{uw}q_u^E}p_{uw}\Delta q_u^E(s,t)\\
&= \Delta^v q_w^E(s,t) + \Delta^{v+u}q_w^E(s,t).
\end{aligned}$$

From theorem 1 and the note above, the objective function of content spread in the marginal gain form can be expressed as $f(X) = \sum_{v\in V}(q_v^E + \sum_{(s,t)\in X}\Delta q_v^{(E\bigcup Xst)}(s,t))$, where Xst denotes the edge set that have already been added into the network before edge (s,t). The order-independent property shows the definition of $f(X)$ expressed above is well-defined. More importantly, this definition is consistent with the content propagation process and there is no loss during the content spread process.

IV. Submodular Bounds and MIS Algorithm

In the GCMP, the objective function $f(X) = \sum_{v\in V}(q_v^E + \sum_{(s,t)\in X}\Delta q_v^{(E\bigcup X^{st})}(s,t))$ is nonsubmodular which increases the challenges of designing efficient and effective algorithm for the GCMP. In order to elaborate on the structure of $f(X)$, we can rewrite it in marginal increment form. Denote $X = \{(s_1,t_1),(s_2,t_2),\ldots,(s_{\hat{k}},t_{\hat{k}})\}$, $X^k = \{(s_1,t_1),\ldots,(s_k,t_k)\}$, $k = 1,2,\ldots,\hat{k}$ and $X^0 = \emptyset$ for convenience. Then, we have $f(X) = f(X^0) + \sum_{k=1}^{\hat{k}}\Delta_k f(X^{k-1})$, where $f(X^0) = \sum_{v\in V}q_v^E$ and $\Delta_k f(X^{k-1}) = \sum_{v\in V}\Delta q_v^{(E\bigcup X^{k-1})}(s_k,t_k)$, $k = 1,\ldots,\hat{k}$. Fortunately, each term $\Delta q_v^{(E\bigcup X^{k-1})}(s_k,t_k)$ in $\Delta_k f(X^{k-1})$ is monotone decrease with $q_v^{(E\bigcup X^{k-1})}$. Thus, we have the following monotone decrease property of $f(X)$.

Property 1: Content spread function $f(X) = f(X^0) + \sum_{k=1}^{\hat{k}}\Delta_k f(X^{k-1})$ is monotone decrease with $q_v^{(E\bigcup X^{k-1})}$, for $v \in V$ and $k = 1,\ldots,\hat{k}$.

Proof: First note that $\Delta q_{t_k}^{(E\bigcup X^{k-1})}(s_k,t_k) = (1-q_{t_k}^{(E\bigcup X^{k-1})})p_{s_kt_k}q_{s_k}^{(E\bigcup X^{k-1})}$ is monotonically decreasing with $q_{t_k}^{(E\bigcup X^{k-1})}$, respectively. Second, $\Delta q_v^{(E\bigcup X^{k-1})}(s_k,t_k) = ((1-q_v^{(E\bigcup X^{k-1})})/(1-p_{t_kv}q_{t_k}^{(E\bigcup X^{k-1})}))p_{t_kv}\Delta q_{t_k}^{(E\bigcup X^{k-1})}(s_k,t_k)$ is monotone decrease with both $q_v^{(E\bigcup X^{k-1})}$ and $q_{t_k}^{(E\bigcup X^{k-1})}$. In fact, the former is obvious. To show the latter, we consider the derivative of $\Delta q_v^{(E\bigcup X^{k-1})}(s_k,t_k)$ with respect to $q_{t_k}^{(E\bigcup X^{k-1})}$, that is,

$$\begin{aligned}
&\frac{\partial\Delta q_v^{(E\bigcup X^{k-1})}(s_k,t_k)}{\partial q_{t_k}^{(E\bigcup X^{k-1})}}\\
&= \frac{\partial}{\partial q_{t_k}^{(E\bigcup X^{k-1})}}\left\{\frac{1-q_v^{(E\bigcup X^{k-1})}}{1-p_{t_kv}q_{t_k}^{(E\bigcup X^{k-1})}}p_{t_kv}\Delta q_{t_k}^{(E\bigcup X^{k-1})}(s_k,t_k)\right\}\\
&= \frac{\partial}{\partial q_{t_k}^{(E\bigcup X^{k-1})}}\left\{\frac{1-q_v^{(E\bigcup X^{k-1})}}{1-p_{t_kv}q_{t_k}^{(E\bigcup X^{k-1})}}\right.\\
&\qquad\left.\times p_{t_kv}\left(1-q_{t_k}^{(E\bigcup X^{k-1})}\right)p_{s_kt_k}q_{s_k}^{(E\bigcup X^{k-1})}\right\}\\
&= p_{t_kv}\left(1-q_v^{(E\bigcup X^{k-1})}\right)p_{s_kt_k}q_{s_k}^{(E\bigcup X^{k-1})}\\
&\quad\times\frac{p_{t_kv}-1}{\left(1-p_{t_kv}q_{t_k}^{(E\bigcup X^{k-1})}\right)^2} \le 0.
\end{aligned}$$

for $v \in V\setminus\{t_k\}$ and $k = 1,\ldots,\hat{k}$. Because monotonically decreasing property is still hold under summation operations, so we complete the proof. ∎

However, the monotone decrease property does not guarantee the submodularity of the objective function $f(X)$, this is just because the neighbor relationship will

change during the new edges are added into the network. In order to see this structural change clearly, we denote $N_E^{\text{out}}(v) = \{u \in V | (v,u) \in E\}$ the out-neighbor of vertex v under edge set E. Obviously, we have the inclusion relationship $N_E^{\text{out}}(v) \subseteq N_{E \bigcup X^1}^{\text{out}}(v) \subseteq N_{E \bigcup X^2}^{\text{out}}(v) \subseteq \cdots \subseteq N_{E \bigcup X^K}^{\text{out}}(v) \subseteq N_{E \bigcup \bar{X}}^{\text{out}}(v)$. With this notation, $\Delta_k f(X^{k-1})$ can be rewritten in a more detailed expression

$$\begin{aligned}
&\Delta_k f(X^{k-1}) \\
&= \sum_{v \in V} \Delta q_v^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&= \Delta q_{t_k}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \sum_{v_1 \in N_{E \bigcup X^{k-1}}^{\text{out}}(t_k)} \Delta q_{v_1}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \sum_{v_2 \in N_{E \bigcup X^{k-1}}^{\text{out}}(v_1), v_1 \in N_{E \bigcup X^{k-1}}^{\text{out}}(t_k)} \Delta q_{v_2}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \cdots + \sum_{v_D \in N_{E \bigcup X^{k-1}}^{\text{out}}(v_{D-1}), v_{D-1} \in N_{E \bigcup X^{k-1}}^{\text{out}}(v_{D-2})} \Delta q_{v_D}^{(E \bigcup X^{k-1})} \\
&\quad \times (s_k, t_k)
\end{aligned}$$

where $D = D_k$ is the largest hops number among all the paths originate from the vertex t_k and $k = 1, 2, \ldots, \hat{k}$. The reason that $f(X)$ is not submodular lies in that the out-neighbor set of a vertex $v \in V$ may become larger and larger with new edges added in the maximize content spread update process. However, for a given GCMP, initial edge set E, propagation probability P and initial seed set S that contains content c are all fixed. When we further fix the out-neighbor relationship of all vertexes $v \in V$ during the whole updating process, the number of marginal gain terms will remain unchanged during the whole procedure. Due to the monotonically increasing property of $q_v^{(E \bigcup X^{k-1})}$ with respect to edge set X^{k-1} and from property 1, for each newly added edge (s_k, t_k), the resultant marginal gain term of content spread $\Delta q_v^{(E \bigcup X^{k-1})}(s_k, t_k)$ becomes monotonically decreasing with $q_v^{(E \bigcup X^{k-1})}$, for $v \in V$ and thus further make it possible to guarantee that the associate content spread function is submodular. In the following subsection, we construct submodular lower bound and submodular upper bound of the objective functions by reasonably imposing restriction on the neighborhood structure of each vertex $v \in V$.

A. Submodular Lower Bound and Upper Bound

Although the content spread function we obtained is not submodular, we can fortunately construct a submodular lower bound and submodular upper bound in a marginal gain incremental way. Based on the monotone analysis of marginal gain term of content spread above, we can construct the lower bound of the objective function as follows: $\underline{f}(X) = f(X^0) + \sum_{k=1}^{\hat{k}} \Delta_k \underline{f}(X^{k-1})$, where

$$\begin{aligned}
&\Delta_k \underline{f}(X^{k-1}) \\
&= \sum_{v \in V} \underline{\Delta} q_v^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&= \Delta q_{t_k}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \sum_{v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_1}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \sum_{v_2 \in N_E^{\text{out}}(v_1), v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_2}^{(E \bigcup X^{k-1})}(s_k, t_k) \\
&\quad + \cdots + \\
&\quad \sum_{v_D \in N_E^{\text{out}}(v_{D-1}), v_{D-1} \in N_E^{\text{out}}(v_{D-2})} \Delta q_{v_D}^{(E \bigcup X^{k-1})}(s_k, t_k).
\end{aligned}$$

$k = 1, 2, \ldots, \hat{k}$. $\underline{f}(X)$ defined above is lower bound of $f(X)$ because all the term $\Delta q_v^{(E \bigcup X^{k-1})}(s_k, t_k)$ in $\underline{f}(X)$ is nonnegative and must be included in $f(X)$ due to the inclusion relationship $N_E^{\text{out}}(v) \subseteq N_{E \bigcup X^1}^{\text{out}}(v) \subseteq N_{E \bigcup X^2}^{\text{out}}(v) \subseteq \cdots \subseteq N_{E \bigcup X^K}^{\text{out}}(v) \subseteq N_{E \bigcup \bar{X}}^{\text{out}}(v)$. Furthermore, $\underline{f}(X)$ have the following nice submodular property.

Theorem 2: The lower bound of objective function $\underline{f}(X) = f(X^0) + \sum_{k=1}^{\hat{k}} \Delta_k \underline{f}(X^{k-1})$ defined above is submodular with respect to X.

Proof: For any $X \subseteq Y \subseteq \overline{X}$ and any $(s,t) \in \overline{X} \backslash Y$, we have $\Delta q_v^{(E \bigcup X)}(s,t) \geq \Delta q_v^{(E \bigcup Y)}(s,t)$ due to property 1 and the monotonically increasing property of $q_v^{(E \bigcup X)}$ with respect to edge set X. In addition, the number of marginal gain terms does not change during the whole edges added process. Therefore, we have

$$\begin{aligned}
&\underline{f}\big(X \bigcup \{(s,t)\}\big) - \underline{f}(X) \\
&\quad = \Delta q_{t_k}^{(E \bigcup X)}(s,t) + \sum_{v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_1}^{(E \bigcup X)}(s,t) \\
&\qquad + \sum_{v_2 \in N_E^{\text{out}}(v_1), v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_2}^{(E \bigcup X)(s,t)} \\
&\qquad + \cdots + \sum_{v_D \in N_E^{\text{out}}(v_{D-1}), v_{D-1} \in N_E^{\text{out}}(v_{D-2})} \Delta q_{v_D}^{(E \bigcup X)}(s,t) \\
&\quad \geq \Delta q_{t_k}^{(E \bigcup Y)}(s,t) + \sum_{v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_1}^{(E \bigcup Y)}(s,t) \\
&\qquad + \sum_{v_2 \in N_E^{\text{out}}(v_1), v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_2}^{(E \bigcup Y)}(s,t) + \cdots \\
&\qquad + \sum_{v_D \in N_E^{\text{out}}(v_{D-1}), v_{D-1} \in N_E^{\text{out}}(v_{D-2})} \Delta q_{v_D}^{(E \bigcup Y)}(s,t) \\
&\quad = \underline{f}\big(Y \bigcup \{e_{st}\}\big) - \underline{f}(Y)
\end{aligned}$$

which shows the submodularity of $\underline{f}(X)$. ■

Now, we can construct the upper bound of objective function $\overline{f}(X) = f(\overline{X}) - \sum_{(s,t) \in \overline{X} \backslash X} \Delta_{(st)} f(\overline{X})$, where

$$f(\overline{X}) = \sum_{v \in V} q_v^{(E \bigcup \overline{X})}$$

and $\forall e_{st} \in \overline{X}\backslash X$

$$\begin{aligned}
&\Delta_{(st)} f(\overline{X}) \\
&= \sum_{v \in V} \Delta q_v^{(E \bigcup \overline{X})}(s,t) \\
&= \Delta q_{t_k}^{(E \bigcup \overline{X})}(s,t) + \sum_{v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_1}^{(E \bigcup \overline{X})}(s,t) \\
&\quad + \sum_{v_2 \in N_E^{\text{out}}(v_1), v_1 \in N_E^{\text{out}}(t_k)} \Delta q_{v_2}^{(E \bigcup \overline{X})}(s,t) \\
&\quad + \cdots + \sum_{v_D \in N_E^{\text{out}}(v_{D-1}), v_{D-1} \in N_E^{\text{out}}(v_{D-2})} \Delta q_{v_D}^{(E \bigcup \overline{X})}(s,t))
\end{aligned}$$

$\overline{f}(X)$ defined above is upper bound of $f(X)$ because $\overline{f}(X) - f(X) = (f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash X} \Delta_{(st)} f(\overline{X})) - (f(X^0) + \sum_{k=1}^{\hat{k}} \Delta_k f(X^{k-1})) \geq \sum_{(s,t)\in\overline{X}\backslash X} (\Delta_{(st)} f(X^{l_{st}-1})) - \Delta_{(st)} f(\overline{X}))) \geq 0$, here l_{st} denote the edge (s,t) is the l_{st}th edge added into the recommendations boost network among all the edges in $\overline{X}$. The last inequality holds because $\Delta_{(st)} f(X^{l_{st}-1})$ has at least the same number of items as $\Delta_{(st)} f(\overline{X})$ has and $\Delta q_v^{(E \bigcup X^{l_{st}-1})}(s,t) \geq \Delta q_v^{(E \bigcup \overline{X})}(s,t)$ due to the monotonically decreasing property which is mentioned in property 1. Similarly, $\overline{f}(X)$ have the following nice submodular property.

Theorem 3: The upper bound of objective function $\overline{f}(X) = f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash X} \Delta_{(st)} f(\overline{X})$ defined above is submodular with respect to X.

Proof: For any $X \subseteq Y \subseteq \overline{X}$ and any $(s', t') \in \overline{X}\backslash Y$, by the definition of $\overline{f}(X)$, we have

$$\begin{aligned}
&\overline{f}(X \bigcup \{(s', t')\}) - \overline{f}(X) \\
&= (f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash(X \bigcup \{(s',t')\})} \Delta_{(st)} f(\overline{X})) \\
&\quad - (f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash X} \Delta_{(st)} f(\overline{X})) \\
&= \Delta_{(s't')} f(\overline{X}) \geq \Delta_{(s't')} f(\overline{X}) \\
&= (f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash(Y \bigcup \{(s',t')\})} \Delta_{(st)} f(\overline{X})) \\
&\quad - (f(\overline{X}) - \sum_{(s,t)\in\overline{X}\backslash Y} \Delta_{(st)} f(\overline{X})) \\
&= \overline{f}(Y \bigcup \{(s', t')\}) - \overline{f}(Y)
\end{aligned}$$

which shows the submodularity of $\overline{f}(X)$. ■

B. MIS Algorithm

Generally, there is no effective way to optimize or approximate a nonsubmodular function [21], [22]. Lu *et al.* [17] propose a sandwich approximation strategy, which approximates the nonsubmodular objective function by approximating its submodular lower bound and upper bound. As far as the GCMP is concerned, although the original contend spread function $f(X)$ is nonsubmodular, we have obtained the submodular lower $\underline{f}(X)$ and upper bounds $\overline{f}(X)$ in Section IV-A. Therefore, the sandwich framework can be applied, and we devise a marginal increment-based algorithm (MIS) that guarantees a data-dependent approximation factor. The sandwich approximation strategy works as follows. First, a solution to the original problem with any strategy is found. Then, an approximate solution to the submodular lower bound and the submodular upper bound is found, respectively. Finally, the solution that has the best result for the original problem is returned. Algorithm 1 shows the general framework of MIS.

Algorithm 1 Marginal Increment-Based Sandwich Approximation Framework (MIS)

1: Let X_U be α-approximation to the upper bound $\overline{f}(X)$.
2: Let X_L be β-approximation to the lower bound $\underline{f}(X)$.
3: Let X_A be a solution to the original problem $f(\overline{X})$.
4: $X = \arg\max_{X_0 \in \{X_U, X_L, X_A\}} f(X_0)$.
5: **return** X.

The solution returned by the MIS has a data-dependent approximation factor, which is presented in the following theorem.

Theorem 4: Let X^* be the edge set returned by MIS and X_A^* is the optimal solution maximizing the GCMP, then we have

$$f(X^*) \geq \max\left\{\frac{f(X_U)}{\overline{f}(X_U)}\alpha, \frac{\underline{f}(X_L^*)}{f(X_A^*)}\beta\right\} f(X_A^*).$$

Proof: Let X_L^*, X_U^*, and X_A^* be the optimal solutions maximizing the lower bound, the upper bound, and the original spread of GCMP, respectively. Then, we have

$$\begin{aligned}
f(X_U) &= \frac{f(X_U)}{\overline{f}(X_U)}\overline{f}(X_U) \geq \frac{f(X_U)}{\overline{f}(X_U)}\alpha\overline{f}(X_U^*) \\
&\geq \frac{f(X_U)}{\overline{f}(X_U)}\alpha\overline{f}(X_A^*) \geq \frac{f(X_U)}{\overline{f}(X_U)}\alpha f(X_A^*).
\end{aligned}$$

And

$$f(X_L) \geq \underline{f}(X_L) \geq \beta\underline{f}(X_L^*) = \frac{\underline{f}(X_L^*)}{f(X_A^*)}\beta f(X_A^*).$$

Let $X^* = \arg\max_{X_0 \in \{X_U, X_L, X_A\}} f(X_0)$, then

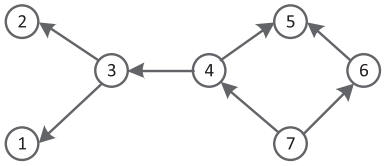
$$f(X^*) \geq \max\left\{\frac{f(X_U)}{\overline{f}(X_U)}\alpha, \frac{\underline{f}(X_L^*)}{f(X_A^*)}\beta\right\} f(X_A^*).$$

■

For both submodular lower bound and upper bound, the greedy hill algorithm can guarantee a $\alpha = \beta = 1 - (1/e)$ approximation factor. Thus, we have the corollary as follows.

Corollary 1: Let X^* be the edge set returned by MIS and X_A^* is the optimal solutions to maximizing the GCMP, then we have

$$f(X^*) \geq \max\left\{\frac{f(X_U)}{\overline{f}(X_U)}, \frac{\underline{f}(X_L^*)}{f(X_A^*)}\right\}\left(1 - \frac{1}{e}\right) f(X_A^*).$$



$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \end{bmatrix}$$

$$\begin{aligned} e &= [1\ 1\ 1\ 1\ 1\ 1\ 1]^T \\ Ae &= [0\ 0\ 1\ 1\ 0\ 0.5\ 1]^T \\ A^2e &= [0\ 0\ 0\ 0.5\ 0\ 0\ 0.75]^T \\ A^3e &= [0\ 0\ 0\ 0\ 0\ 0\ 0.25]^T \\ A^4e &= [0\ 0\ 0\ 0\ 0\ 0\ 0]^T \\ \sigma &= e + Ae + A^2e + A^3e + A^4e \\ &= [1\ 1\ 2\ 2.5\ 1\ 1.5\ 3]^T \end{aligned}$$

Fig. 1. Example with propagation probability $p = 0.5$.

It is worth pointing out that the approximation factor obtained in the above theorem and corollary is highly depend on the quality of both the lower and upper bounds. Compared to the second term inside the $\max\{\cdot\}$, the first term can be calculated efficiently and can be of practical value.

V. IRFA Algorithm

Another key point in the MIS is how to obtain high-quality solution to the original problem $f(X)$. We will present a novel scalable heuristic method IRFA, which is based on influence ranking of a single node and fast adjustment according to the recommendations selected in the network.

A. Influence Ranking of Single Node

Intuitively, it is beneficial to set up the connections between seed nodes with content c and strong influence nodes to boost content spread. Therefore, how to rank the influence of a single node is very important. In order to obtain this useful influence information, denote σ_v^E is the spread factor if $\{v\}$ is the only seed node in the given social network $G = (V, E, P)$ under the edge set E and $\sigma^E = (\ldots, \sigma_v^E, \ldots)^T$ is the corresponding spread vector under the topology determined by E. We denote $\sigma^E = \sum_{0\le l\le D} \sigma^l = \sum_{0\le l\le D} A^l e$, where $e = (1, 1, \ldots, 1)_n^T$ is a n-dimensional vector with all components of 1 and $A = (a_{ij})_{n\times n}$ is the adjacency propagation matrix of G with $a_{ij} = p_{ij} = p_{ij}^c$ if $e_{ij} \in E$ and 0 otherwise. D is the diameter of the network G. Usually $\sigma^l, l = 0, 1, \ldots, D$ reflects the lth hop propagation spread of all vertex in the network. Fig. 1 shows an example to demonstrate how we calculate σ^E.

Using this spread influence information, we know how to establish connections to maximize the spread. Suppose that node 7 is the only node has content c, if we want add another connection among node 7's two-hop neighbors, we say node 3 ($\sigma_3^E = 2$) is better than node 5 ($\sigma_5^E = 1$) because 3 has higher spread capability or node 3 is more influential than node 5. As for the spread vector, we have the following properties.

Property 2: For a directed acyclic graph G, if D = diameter(G), then $A^{D+k} = 0$, for all $k = 1, 2, \ldots$

Property 2 shows that if the diameter of a graph is D, then any node in G must finish its propagation within no more than D hops. $\sigma^l = A^l e$, $l = 0, 1, \ldots, D$ is called l-hop influence vector with the component of which is equal to the expectation of relaxed influence of node $v \in V$ in exactly l hops propagation. Here, relaxed means that we allow multiple counts of influence on some nodes without considering their interactive influence.

Now, turn to the property of the spread vector from an average point of view. Suppose $\overline{d}$ is the average out-degree of the network and $\overline{p}$ is the average propagation probability. Then, we have

Property 3:

1) $\|\sigma^l\|_1 = n(\overline{d}\,\overline{p})^l$, $l = 0, 1, \ldots, D$, where $\|\sigma^l\|_1 = \sum_{u=1}^n \sigma^l(u)$ denotes the 1-norm of n-dimension vector σ^l. In particular, $\|\sigma^0\|_1 = n$, $\|\sigma^1\|_1 = n\overline{d}\,\overline{p}$.
2) $\|\sigma\|_1 = \sum_{l=0}^D \|\sigma^l\|_1 \le (n/(1-\overline{d}\,\overline{p}))$, if $\overline{d}\,\overline{p} < 1$; $\|\sigma\|_1 = \sum_{l=0}^D \|\sigma^l\|_1 = (D+1)n$, if $\overline{d}\,\overline{p} = 1$; $\|\sigma\|_1 = \sum_{l=0}^D \|\sigma^l\|_1 \le (n(\overline{d}\,\overline{p})^{D+1}/(\overline{d}\,\overline{p} - 1))$, if $\overline{d}\,\overline{p} > 1$.
3) If $\overline{d}\,\overline{p} \ll 1$, every node at most have $(1/(1-\overline{d}\,\overline{p}))$ influence in average sense, which can be reach a good approximation by only using 0-step and 1-step influence vector.

Based on these properties, we present the influence ranking algorithm of a single node (IR) in the following.

Algorithm 2 Influence Ranking IR ($G(E)$)

Input: Social network G, diameter D and adjacency propagation matrix A which is determined by E.
Output: spread vector σ^E.

```
initialize σ^E = e
for d = 0 to D do
    Compute σ^E = +Aσ^E
end for
return σ^E.
```

For the GCMP, we aim to add new edges from $\overline{X}$ to further boost the spread of content, therefore what we really care about is how much the spread increment caused by newly added edges, not content spread itself. Therefoer, another important factor that influences the spread increment should be considered at the same time. This factor is q_v^E, the content spread of c contained at $v \in V$ under the topology of E which is equivalent to the accumulated activate probability of each node received before the edge added.

B. IRFA Algorithm

Based on the analysis in Section V-A, we present our IRFA algorithm in this section.

The main idea of scalable heuristic method IRFA is to select the node with the maximum weighted influence in the sense of $(1 - q_t^E)\sigma^E(t)$ from the candidate set of vertex s, and add the connection (s, t) into edge set. That is, add edge (s, t) into X such that $t = \arg\max_{v:(s,v)\in\overline{X}_s}(1 - q_v^{(E\bigcup X)})\sigma^E(v)$, where $q_v^{(E\bigcup X)}$ is the probability that node v becomes activated after the diffusion process when the edge set is $E \bigcup X$. In order

to maintain the current influence of each node, a fast update adjustment procedure is needed.

The influence **fast update adjustment procedure** [FA(s,t)] is as follows.

FA(s,t): After the edge (s,t) is added into the current edge set, first update $\sigma^{E\bigcup(X\bigcup\{(s,t)\})}(s) = \sigma^{E\bigcup X}(s) + p_{st}\sigma^{E\bigcup X}(t)$; then $\sigma^{E\bigcup(X\bigcup\{(s,t)\})}(v) = \sigma^{E\bigcup X}(v)$ remains unchanged for the descendant node of node t (include t itself); at last, reversely update the ancestor node according to $\sigma^{E\bigcup(X\bigcup\{(s,t)\})}(v) = \sigma^{E\bigcup X}(v) + \Delta_{(st)}\sigma^{E\bigcup X}(v)$, here $\Delta_{(st)}\sigma^{E\bigcup X}(v) = p_{vu}\Delta_{(st)}\sigma^{E\bigcup X}(u)$ for $u \in N^{\text{out}}_{E\bigcup X}(v)$.

Now, we present the overall algorithm of IRFA to compute the solution of GCMP.

Algorithm 3 Algorithm IRFA $(G, \hat{k}, \overline{X})$

Input: the social network $G = (V, E, P)$, candidate edge set $\overline{X}$, positive number $\hat{k}$ and content seed set S of the content c.

Output: X satisfying constraint of no larger than $\hat{k}$ edges and maximizing the boost influence spread.

1: Initialize $X = \emptyset$
2: Run Algorithm IR$(G(E))$ to obtain all $\sigma^{E}(v)$ and let $\sigma^{E\bigcup X}(v) = \sigma^{E}(v)$
3: **for** $v \in V$ **do**
4: Compute q_v^E and let $q_v^{(E\bigcup X)} = q_v^E$
5: **end for**
6: **for** $k = 1$ to $\hat{k}$ **do**
7: Select Edge $(s,t) = \arg\max_{v:(u,v)\in\overline{X}_u, u\in S}(1 - q_v^{(E\bigcup X)})\sigma^{E\bigcup X}(v)$
8: Run influence update procedure FA(s,t) to obtain $\sigma^{E\bigcup(X\bigcup(s,t))}(v)$
9: Compute $\Delta q_v^{(E\bigcup X)}(s,t)$ and Update $q_v^{(E\bigcup X\bigcup(s,t))}$
10: Update $X = X\bigcup(s,t)$
11: **end for**
12: **return** X as the solution to the GCMP.

Complexity Analysis: Our proposed algorithm mainly consists of two parts: initialization and iteration. In the initialization, we first calculate the influence ranking σ and initial spread q_v^E of each node. In property 2, we show that any node will propagate at most D-hop neighbors. Therefore, the time complexity of algorithm 2 will be $O(D \cdot m)$ where m is the number of edges in the graph. For a given c, the time complexity to compute q_v^E is $O(|S|m)$ in marginal increment way recursively. The next iteration step consists of $\hat{k}$ loops. Each loop selects the best candidate edge which costs $|\overline{X}|$ comparisons. The update procedure of both $\sigma^{E\bigcup X}$ and $q_v^{(E\bigcup X)}$ cost at most $O(m)$. Therefore, the time complexity of our algorithm costs $O(D \cdot m + |S|m + \hat{k}(|\overline{X}| + 2m))$ time.

VI. Experiments

In this section, we conduct experiments on four data sets to test the effectiveness of MIS and IRFA and compare our adding edges approach with other different strategies.

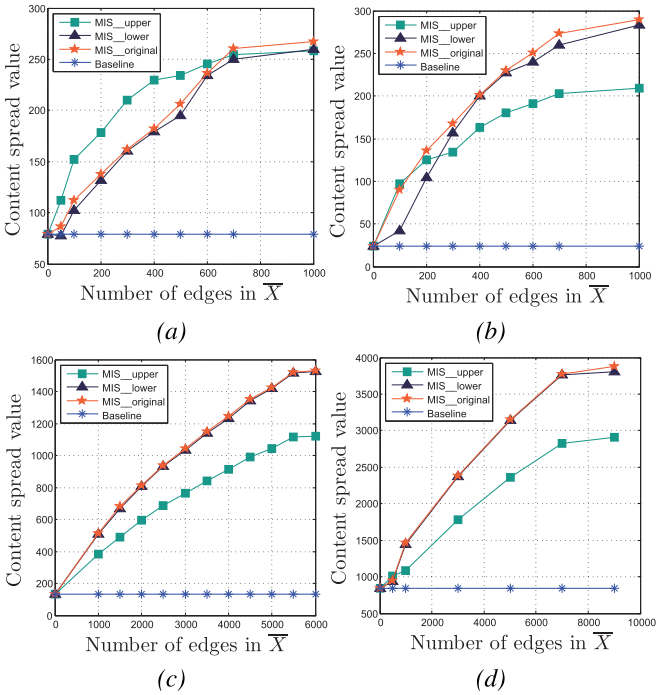


Fig. 2. Content spread value versus incremental edges under uniform and trivalency propagation. (a) Synthetic p = 0.05. (b) Facebook p = 0.05. (c) Wikipedia p = trivalency. (d) HEP-TH p = trivalency.

A. Experiment Setup

We use one synthetic graph and three real-world social graphs in our experiment which is described in the following. These social graphs represent a wide variety of relationship.

1) Synthetic: We randomly generated a relatively small acyclic directed graph with 2000 nodes and 5000 edges used to validate our experiment results.

2) Facebook: This data set includes 1899 users and a total number of 59 835 online messages were sent over 20 296 directed ties among these users. This data set represents an online community for students at a university. The directed edges indicate the friend relation between two users. [28]

3) Wikipedia: The Wikipedia data set is generated by a voting activity which Wikipedia community discuss and vote for the people who to promote to become an administrator. There are 7115 nodes and 103 689 edges. Each node in the graph represents a user attend the voting procedure. Each directed edge denotes who vote for whom.

4) HEP-TH: This is a citation graph which from the e-print arXiv and covers all the citations within a data set of 27 770 papers with 352 807 edges. If a paper i cites paper j, the graph contains a directed edge from i to j [26], [27].

All of our data sets only have the relationship between two nodes but no other information on each node or edge can be used directly. For ease of comparison, we assume that a content c is determined by a given distribution topics. We further assume the following seed set generation process. Seed set S whose node contains the content c is selected randomly and uniformly by a rate from each data set, e.g., 1% of total nodes. For edge selections, we first generate a candidate edge set $\overline{X}$ in which the candidate edges are selected from each seed node to its two-hop and three-hop neighbors to

the ratio of 4:1. Then, we select the edges from $\overline{X}$ by the rank of the two-hop and three-hop neighbors' weighted influence. We update the weighted influence after adding each edge until the final adding edges is $0.2|\overline{X}|$.

The influence propagation model we adopt in our experiment is IC model, which is a special case of topic-aware IC model. We avoid cycles on networks by terminating the propagation process when cycle appears. For the propagation probability, first we consider to uniformly assign the weights which assume all nodes have the same synthetic probability $p = p_{vu} = \Sigma_{z=1}^{K} \gamma_c^z p_{vu}^z$ of sharing content. We set $p = 0.05$ with the same setting as in [2]. We also use a trivalency model [5], [24], [25]. For each edge, we uniformly select a value from (0.1, 0.01, 0.001) at random, which corresponds to high, medium, and low influences.

We use greedy algorithm to compute the upper and lower bounds of MIS algorithm. Since the bounds have been shown submodular in Section IV, thus we can achieve $1-1/e$ approximation guarantee. The content spread value is calculated as performance evaluation because our proposed algorithm could compute the objective function highly close to accurate value.

B. Edge Selection Methods

We compare with three heuristic edges selection strategies and other two methods proposed in [2] and [12].

Random, we select $\hat{k}$ edges from $|\overline{X}|$ randomly to add in the graph.

Maxdegree, where the added edges are based on nodes' degree. We first select top $\hat{k}$ degree nodes in $N(\overline{X}) = \{u|(v,u) \in \overline{X}\}$ and each seed node with content c is connected to $\hat{k}/|S|$ selected nodes randomly. Intuitively, this strategy is very competitive because high degree nodes could have high potential to influence more nodes.

PageRank, which is widely known Google Page Rank measure [30]. The pagerank score indicates the importance of a node. To calculate the pagerank of each node, we set the damping factor to 0.9. Nodes with top $\hat{k}$ pagerank score in $N(\overline{X})$ will be connected with seed nodes in the same way as Maxdegree.

R1 and **R2** indicate the edge selection methods in [2] and [12], respectively.

C. Experiment Results

In this section, we evaluate the effectiveness of MIS. First, we compare MIS upper and lower bounds with spread value of GCMP by varying the number of added edge and seed nodes under two probability setting. Second, we show the distribution of added edges on each seed node. Third, we compare with other edges selection methods. Finally, we give a table that concludes our MIS algorithm.

1) Content Spread Value With the Increase in the Number of Added Edges: In Fig. 2, we evaluate MIS algorithm in terms of the content spread over network by varying the number of edges added to graph. We perform experiments on four data sets under two propagation probability settings. Since the content spread grows in very similar form, we present two data sets under uniform probability and two under trivalency

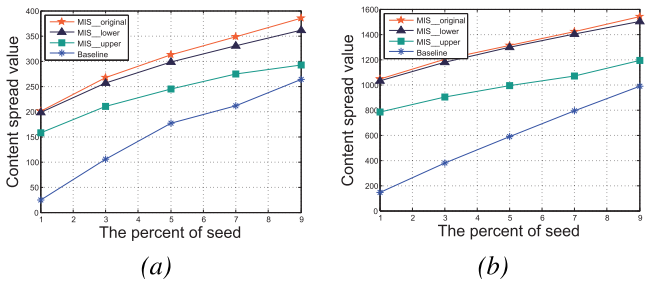


Fig. 3. Content spread value versus incremental seed set size. (a) Facebook p = trivalency. (b) Wikipedia p = 0.05.

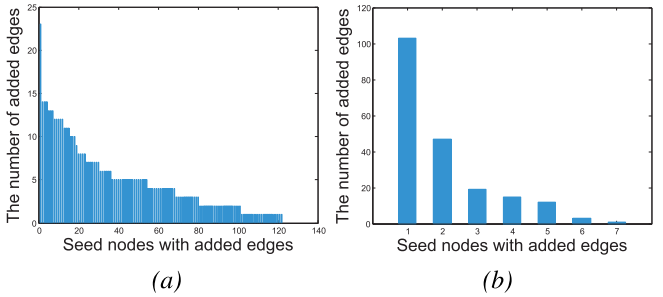


Fig. 4. Number of edges added on each seed node. (a) HEP-TH p = trivalency. (b) Facebook p = 0.05.

model. The x-axis holds the number of candidate edges. The y-axis holds the content spread over whole network after adding these edges. Baseline shows the content spread without any adding edges. In Fig. 2(a)–(d), the content spread value grows dramatically as number of edges increase. This is to say, the edges we select have large impact on the content spread all over the network. Wikipedia data set shows when we add 200 edges (the number of candidate edges is 1000) which is only 0.2% of total 100 000 edges the content spread under our MIS-original becomes 3 times compared to baseline.

We also observe that the increment of content spread value become slow with the increasing number of added edges. It is well explained that although our original problem is not submodular theoretically, two submodular bounds could ensure the submodularity of GCMP to a large extent.

Fig. 2(b)–(d) also shows that the content spread of upper bound solution is not as good as that of lower bound or original. It is because during the process of calculation upper bound content spread, we first add all the candidate edges then remove the edges with small marginal gain without updating $\sigma^{E \bigcup X}$ and $q^{E \bigcup X}$ thus content spread value might loss during removing edges process. This procedure differs from the calculation of the lower bound and original problem which add edges in a dynamic iterative way.

2) Content Spread Value With the Increase in the Number of Seed Nodes: We also conduct experiments on content spread by varying the number of seed nodes. Fig. 3 shows the content spread grows as the size of seed set increases. It is further shown the trend that spread gain is getting slower when the seed size getting larger as the same as Fig. 2. We also observe that the gap between MIS algorithm and baseline becomes smaller when the seed nodes increase from 1% to 9%. As expected, the content spread will converge as the seed

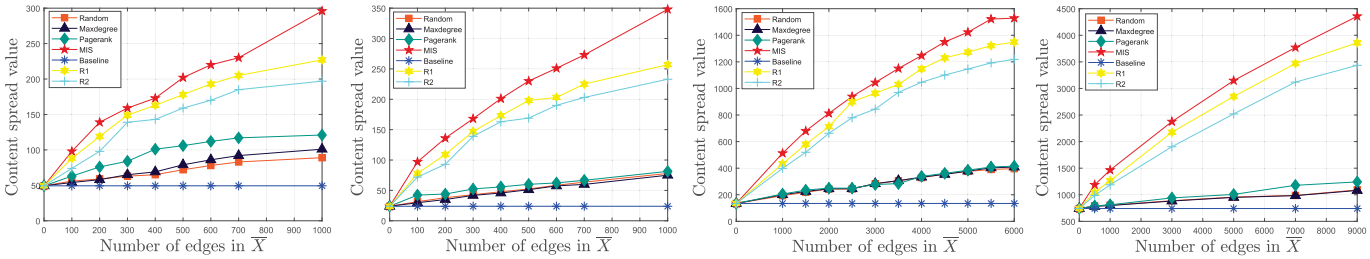


Fig. 5. Comparison with other methods. (a) Synthetic $p = 0.05$. (b) Facebook $p = 0.05$. (c) Wikipedia p = trivalency. (d) HEP-TH p = trivalency.

set becomes very large. The upper bound is also relatively smaller here which is due to the same reason as we discussed above.

3) Number of Edges Added to Each Seed Node: In Fig. 4, we show the number of edges added to each seed node by setting probability to trivalency on data set high energy physics-theory and Facebook. The results are very similar to uniform probability. The x-axis holds the seed nodes with added edges and the y-axis holds how many number of edges we add on each seed node. From the figure, we show that not all the seed nodes have been added edges. In Facebook data set, we set the seed nodes to 0.01 of total nodes on graph. 200 edges are added only on 7 nodes of 18 seed nodes. The added edges are not uniformly distributed on each seed. There are more than half of the edges are added on one seed. Since the inequality role of each edge and node in social network structure, the uneven distribution of added edges is reasonable. This is why in our problem GCMP, we set the added edges to be total $\hat{k}$ which is different from the RMPP problem of Chaoji *et al.* [2]. If the edges are added uniformly for each seed node, it may fail to achieve an optimal solution in terms of power law property of real-world social network.

4) Comparison With Other Edge Selection Strategies: We compare our MIS algorithm with random, Maxdegree, Pagerank score, R1 [2]- and R2 [12]-based selection methods in Fig. 5. Our algorithm outperforms the other heuristics with increasing added edges. The content spread grows rapidly using MIS, R1, and R2 whereas the others increase slowly. It is notable that the difference between ours and heuristics (random, Maxdegree, and Pagerank score) becomes larger when adding more edges. As shown in Fig. 5(d), at the beginning, when we add 100 edges (500 candidate edges) our algorithm is $1/2$ times more than heuristics, when the added edges increase to 1800 (9000 candidate edges) ours perform three times of the Pagerank which is the best of heuristics. Among heuristics, Pagerank performs better than the other two because high Pagerank score means the node is more important than the others on network. When the edges connect to important nodes, the content spread will increase. The results of Maxdegree and random are very similar. It is intuitively that high degree nodes have greater spreading capability, but it is not always the case. The spread capability of a node is also highly relied on the probability of being activated. In addition, our MIS is better than R1 and R2 from the figure since the content spread value is larger.

TABLE I
SET OF OBJECTIVE FUNCTION VALUES OF FACEBOOK

	f_l	f_o	f_u
X_l	329.2	330.3	336.5
X_o	331.5	335.9	339.4
X_u	260.3	264.7	271.3

5) Objective Function Values: Table I represents a set of values for the objective functions by setting the size of candidate edge set $|\bar{X}| = 1000$ and $p = 0.05$ under Facebook data set. Results from all data sets with two different probability settings are very similar. Table I demonstrates the efficiency of both lower bound and upper bound we proposed. We can observe that the expected spread value in each row is getting larger when calculating from lower bound, original to upper bound function. It is also shown in the table that from each column, the solution of the upper bound performs not as good as that of lower or of original. We have discussed the reason of this in the first result analysis part of this section, which is due to the static edge-adding process of upper bound; whereas in lower bound and original problem, we have a dynamic adding edges process.

VII. CONCLUSION

In this paper, we formulate content spread maximization problem from an incremental marginal gain perspective. We reveal the reason why the objective function of this problem lacks submodularity. We derive the submodular upper bound and lower bound of the original problem. Using the sandwich framework, a marginal increment-based algorithm (MIS) that guarantees a data-dependent approximation factor is devised. We also propose a novel algorithm IRFA which is based on influence raking of a single node and fast adjustment according to the recommendations selected in the network. These algorithms calculate the content spread function value as accurate as possible. Simulation results on real social graphs demonstrate the property of realistic network and superiority of our algorithms.

There are several directions on maximizing the content spread problem that deserves further study, for example, if the content propagation can be extended using the linear threshold dissemination model or on a general graph and it is challenging if the network has cycles when we recursively update the spread gain for each node. Since we use sandwich framework, how to find high quality submodular upper and lower bounds

is a very important issue. Finding tighter bounds especially submodular upper bound is still a topic worth studying in the future work.

ACKNOWLEDGMENT

The authors would like to thank the constructive amendments provided by three anonymous reviewers.

REFERENCES

[1] Brandwatch. (2017). *105 Amazing Social Media Statistics and Facts*. Accessed: Dec. 9, 2017. [Online]. Available: https://www.brandwatch.com/blog/96-amazing-social-media-statistics-and-facts-for-2016/

[2] V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt, "Recommendations to boost content spread in social networks," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, p. 529–538.

[3] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the Flickr social network," in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 721–730.

[4] P. Zhang, W. Chen, X. Sun, Y. Wang, and J. Zhang, "Minimizing seed set selection with probabilistic coverage guarantee in a social network," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1306–1315.

[5] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2009, pp. 199–208.

[6] Blog.facebook.com. (2017). *Facebook Newsroom*. Accessed: Dec 9, 2017. [Online]. Available: http://blog.facebook.com/blog.php?post=15610312130

[7] (2017). *Anon*. Accessed: Dec 9, 2017. [Online]. Available: http://blog.twitter.com/2010/07/ discovering-who-to-follow.html

[8] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *Proc. 4th ACM Conf. Rec. Syst.*, Sep. 2010, pp. 199–206.

[9] 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*, Jul. 2010, pp. 1029–1038.

[10] K. Liontis and E. Pitoura. (2016). "Boosting nodes for improving the spread of influence." [Online]. Available: https://arxiv.org/abs/1609.03478v1

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

[12] 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.*, Nov. 2012, p. 245–254.

[13] S. Antaris, D. Rafailidis, and A. Nanopoulos, "Link injection for boosting information spread in social networks," *Social Netw. Anal. Mining*, vol. 4, no. 1, pp. 1–16, 2014.

[14] D. Rafailidis, A. Nanopoulos, and E. Constantinou, "'With a little help from new friends': Boosting information cascades in social networks based on link injection," *J. Syst. Softw.*, vol. 98, pp. 1–8, Dec. 2014.

[15] D. Rafailidis and A. Nanopoulos, "Crossing the boundaries of communities via limited link injection for information diffusion in social networks," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 97–98.

[16] D. Li, Z. Xu, S. Li, X. Sun, A. Gupta, and K. Sycara, "Link recommendation for promoting information diffusion in social networks," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 185–186

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

[18] K. Liontis and E. Pitoura. (2016). "Boosting nodes for improving the spread of influence." [Online]. Available: https://arxiv.org/abs/1609.03478?context=cs

[19] Y. Lin, W. Chen, and J. C. S. Liu. (2017). "Boosting information spread: An algorithmic approach." [Online]. Available: https://arxiv.org/abs/1602.03111

[20] D.-N. Yang, H.-J. Hung, W.-C. Lee, and W. Chen, "Maximizing acceptance probability for active friending in online social networks," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 713–721.

[21] Y. Yang, X. Mao, J. Pei, and X. He, "Continuous influence maximization: What discounts should we offer to social network users?" in *Proc. Int. Conf. Manage. Data.*, Jul. 2016, pp. 727–741.

[22] W. Chen, F. Li, T. Lin, and A. Rubinstein, "Combining traditional marketing and viral marketing with amphibious influence maximization," in *Proc. 16th ACM Conf. Econ. Comput.*, Jun. 2015, pp. 779–796.

[23] Z. Yu, C. Wang, J. Bu, X. Wang, Y. Wu, and C. Chen, "Friend recommendation with content spread enhancement in social networks," *Inf. Sci.*, vol. 309, pp. 102–118, Jul. 2015.

[24] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," in *Proc. VLDB Endowment*, Sep. 2011, pp. 73–84.

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

[26] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2005, pp. 177–187.

[27] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 KDD cup," *ACM SIGKDD Explor. Newslett.*, vol. 5, no. 2, pp. 149–151, Dec. 2003.

[28] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Netw.*, vol. 31, no. 2, pp. 155–163, May 2009.

[29] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 641–650

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the Web," Stanford InfoLab., Stanford, CA, USA Tech. Rep., 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[31] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," *Knowl. Inf. Syst.*, vol. 37, no. 3, pp. 555–584, 2013.

[32] Q. Shi, C. Wang, J. Chen, Y. Feng, and C. Chen, "Location driven influence maximization: Online spread via offline deployment," *Knowl. Based Syst.*, vol. 166, Feb. 2019, pp. 30–41.

[33] Q. Shi, C. Wang, J. Chen, Y. Feng, and C. Chen, "Post and repost: A holistic view of budgeted influence maximization," *Neurocomputing*, vol. 338, Apr. 2019, pp. 92–100.

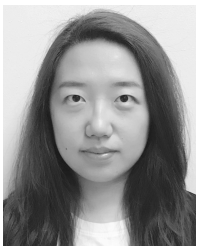
[34] R. Yan, Y. Li, W. Wu, D. Li, and Y. Wang, "Rumor blocking through Online link deletion on social networks," *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 13, no. 2, Apr. 2019, Art. no. 16.

Wenguo Yang received the M.A. degree in operation research and control theory from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in operation research and control theory from the Graduate University of the Chinese Academy of Sciences, Beijing, in 2006.

He is currently a Professor with the School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing. His current research interests include social networks, robust optimization, nonlinear combinatorial optimization, and telecommunication network optimization. He has supervised many M.Sc. and Ph.D. students in these areas.

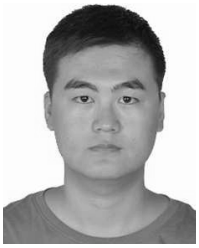
Jianmin Ma received the Ph.D. degree in mathematics from Colorado State University, Fort Collins, CO, USA.

He is currently a Professor of mathematics with the College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, China. His current research interests include machine learning, data mining, social networks, and discrete mathematics.



Yi Li received the M.S. degrees in digital communication and multimedia and computer science and the Ph.D. degree in computer science from the University of Texas at Dallas, Richardson, TX, USA.

Her current research interests include social influence maximization/minimization, information propagation, and data science.



Ruidong Yan received the B.S. degree in information and computing Sciences from Inner Mongolia University, Hohhot, China, in 2014. He is currently pursuing the Ph.D. degree from the Department of Computer Science, Renmin University of China, Beijing, China.

His current research interests include social networks, algorithm design, and analysis.

Jing Yuan received the B.S. degree in computer science from Nanjing University, Nanjing, China, in 2008. She is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA.

Her current research interests include online social networks, cyber physical systems, and cloud computing.

Weili Wu received the Ph.D. and M.S. 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 current research interests include data communication, data management, the design and analysis of algorithms for optimization problems that occur in wireless networking environments, and various database systems.

Deying Li received the B.S. and M.S. degrees in mathematics from Huazhong Normal University, Wuhan, China, in 1985 and 1988, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2004.

She is currently a Professor with the Renmin University of China, Beijing, China. Her current research interests include wireless networks, *ad hoc* and sensor networks mobile computing, distributed network system, social networks, and algorithm design.