ABSTRACT Crowd psychology is a critical factor when considering information diffusion, which has been modeled as composed influence. The composed influence is represented as a hyperedge in a graph model. A hyperedge $e = (H_e, v)$ contains the head node set H_e and tail node v. Then a social network is modeled as a hypergraph. e could only propagate this influence when all nodes in H_e first become all active. In this paper, the Composed Influence Maximization(CIM) also aims to select k initially-influenced seed users in such a social network. The objective is to maximize the expected number of eventually-influenced users. We present an approximating method for this objective function by formulating a series of submodular functions and these functions are convergent. Then, we develop a lower bound and an upper bound problems which objective functions are submodular. We design a greedy strategy based on lower bound maximization for solving CIM. We formulate a sandwich approximation framework, which preserves a theoretical analysis result. Finally, we evaluate our algorithm on real world data sets. The results show the effectiveness and the efficiency of the proposed algorithm.

INDEX TERMS Composed Influence, Influence Maximization, Independent Cascade, Social Networks

I. INTRODUCTION

People's personal behaviours always follow their friends or groups, which is called conformity behaviour [1]. Especially in marketing, when one plans to buy some product, he usually ask his friends' opinions [2]. If more than one friends recommend the same product, he may be in conformity with that. Also this influence from group is different from personal influence, which is called composed influence in this paper. The composed influence is an additional influence, and it is not simple accumulative of personal influence. Then, we model such a social network by combining the personal influence with the composed influence. In which, we consider Influence Maximization (IM) problem aims to select k initially-influenced seed users to maximize the expected number of eventually-influenced users. IM problem has many applications, such as viral marketing [3], epidemic control and assessing cascading failures within complex systems. Then Influence Maximization (IM) in normal directed graph was extended to Composed Influence Maximization (CIM) [4] in social network.

When considering information diffusion in social network, Independent Cascade (IC) model is most popular. An user is active if he receives the information and tries to activate his neighbors (friends). For each inactive user, every active neighbor will try to activate him with a certain active probability. IC model assumes the active events from all his active neighbors are independent. IC model also assumes each activated user has only one chance to activate his inactive neighbors. For example, there are three guys: John, Mike and Bob. They are friends each other. Assume the influence probability from John to Bob is 0.5, while that from Mike to Bob is 0.6. Then Bob will be activated by his 2 friends with the probability of $1 - (1 - 0.5)(1 - 0.6) = 0.8$ when John and Mike are active. Besides the personal influence, the composed influence comes from John and Mike when they become active. Bob will be activated by this composed

influence with a certain probability which is quite different from personal influence. Assume the probability of composed influence is 0.4. Then the active probability of Bob in this model is $1-(1-0.5)(1-0.6)(1-0.4)=0.88$. The composed influence is modeled as a hyperedge and such a social network is modeled as a hypergraph. A hyperedge is defined as a directed edge from a user set to one user in this paper.

This conformity phenomenon leads to non-submodularity of the objective function in CIM which will be proved in [4]. Note that influences through hypergraphs are not submodular, we cannot adapt existing social influence maximization methods to solve the CIM. Therefore, challenges are posed to solve the CIM. The first challenge is to deal with the non-submodularity. New algorithms are needed, since a simple greedy algorithm can no longer guarantee an approximation ratio. Another challenge is the scalability. Since hyperedges change the scalability, it is difficult to reduce their complexities.

A. RELATED WORKS

Kempe et al. [5] were the first to formulate Influence Maximization Problem (IM) as an optimization problem under the IC model. They prove IM to be NP-hard under IC model and design a natural greedy algorithm that yields $(1-1/e-\epsilon)$-approximate solutions for any $\epsilon>0$. Zhu et al. [4] studied the influence maximization problem considering the crowd psychology influence. Motivated by this celebrated work, a fruitful literature for IM [6]–[13] has been developed. However, most of the existing methods are either too slow for billion-scale networks such as Facebook, Twitter and World Wide Web or fail to retain the $(1-1/e-\epsilon)$-approximation guarantees.

TIM/TIM+ [14] and IMM [15] are two scalable methods with $(1-1/e-\epsilon)$-approximation guarantee for IM. Tang et al. [14], [15] utilize a novel Reverse Influence Set (RIS) sampling technique introduced by Borgs et al. [16]. TIM+ and IMM attempt to generate a $(1-1/e-\epsilon)$-approximate solution with minimal numbers of RIS samples. However, they may take days on billion-scale networks.

Later, Nguyen et al. [17] make a breakthrough and proposed two novel sampling algorithms SSA and D-SSA. Unlike the previous heuristic algorithms, SSA and D-SSA are faster than TIM+ and IMM while providing the same $(1-1/e-\epsilon)$-approximate guarantee. SSA and D-SSA are the first approximation algorithms that use minimum numbers of samples, meeting strict theoretical thresholds characterized for IM.

Although there are a large amount of literature for IM, almost all of IM are submodular. Few results [18] are provided when the influence propagation model even slightly violates the submodularity. Note that influences through hypergraphs are not submodular, we cannot adapt existing social influence maximization methods to solve the CIM in this paper. Narasimhan and Bilmes [19] presented an approximation method for submodular+supermodular function by substituting the supermodular function to a modular function. In [20], Bach gave an general idea that any nonsubmodular function could represent as a difference of two submodular function. The latest approach is based on the sandwich [21] approximation strategy, which approximates the objective function by looking for its lower bound and upper bound.

B. CONTRIBUTIONS

Our contributions are summarized as follows:

(1) We develop lower bound and upper bound of the objective function for CIM.

(2) We design an algorithm to estimate the objective of CIM by formulating a series of submodular functions. Theorem 3.3 shows the convergency of these functions.

(3) For solving CIM, Algorithm 3 gives a greedy strategy based on lower bound maximization.

(4) Last, we verify our algorithm on real world data sets. The results show the effectiveness and the efficiency of the proposed algorithm.

The rest of this paper is organized as follows. In Section 2, we formulate a lower bound, upper bound and design an approximation scheme of the objective function. Algorithms for solving CIM are designed in Section 3. Experiment results are shown in Section 4 and we draw a conclusion in Section 5. Table 1 summarizes the frequently used symbols and their meaning.

TABLE 1: Frequently used notation

Notation	Description
$G=(V,E,P)$	A social network, where V is the node set, $E=E_N\cup E_H$ is edge set. E_N is the normal directed edge set and E_H is the directed hyperedge set. Each directed hyperedge in E_H has at least 2 head nodes. Each edge is associated with an influence probability P.
$e=(H_e,v)$	e is a directed hyperedge when $\|H_e\|\geq 2$, while e is a normal directed edge when $\|H_e\|=1$. H_e is called the head and v is called the tail.
$G-E$	A subgraph with the same node set of G by deleting edge set E from G
$G+E$	A graph with the same node set of G by adding edge set E to G
$m=\|E_H\|$	The number of hyperedges
$n=\|V\|$	The number of nodes in G
$E(P)$	An edge set with influence probability P on each edge in E
$\sigma(S\|G)$	The expected number of eventually-influenced nodes with initial seed set S in social network G under diffusion model.

II. PROBLEM FORMULATION

Given a directed social network with composed influence $G=(V,E,P)$, where V is a set of nodes (i.e., users in an OSN), E is a set of directed hyperedges and P is the weight function on edge set E. For each directed hyperedge $e=(H_e,v)$, where H_e is a subset of V and denotes $H_e=\{u_1,u_2,\ldots,u_k\}$, P_e denotes the weight of e, representing the influence propagation probability ($0\leq P_e\leq 1$).

A. COMPOSED INFLUENCE

Composed influence comes from a group of users. When given a $G = (V, E, P)$, we model all relations between users as hyperedges. The hyperedges represent influence propagation directions, including personal and composed influences. For a hyperedge $e = (H_e, v)$, let H_e denote its head set of nodes and v be the tail node (i.e., e connects nodes in H_e to node v). If H_e contains only one node u, it means e is a normal directed edge and the influence is personal. While H_e contains more than one node, the hperedge e means there is crowd influence from H_e to v. Let P_e denote the weight of e, representing the influence propagation probability ($0 \leq P_e \leq 1$). Specifically, P_e is the probability that v is activated by H_e after each node in H_e is activated. For simplification, we separate E into two parts: normal directed edges E_N and directed hyperedges with at least two head nodes E_H.

B. INDEPENDENT CASCADE MODEL

The Independent Cascade (IC) model is the most widely used information diffusion model. Our Composed Influence Maximization (CIM) Problem is based on IC model.

IC model assumes a seed set $S \subseteq V$. Let S_t be the nodes that are activated in step $t (t = 0, 1, \ldots)$ and $S_0 = S$. For a hyperedge $e = (H_e, v)$, e is activated for the first time at step t only if $H_e \subseteq S_t$ and $H_e \setminus S_{t-1} \neq \emptyset$. The diffusion process is as follows. At step t, each activated hyperedge $e = (H_e, v)$ for the first time has only one chance to activate the inactivated node v with the probability of P_e. Note that a hyperedge e could only propagate the influence when all nodes in H_e first become all active. If H_e includes an inactive node, e cannot propagate the influence since hyperedges represent crowd influences. All seed nodes in S are initially-influenced, while all active nodes at the end of the process are eventually influenced. We use $\sigma(S|G)$ to denote the expected number of eventually-influenced nodes.

C. COMPOSED INFLUENCE MAXIMIZATION PROBLEM

The Composed Influence Maximization (CIM) Problem $G = (V, E, P)$, the objective is to select k initially-influenced seed users to maximize the expected number of eventually-influenced users:

$$\max \sigma(S|G) \tag{1}$$

$$s.t. |S| \leq k \tag{2}$$

Where S is the initial seed set and $\sigma(S|G)$ the expected number of eventually-influenced nodes.

D. PROPERTIES OF COMPOSED INFLUENCE MAXIMIZATION PROBLEM

It is known that any generalization of a NP-hard problem is also NP-hard. The Composed Influence Maximization (CIM) Problem was proved to be NP-hard in [4]. The social influence maximization problem is a special case of CIM when the head set H_e just contains one node. Also, we can get the following result of computing $\sigma(S)$.

A set function $f : 2^V \leftarrow \mathbb{R}$ is said to be submodular [22] if for all subsets $S, T \in V$, it holds that $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. While if for all subsets $S, T \in V$, it holds that $f(S) + f(T) \leq f(S \cup T) + f(S \cap T)$, f is supermodular.

The objective function of CIM is neither submodular nor supermodular under IC model [4].

III. UPPER BOUND AND LOWER BOUND

Since the objective function $\sigma(S|G)$ is nonsubmodular and nonsupermodular, the greedy strategy can not guarantee an efficient approximation bound. We will introduce the sandwich approximation framework to bound the approximation by an upper bound function and a lower bound function. At the same time, we try to reduce the gap between upper bound and lower bound.

A. UPPER BOUND

For each hyperedge, CIM requires all nodes H_e are activated when try to activate the tail node with a given probability. Now we construct an upper bound problem, which only needs at least one node in H_e is activated. Furthermore, if specify a certain node u in H_e, H_e will try to activate the tail node after u is activated. It can be shown this problem is also an upper bound and it is more tight upper bound especially.

Given an instance of CIM $G = (V, E, P)$, assume $E_H = \{(H_1, v_1), (H_2, v_2), \ldots, (H_m, v_m)\}$ is the hyperedge set, where $H_i = \{u_{i1}, u_{i2}, \ldots, u_{ik_i}\}$ is the head node set for hyperedge (H_i, v_i) for $i = 1, 2, \ldots, m$. Particularly, the number of nodes in H_i is at least 2 according to our assumption. We generate new directed edges $E_i = \{(u_{i1}, v_i), (u_{i2}, v_i), \ldots, (u_{ik_i}, v_i)\}$ for $i = 1, 2, \ldots, m$. Let $E' = E_1 \cup E_2 \cup \ldots \cup E_m$. Randomly picking exactly one edge $e_i = (u_{ij_i}, v_i)$ from each E_i, and let $E^* = \{e_i | i = 1, 2, \ldots, m\}$ be the collection of these edges. For each $e_i \in E^*$, $P_{e_i} = P_{(H_i, v_i)}$ is the influence probability from u_{ij_i} to v_i. If there already have some edges in E which are the same as in E^*, they will be considered as multi-edge and activate the their tail nodes under IC model. Let $G_0 = G - E_H$ and $\sigma(S|G_0 + E^*)$ denote the totally expected number of activated nodes when spread in $G_0 + E^*$ with initial seed set S under IC model. Since, $G_0 + E^*$ is a graph without hyperedges, $\sigma(S|G_0 + E^*)$ is monotone and submodular [5]. Moreover, we have the following result.

Theorem 3.1: Given an instance of CIM $G = (V, E, P)$ and a seed set S, where $E = E_N \cup E_H$. Let $G_0 = G - E_H$, then we have $\sigma(S|G_0 + E^*) \geq \sigma(S|G)$.

Proof 3.1: Given an instance of CIM $G = (V, E, P)$ and a seed set S, let $f_{G_0+E^*}(v)$ and $f_G(v)$ denote the eventually activated probability of v in $G_0 + E^*$ and G respectively. We only need to prove $f_{G_0+E^*}(v) \geq f_G(v)$.

Assume $E(v)$ is the edge set in $G_0 + E^*$ that contains all edges with tail node v. $E(v)$ can be divided into two parts $E_0(v)$ and $E^*(v)$, where $E_0(v)$ belong to $G_0 + E^*$ and $E^*(v) \subseteq E^*$. $E'(v)$ is the edge set in G that contains all edges with tail node v. Then, $E'(v)$ can also divided into

two parts according to the number of nodes in their head set $E'_1(v) = \{(H_e, v)||H_e| = 1\}$ and $E'_2(v) = \{(H_e, v)||H_e| > 1\}$. Under IC model, the diffusion process is step by step. In the first step, S is the seed set. In $G_0 + E^*$, v will be activated by nodes in $\{u|(u, v) \in E(v)\} \cap S$ with probability P. In G, v will be activated by nodes in $\{u|(u, v) \in E'(v)\} \cap S$ with probability P. The difference between this two set is edges in $E^*(v)$ and $E'_2(v) = \{(H_e, v)||H_e| > 1\}$. If one edge e' in $E'_2(v)$ is activated, all nodes in head set $H_{e'}$ of e' must belong to S. On the other hand, there must exist $(u_{ij}, v) \in E^*(v)$ where $u_{ij} \in H_{e'}$. Obviously, $u_{ij} \in S$ and $P_{(u_{ij},v)} = P_{e'}$. Then we have $f_{G_0+E^*}(v) \geq f_G(v)$. In the following step, the activated nodes in G are always a subset of that in $G_0 + E^*$. $f_{G_0+E^*}(v) \geq f_G(v)$ must keep after whole diffusion process.

From Theorem 3.1, E^* are picked arbitrarily. Meanwhile, we may gain the better one when $\min_{E^*} \sigma(S|G_0 + E^*)$. Then, a better upper bound will be obtained.

B. LOWER BOUND

If the composed influences are omit, which means all hyperedges are deleted from CIM, then we obtain a trivial lower bound for CIM. Obviously, this lower bound function is monotone and submodular.

C. APPROXIMATION SCHEME OF THE OBJECTIVE FUNCTION

In this subsection, we will present an approximation scheme for the objective function in CIM, then RIS method can be applied, which has been proved an efficient method in computing the objective function of Influence Maximization problems. One big challenge of computing objective function for CIM comes from the hyperedge structure which makes the RIS method unusable. Our basic idea is to formulate a series of functions which is convergence to the objective function in CIM for any given seeds set S. RIS method can be applied in each one of these functions. And the number of these function may be theoretically bounded.

Given an instance of CIM $G = (V, E, P)$, assume $E_H = \{(H_1, v_1), (H_2, v_2), \ldots, (H_m, v_m)\}$ is the hyperedge set, where $H_i = \{u_{i1}, u_{i2}, \ldots, u_{ik_i}\}$ is the head node set for hyperedge (H_i, v_i) for $i = 1, 2, \ldots, m$. Particularly, the number of nodes in H_i is at least 2 according to our assumption. We generate new directed edges $E_i = \{(u_{i1}, v_i), (u_{i2}, v_i), \ldots, (u_{ik_i}, v_i)\}$ for $i = 1, 2, \ldots, m$. Let $E' = E_1 \cup E_2 \cup \ldots \cup E_m$. Randomly pick exactly one edge $e_i = (u_{ij_i}, v_i)$ from each E_i, and let $E^* = \{e_i|i = 1, 2, \ldots, m\}$ be the collection of these edges.

Let $G_0 = G - E_H$, and assume $f_0(v)$ is the expected activation probability of v when spreading the influence on G_0 for a given seed set S. For each $e_i \in E^*$, the influence probability $P^1_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_0(u_{ij_i})$. Let $E^*(P^1)$ denote the edge set with influence probability $P^1_{e_i}$ on each edge. Then we have the following theorem.

Theorem 3.2: Given an instance of CIM $G = (V, E, P)$ and a seed set S, where $E = E_N \cup E_H$. Let $G_0 = G - E_H$, then we have $\sigma(S|G_0 + E^*(P^1)) \leq \sigma(S|G)$.

Proof 3.2: Given an instance of CIM $G = (V, E, P)$ and a seed set S, let $f_{G_0+E^*(P^1)}(v)$ and $f_G(v)$ denote the eventually activated probability of v in $G_0 + E^*(P^1)$ and G respectively. We only need to prove $f_{G_0+E^*(P^1)}(v) \leq f_G(v)$. For simplicity, let $G' = G_0 + E^*(P^1)$.

Assume $f_0(v)$ is the expected activation probability of v when spread the influence on G_0 for a given seed set S. $f_0(v) \leq f_G(v)$ $f_0(v) \leq f_{G'}(v)$ are obviously true since G_0 is a subgraph of G and subgraph of G'. For any node v, let $E_1(v)$ and $E_2(v)$ be the incoming normal directed edge set and hyperedge set respectively in G. Since the diffusion model in this paper is based on IC model, which means the influence from different neighbors are independent. Then, the expected activation probability can be estimated as follows

$$f_G(v) = 1 - \prod_{(u,v)\in E_1(v)} (1 - P_{(u,v)} f_G(u)) \times \prod_{(H_e,v)\in E_2(v)} (1 - P_{(H_e,v)} \prod_{u\in H_e} f_G(u)) \tag{3}$$

In $G_0 + E^*(P^1)$, there are only normal directed edges left according to the formulation process. We can still divide the incoming normal directed edge set of v into two parts: edges in G_0 and in $E^*(P^1)$. Let $E'_1(v)$ and $E'_2(v)$ denote these two parts respectively. Then,

$$f_{G'}(v) = 1 - \prod_{(u,v)\in E'_1(v)} (1 - P_{(u,v)} f_{G'}(u)) \times \prod_{(u,v)\in E'_2(v)} (1 - P^1_{(u,v)} f_{G'}(u)) \tag{4}$$

Obviously, $E_1(v) = E'_1(v)$ since G_0 is a subgraph of G. For each edge $(u, v) \in E'_2(v)$, there must exist $(H_e, v) \in E_2(v)$ such that $u \in H_e$ according to the definition of E^*. Then, $P^1_{(u,v)} = P_{(H_e,v)} \cdot \prod_{u'\in H_e, u'\neq u} f_0(u')$. Equation 4 is as follows:

$$f_{G'}(v) = 1 - \prod_{(u,v)\in E'_1(v)} (1 - P_{(u,v)} f_{G'}(u)) \times \prod_{(u,v)\in E'_2(v)} (1 - P_{(H_e,v)} \cdot \prod_{u'\in H_e, u'\neq u} f_0(u') f_{G'}(u)) \tag{5}$$

Under IC model, the diffusion process is step by step. In the first step, S is the seed set. For each node $u \in S$, $f_0(u) = f_{G'}(u) = f_G(u) = 1$, then $f_{G'}(v) = f_G(v)$ for any inactivated node v after the fist step. While the following steps, $f_{G'}(v) \leq f_G(v)$ always keeps since we combine Equation 5 with $f_0(u) \leq f_{G'}(u)$.

Furthermore, we can update the influence probability on each edge in E^* according to the output of the expected activation probability of each node v, i.e. $f_1(v)$. Then, for each $e_i \in E^*$, the influence probability $P^2_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_1(u_{ij_i})$ and formulate a new IM problem $G_0 + E^*(P^2)$. A series of IM problems can be generated according to the updated activation probability, and the following theorem shows their convergence to the objective function of CIM for a given S.

Theorem 3.3: Given an instance of CIM $G = (V, E, P)$ and a seed set S, where $E = E_N \cup E_H$. $f_{t-1}(v)$ is the expected activation probability of v when spread the influence on $G_0 + E^*(P^{t-1})$ for the given seed set S. For each $e_i = (u_{ij_i}, v_i) \in E^*$, the influence probability $P^t_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_{t-1}(u_{ij_i})$ and then generate IM problem $G_0 + E^*(P^t)$. Then,

$$\lim_{t\to\infty} \sigma(S|G_0 + E^*(P^t)) = \sigma(S|G) \tag{6}$$

Where the iterations are at most $m = |E_H|$.

Proof 3.3: Given an instance of CIM $G = (V, E, P)$ and a seed set S, let $f_{G_0+E^*(P^t)}(v)$ and $f_G(v)$ denote the eventually activated probability of v in $G_0 + E^*(P^t)$ and G respectively. We need to prove the following three statements is true for any t:

$$f_{G_0+E^*(P^t)}(v) \leq f_G(v) \tag{7}$$

$$f_{G_0+E^*(P^t)}(v) \leq f_{G_0+E^*(P^{t+1})}(v) \tag{8}$$

And

$$f_{G_0+E^*(P^m)}(v) = f_G(v) \tag{9}$$

For simplicity, let $G^t = G_0 + E^*(P^t)$.

Firstly, $f_{G^1}(v) \leq f_G(v)$ is obviously true according to Theorem 3.2. Assume $f_{G^t}(v) \leq f_G(v)$, we will prove $f_{G^{t+1}}(v) \leq f_G(v)$. For any node v, let $E_1(v)$ and $E_2(v)$ be the incoming normal directed edge set and hyperedge set respectively in G. Then, the expected activation probability can be estimated as follows

$$\begin{aligned} f_G(v) = 1 - &\prod_{(u,v)\in E_1(v)} (1 - P_{(u,v)} f_G(u)) \\ \times &\prod_{(H_e,v)\in E_2(v)} (1 - P_{(H_e,v)} \prod_{u\in H_e} f_G(u)) \end{aligned} \tag{10}$$

In $G_0 + E^*(P^{t+1})$, there are only normal directed edges left according to the formulation process. We can still divide the incoming normal directed edge set of v into two parts: edges in G_0 and in $E^*(P^{t+1})$. Let $E'_1(v)$ and $E'_2(v)$ denote these two parts respectively. Then,

$$\begin{aligned} f_{G^{t+1}}(v) = 1 - &\prod_{(u,v)\in E'_1(v)} (1 - P_{(u,v)} f_{G^{t+1}}(u)) \\ \times &\prod_{(u,v)\in E'_2(v)} (1 - P^t_{(u,v)} f_{G^{t+1}}(u)) \end{aligned} \tag{11}$$

Obviously, $E_1(v) = E'_1(v)$ since G_0 is a subgraph of G. For each edge $(u, v) \in E'_2(v)$, there must exist $(H_e, v) \in E_2(v)$ such that $u \in H_e$ according to the definition of E^*. Then, $P^t_{(u,v)} = P_{(H_e,v)} \cdot \prod_{u'\in H_e, u'\neq u} f_t(u')$. Equation 11 is as follows:

$$\begin{aligned} f_{G^{t+1}}(v) = 1 - &\prod_{(u,v)\in E'_1(v)} (1 - P_{(u,v)} f_{G^{t+1}}(u)) \\ \times &\prod_{(u,v)\in E'_2(v)} (1 - P_{(H_e,v)} \cdot \prod_{u'\in H_e, u'\neq u} f_{G^t}(u') f_{G^{t+1}}(u)) \end{aligned} \tag{12}$$

Under IC model, the diffusion process is step by step. In the first step, S is the seed set. For each node $u \in S$, $f_{G^t}(u) = f_{G^{t+1}}(u) = f_G(u) = 1$, then $f_{G^{t+1}}(v) = f_G(v)$ for any inactivated node v after the first step. While the following steps, $f_{G^{t+1}}(v) \leq f_G(v)$ always keeps since we combine Equation 12 with $f_{G^t}(u) \leq f_G(u)$. Then, $f_{G_0+E^*(P^t)}(v) \leq f_G(v)$ for any t.

Secondly, assume $f_0(v)$ is the expected activation probability of v when spread the influence on G_0 for a given seed set S. $f_0(v) \leq f_{G^1}(v)$ are obviously true since G_0 is a subgraph of G^1. Assume $f_{G^{t-1}}(v) \leq f_{G^t}(v)$, we will prove $f_{G^t(v)} \leq f_{G^{t+1}}(v)$. In G^{t+1}, $P^{t+1}_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_t(u_{ij_i})$ and $P^t_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_{t-1}(u_{ij_i})$ in G^t. Then, we have $P^t_{e_i} \leq P^{t+1}_{e_i}$. That means the influence probability on each directed edge in G^{t+1} is at least that of G^t. $f_{G_0+E^*(P^t)}(v) \leq f_{G_0+E^*(P^{t+1})}(v)$ is true for any t.

Finally, we will prove the iterations are at most m. Since the activated nodes can only try to activate their inactivated neighbors once, the influence diffusion path from each node will be acyclic. According to the definition of IM problem G^t, only edges in E^* will change at each iteration $P^t_{e_i} = P_{(H_i,v_i)} \cdot \prod_{j=1,j\neq j_i}^{j=k_i} f_{t-1}(u_{ij_i})$. If $f_{t-1}(u)$ does not change for all nodes in H_i, then $P^t_{e_i} = P^{t-1}_{e_i}$. Since the influence diffusion is acyclic, once $f_{t-1}(u)$ dose not change in iteration t, $f_{t-1}(u)$ will never change in the following iterations. In the first iteration, assume hyperedge (H_e, v) is closest to seed set S, that means each node in H_e is not a tail of another hyperedge. Then, $f_1(u) = f_0(u)$ for each $u \in H_e$ and $P^1_{(H_e,v)} = P^0_{(H_e,v)}$. If there does not exist another hyperedge in which v is the tail, otherwise we can used the same strategy as stated above, then $f_2(v) = f_1(v)$. And so on, there must exist at least one hyperedge where each node in the head set of this hyperedge will be unchangeable. Then there are at most m iterations where m is the number of hyperedges in G.

IV. ALGORITHM

A. (ϵ, δ)-APPROXIMATION

We recall the (ϵ, δ)-approximation in [23] that will be used in our algorithm. ϵ is relative error of estimation and $(1 - \delta)$ is confidence.

Definition 4.1: ((ϵ, δ)-approximation). Let $Z_1, Z_2, \ldots$ be independently and identically distributed samples according to Z in the interval [0,1] with mean μ_z and variance σ^2_Z. A Monte Carlo estimator of μ_z,

$$\hat{\mu}_Z = \frac{1}{T} \sum_{i=1}^{T} Z_i \tag{13}$$

is said to be an (ϵ, δ)-approximation of μ_Z if

$$Pr[(1 - \epsilon)\mu_Z \leq \hat{\mu}_Z \leq (1 + \epsilon)\mu_Z] \geq 1 - \delta \tag{14}$$

Define $\Upsilon = 4(e-2)\ln(2/\delta)/\epsilon^2$ and $\Upsilon_1 = 1+(1+\epsilon)\Upsilon$, then the Stopping Rule Algorithm given in [16] has been proved to be (ϵ, δ)-approximation.

Lemma 4.1: Let $Z_1, Z_2, \ldots$ be independently and identically distributed samples according to Z in the interval [0,1] with mean μ_z. Let $SumZ = \sum_{i=1}^{N} Z_i$, $\hat{\mu}_Z = \frac{SumZ}{N}$, $\Upsilon = 4(e-2)\ln(2/\delta)/\epsilon^2$ and $\Upsilon_1 = 1 + (1+\epsilon)\Upsilon$. If N is the number of samples when $SumZ \geq \Upsilon_1$, then $Pr[(1-\epsilon)\mu_Z \leq \hat{\mu}_Z \leq (1+\epsilon)\mu_Z] \geq 1-\delta$ and $\mathbb{E}[N] \leq \Upsilon_1/\mu_Z$.

B. INFLUENCE ESTIMATION METHOD IN CIM

The Reverse Influence Set (RIS) sampling method [16] cannot applied directly to CIM problem since reverse reachable probability can not be used to estimate $\sigma(\cdot)$. On the other hand, Mento Carlo method always cost too much running time [4]. Then, we will introduce a new method to estimate the influence when given a seed seed S for CIM problem. The main idea comes from Theorem 3.3.

Firstly, we propose an estimation procedure for activation probability of each node when given the initial seed set S in a normal directed social network G. For each inactivated node v, we define a reverse reachable set (R_v) as follows [16]: (1) generating a sample graph g from G; (2) returning R_v as the set of nodes that can reach v in g. Let $\mathcal{R}_v$ denote a set of random reverse reachable set for node v. Let $Cov_{\mathcal{R}_v}(S)$ denote the number of reverse reachable sets in $\mathcal{R}_v$ that intersection with S is not empty. Then, we define an estimator of active probability for any node v as $\widehat{f}(v) = Cov_{\mathcal{R}_v}(S)/|\mathcal{R}_v|$. According to Lemma 4.1, for each node v, we need to generate $N = |\mathcal{R}_v(S)|$ reverse reachable sets that satisfy the stopping rule $Cov_{\mathcal{R}_v}(S) \geq 1 + 4(1+\epsilon)(e-2)\ln(2/\delta)/\epsilon^2$. Then, we present the following estimation procedure Algorithm 1 which will output $\widehat{f}(v)$ as an estimation of activation probability $f(v)$.

Algorithm 1 Activation Probability Estimation (APE)

Input: a directed social network $G = (V, E, P)$ without hyperedges, $0 \leq \epsilon, \delta \leq 1$, seed set S, an inactive node v.
Output: $\widehat{f}(v)$ for each node such that $\widehat{f}(v) \leq (1+\epsilon)f(v)$ with at least $(1-\delta)$-probability.

$\Upsilon_1 = 1 + 4(1+\epsilon)(e-2)\ln(2/\delta)/\epsilon^2$
$\mathcal{R}_v \leftarrow$ generate Υ_1 random reverse reachable sets
while $Cov_{\mathcal{R}_v}(S) \leq \Upsilon_1$ **do**
  $R \leftarrow$ generate a random reverse reachable set for node v
  Add R to $\mathcal{R}_v$
end while
$\widehat{f}(v) \leftarrow Cov_{\mathcal{R}_v}(S)/|\mathcal{R}_v|$
return $\widehat{f}(v)$

Secondly, we will formulate a group of IM problems which are lowerbounds for the original CIM. Given an instance of CIM $G = (V, E, P)$, assume $E_H = \{(H_1, v_1), (H_2, v_2), \ldots, (H_m, v_m)\}$ is the hyperedge set, where $H_i = \{u_{i1}, u_{i2}, \ldots, u_{ik_i}\}$ is the head node set for hyperedge (H_i, v_i) for $i = 1, 2, \ldots, m$. Particularly, the number of nodes in H_i is at least 2 according to our assumption. We generate new directed edges $E_i = \{(u_{i1}, v_i), (u_{i2}, v_i), \ldots, (u_{ik_i}, v_i)\}$ for $i = 1, 2, \ldots, m$. Let $E' = E_1 \cup E_2 \cup \ldots \cup E_m$. Randomly pick exactly one edge $e_i = (u_{ij_i}, v_i)$ from each E_i, and let $E^* = \{e_i | i = 1, 2, \ldots, m\}$ be the collection of these edges.

Then we present the following influence estimation procedure Algorithm 2 for CIM based on Theorem 3.3.

Algorithm 2 Influence Estimation Procedure (IEP)

Input: Given an instance of CIM $G = (V, E, P)$, $E = E_N \cup E_H$, $0 \leq \epsilon, \delta \leq 1$ and seed set S.
Output: $\widehat{\sigma}(S|G)$ such that $\widehat{\sigma}(S|G) \leq (1+\epsilon)\sigma(S|G)$ with at least $(1-\delta)$-probability.

$m = |E_H|$; $G_0 = G - E_H$
for each node v **do**
  $\widehat{f}(v) \leftarrow$ APE($G_0, \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, S, v$)
end for
Generate new directed edges $E_i = \{(u_{i1}, v_i), (u_{i2}, v_i), \ldots, (u_{ik_i}, v_i)\}$ for $i = 1, 2, \ldots, m$, where $u_{ij} \in H_i$
Randomly pick exactly one edge $e_i = (u_{ij_i}, v_i)$ from each E_i, and let $E^* = \{e_i | i = 1, 2, \ldots, m\}$
for $t = 0$ to $t = m$ **do**
  $P^t_{e_i} = P_{(H_i, v_i)} \cdot \prod_{j=1, j \neq j_i}^{j=k_i} \widehat{f}(u_{ij_i})$, for each $e_i \in E^*$
  for each node v **do**
    $\widehat{f}(v) \leftarrow$ APE($G_0 + E^*(P^t), \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, S, v$)
  end for
end for
$\widehat{\sigma}(S|G) \leftarrow \sum_{v \in V} \widehat{f}(v)$
return $\widehat{\sigma}(S|G)$

C. GREEDY STRATEGY FOR CIM BASED ON LOWER BOUNDS MAXIMIZATION

In this subsection, we will present a greedy strategy for CIM based on updating the influence probabilities on each edge in E^*. At each iteration, an IM problem will be solved by RIS method. Algorithm D-SSA [4] will apply in our algorithm.

Algorithm 3 may not cost too much time since some data at first iteration can be reused in the following iterations and we only need to estimate the activation probability for node in head set of hyperedges. According to Lemma 4.1, define $\Upsilon = 4(e-2)\ln(2/\delta)/\epsilon^2$ and $\Upsilon_1 = 1+(1+\epsilon)\Upsilon$ for given ϵ, δ. The sample complexity of Algorithm 3 is $O(mn\Upsilon_1)$ where n is the number of nodes and m is the number of hyperedges.

D. COMPARISON EXPERIMENTS

Experiment results will be shown in the next section by comparing our algorithm with others. Firstly, we will present several iterations of Algorithm 2 and compare with the upper bound. Secondly, an heuristic strategy, by selecting the first k Maximum Out-degree (MO) nodes as the seed set, will be presented for comparison with our algorithm.

V. EXPERIMENTS

In this paper, we have used three datasets [4] and one dataset in [24]. Each dataset had both one mode and two mode data

Algorithm 3 Greedy Strategy for CIM based on Lower Bounds Maximization (GSLBM)

Input: Given an instance of CIM $G = (V, E, P)$, $E = E_N \cup E_H$, $0 \leq \epsilon, \delta \leq 1$ and k.
Output: a set of seed nodes, S_k.

$m = |E_H|$; $G_0 = G - E_H$
$S_k^0 \leftarrow$ D-SSA($G_0, \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, k$)
$\widehat{f}(v) \leftarrow$ APE($G_0, \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, S_k^0$)
for $t = 0$ to $t = m$ **do**
  Generate new directed edges $E_i = \{(u_{i1}, v_i), (u_{i2}, v_i), \ldots, (u_{ik_i}, v_i)\}$ for $i = 1, 2, \ldots, m$, where $u_{ij} \in H_i$
  Randomly pick exactly one edge $e_i = (u_{ij_i}, v_i)$ from each E_i, and let $E^* = \{e_i | i = 1, 2, \ldots, m\}$
  $P_{e_i}^t = P_{(H_i, v_i)} \cdot \prod_{j=1, j \neq j_i}^{j=k_i} \widehat{f}(u_{ij_i})$, for each $e_i \in E^*$
  $S_k^t \leftarrow$ D-SSA($G_0 + E^*(P^t), \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, k$)
  for each node v **do**
    $\widehat{f}(v) \leftarrow$ APE($G_0 + E^*(P^t), \frac{\epsilon}{m+1}, \frac{\delta}{m+1}, S_k^t, v$)
  end for
end for
return $S_k \leftarrow S_k^m$

which were preprocessed to give the edges (both simple and hyperedges) of the graph on which the experiments were performed. The first three datasets were a Facebook-like Forum network (FF), Newmanąfs Scientific Collaboration network (NSC) and Norwegian Interlocking Directorate (NID) [4]. The fourth dataset is TheMarker [24].

All datasets contain both one mode and two mode data or group information. For Facebook-like Forum Network dataset, we found the composed influence from the topics and got 479 hyperedges. There were 95188 normal edges and 3668 hyperedges in dataset NSC, and 7710 normal edges and 4977 hyperedges in dataset NID. While we got 13263 hyperedges for dataset TheMarker. All the programs were written in Python 3.6.3 and run on a Linux server with 16 CPUs and 251GB RAM. The data statistics are compiled in the Table 2 below.

TABLE 2: Data Statistics

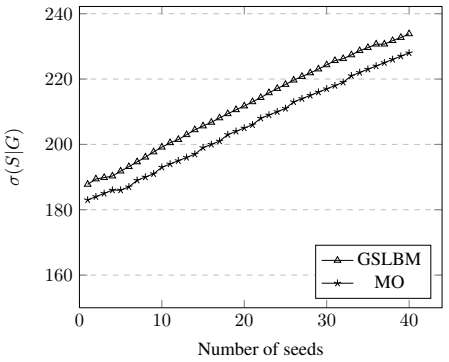
	Normal Directed Edges	Hyperedges	Nodes
Dataset FF	142760	479	897
Dataset NSC	95188	3668	16264
Dataset NID	7710	4977	2045
TheMarker	1048576	13263	65097

The influence probability on each hyperedge was generated according to the in-degree of the tail node.

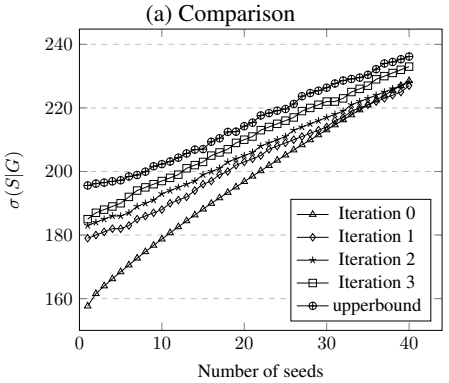
A. EFFECTIVENESS

The graphs in Figure 1, 2, 3, 4 show the experimental results for Dataset FF, NSC, NID, and TheMarker respectively. We run Algorithm 3 to get the seed set S. Then, compare the objective of S with that of seed set from Heuristic strategy MO. This results are shown in graph (a) for each dataset. Secondly, we show the the convergency of Algorithm 2 for seed set S and compare with the upper bound proposed in section III-A. From all these experiments, we can get the following results.

Graph (a) shows our algorithm is much better than the heuristic strategy. Graph (b) shows the convergency of Algorithm 2 for given seed set S and it is not too far from the upper bound of this problem.



(a) Comparison

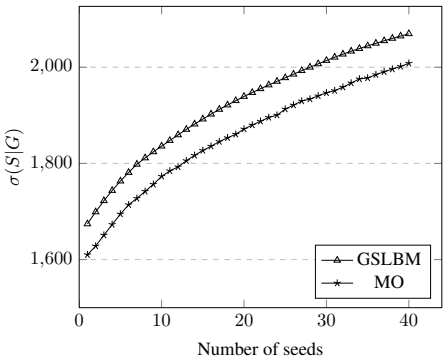


(b) Convergency of Algorithm IEP

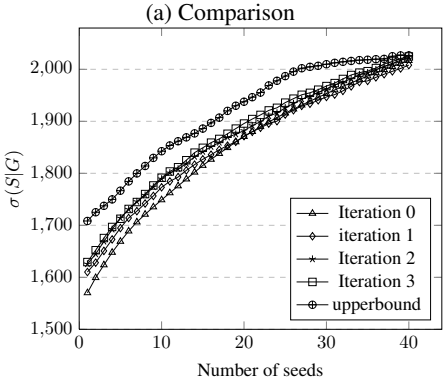
FIGURE 1: Experimental Results for Dataset FF

VI. CONCLUSIONS

In this paper, we modeled the crowd influence in information diffusion process by using a hyperedge. The Composed Influence Maximization (CIM) was formulated to select initially-influenced seed users under Independent Cascade (IC) model to maximize the expected number of eventually-influenced users. We showed CIM was NP-hard and the objective function was neither submodular nor supermodular. An algorithm to estimate the objective of CIM by formulating a series of submodular functions was designed and he convergency of these functions was proved. We developed lower bound and upper bound of the objective function. For solving CIM, a greedy strategy based on lower bound maximization was presented. Finally, the experiment results showed effectiveness and the efficiency of the proposed algorithm. For future research, we are looking for an efficient method to solve nonsubmodular problems, such as difference of submodular

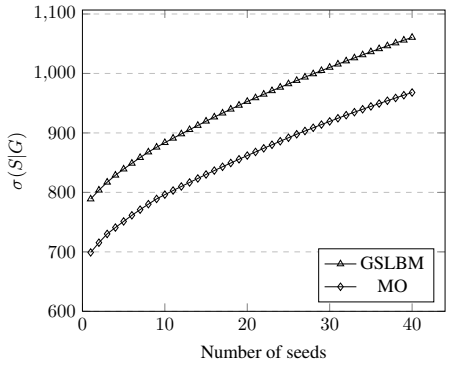


(a) Comparison

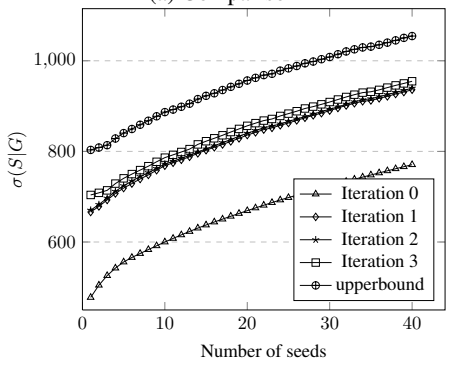


(b) Convergency of Algorithm IEP

FIGURE 2: Experimental Results for Dataset NSC



(a) Comparison



(b) Convergency of Algorithm IEP

FIGURE 3: Experimental Results for Dataset NID

decomposition since every nonsubmodular function can be written as a difference of two submodular functions.

REFERENCES

[1] R. B. C. And and N. J. Goldstein, "Social influence: Compliance and conformity," Annual Review of Psychology, vol. 55, no. 1, p. 591, 2004.

[2] M. Edelson, T. Sharot, R. J. Dolan, and Y. Dudai, "Following the crowd: brain substrates of long-term memory conformity," science, vol. 333, no. 6038, pp. 108–111, 2011.

[3] H. Nguyen and R. Zheng, "On budgeted influence maximization in social networks," IEEE Journal on Selected Areas in Communications, vol. 31, no. 6, pp. 1084–1094, 2013.

[4] J. Zhu, J. Zhu, S. Ghosh, W. Wu, and J. Yuan, "Social influence maximization in hypergraph in social networks," IEEE Transactions on Network Science and Engineering, pp. 1–1, 2018.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003, pp. 137–146.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp. 420–429.

[7] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 1029–1038.

[8] A. Goyal, W. Lu, and L. V. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011, pp. 211–220.

[9] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014, pp. 629–638.

[10] N. Ohsaka, T. Akiba, Y. Yoshida, and K.-i. Kawarabayashi, "Fast and accurate influence maximization on large networks with pruned monte-carlo simulations." in AAAI, 2014, pp. 138–144.

[11] N. Du, Y. Liang, M.-F. Balcan, M. Gomez-Rodriguez, H. Zha, and L. Song, "Scalable influence maximization for multiple products in continuous-time diffusion networks," Journal of Machine Learning Research, vol. 18, no. 2, pp. 1–45, 2017.

[12] Y. Yang, Z. Lu, V. O. Li, and K. Xu, "Noncooperative information diffusion in online social networks under the independent cascade model," IEEE Transactions on Computational Social Systems, vol. 4, no. 3, pp. 150–162, 2017.

[13] C. Aslay, L. V. Lakshmanan, W. Lu, and X. Xiao, "Influence maximization in online social networks," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018, pp. 775–776.

[14] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014, pp. 75–86.

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

[16] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. SIAM, 2014, pp. 946–957.

[17] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in Proceedings of the 2016 International Conference on Management of Data. ACM, 2016, pp. 695–710.

[18] H.-J. Hung, H.-H. Shuai, D.-N. Yang, L.-H. Huang, W.-C. Lee, J. Pei, and M.-S. Chen, "When social influence meets item inference," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016, pp. 915–924.

[19] M. Narasimhan and J. A. Bilmes, "A submodular-supermodular procedure