Influence Maximization Problem: properties and algorithms

Wenguo Yang
School of Mathematics,UCAS
Joint work with
Yapu Zhang;Dingzhu Du

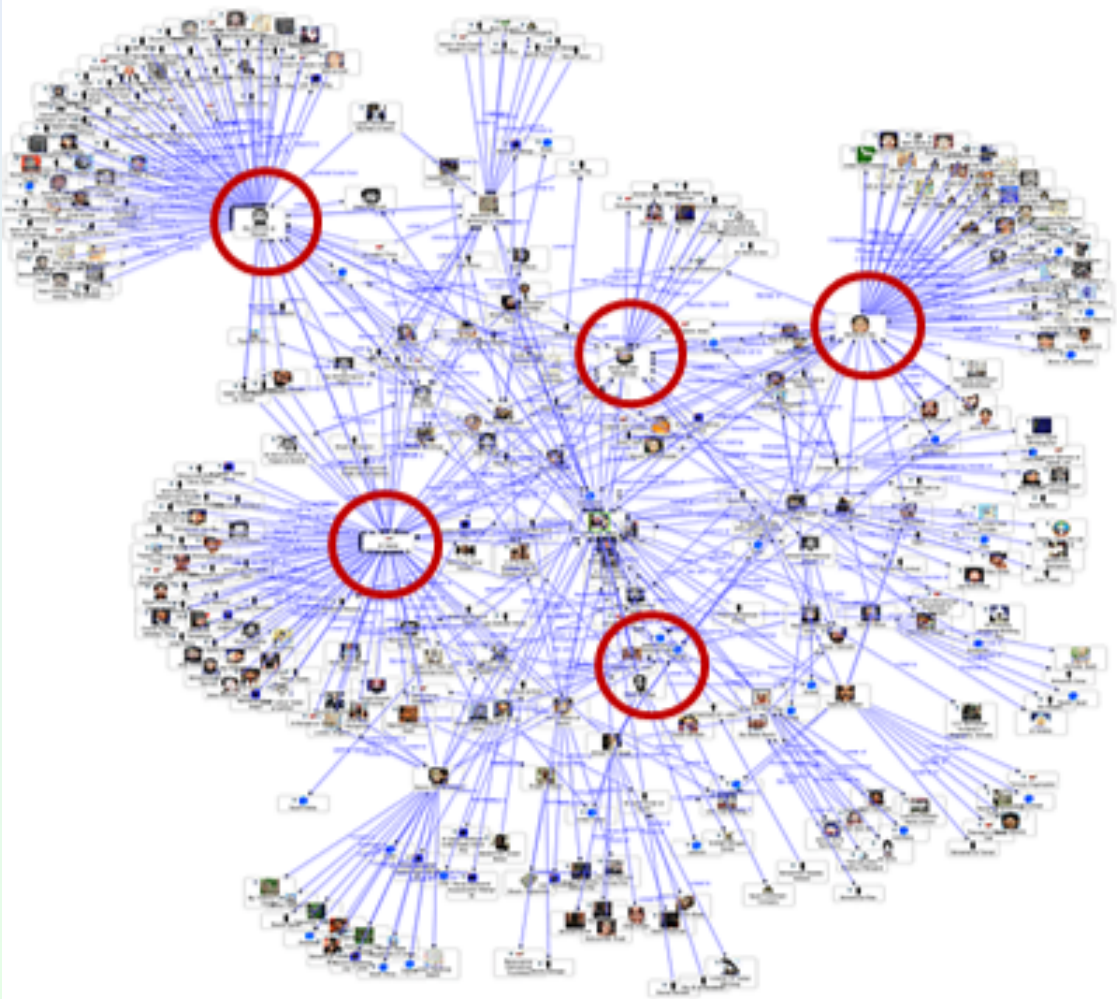
08/20/19

Outline

Outline

Outline

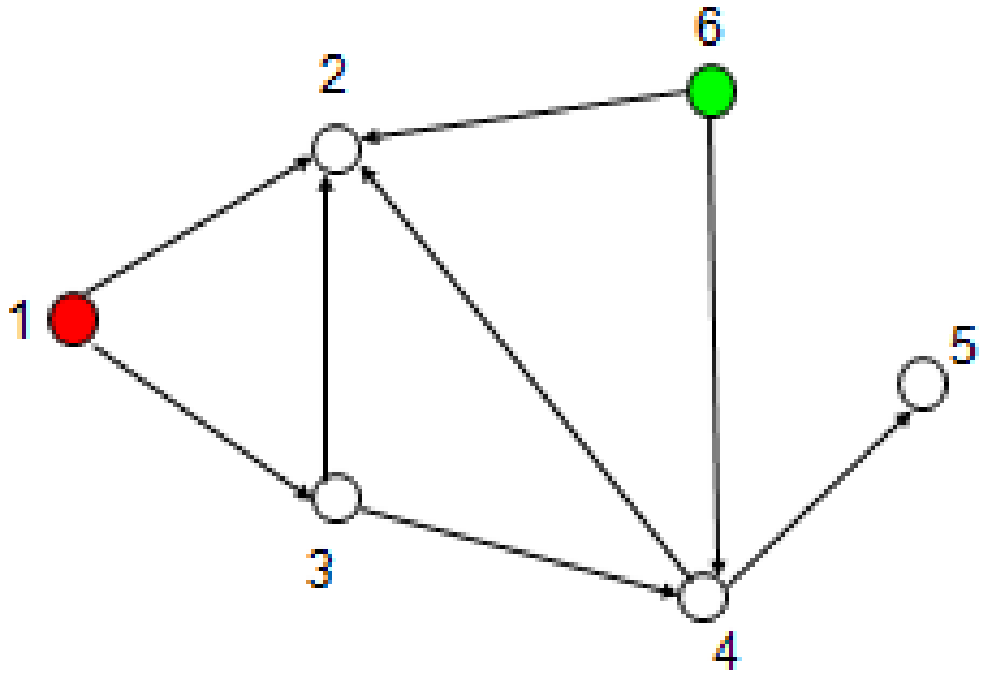
Influence Maximization



- Given a digraph and $k>0$,
- Find k seeds (Kates) to maximize the number of influenced persons (possibly in many steps).

Propagation Model

* IC
* LT



both 1 and 6 are source nodes.

Step 1: 1--2,3; 6--2,4. .

IC Model

Use the directed network to represent a social network $G = (V, E)$.

Each node has two states, active and inactive.

Each arc (u, v) is assigned with a probability p_{uv}, when u is active, v is also activated with probability p_{uv}.

Initially, every node is inactive. To start an information diffusion process, a set of nodes, called *seeds*, are activated.

The process consists of discrete steps.

In each step, each newly activated node would try to influence its out-neighbors.

An active node has only one chance to influence its out-neighbors.

The process ends when no node is activated in current step.

Influence function maximization problem

Influence function maximization problem is proposed as follow:

$$\max_{|S|\leq k} f(S)$$

where $f(S)$ is nonnegative increase set function.

Theorem

Sub-modular influence function maximization problem $\max_{S\subseteq V,|S|\leq k} \sigma(S)$ *is NP hard.*

Theorem

Greedy Algorithm for sub-modular function maximization problem $\max_{S\subseteq V,|S|\leq k} \sigma(S)$ *returns* $1-1/e$ *-approximate solution.*

Marginal Gain

Computation the spread value of a given S is #p hard.

Traditionally, $f(S)$ is evaluated through random sampling method, e.g. Monte Carlo simulation, simulation number?

Is there effective way to evaluate $f(S)$ directly for any given S?

YES! A successive iteration update method (SIUM) to compute $f(S)$ from the marginal gain perspective is given.

Definition

(marginal gain). Suppose that $f : 2^V \to R^+$ is non-negative set value function, where V is ground set. For any subset S of V , $\Delta_v f(S) = f(S \cup \{v\}) - f(S)$ is called marginal gain of $v \in V \backslash S$ at S. Additionally, we can define $\Delta_T f(S) = f(S \cup T) - f(S)$ as marginal gain of $T \subseteq V \backslash S$ at S in the similar way.

Marginal Gain

Based on the definition of marginal gain, we have the following property:

For a given set $S \subseteq V$ and any subset $T \subseteq S$ then we have

$$f(S) = f(S \backslash T) + \Delta_T f(S \backslash T) = f(T) + \Delta_{S \backslash T} f(T)$$

for any given set value function f.

Definition (submodular function)

Suppose that $f : 2^V \rightarrow R^+$ is non-negative set value function, where V is ground set. f is called sub-modular if for any subset S_1 , S_2 of V with $S_1 \subseteq S_2 \subseteq V$, and any $v \in V \backslash S_2$, we have $\Delta_v f(S_1) \geq \Delta_v f(S_2)$, where $\Delta_v f(S) = f(S \cup \{v\}) - f(S)$.

Marginal Gain Formulation

Theorem

(Marginal Gain Formulation) Let $\sigma(S) = \sum_{u \in V} q_u^S$ *is the influence function value (or the spread value) of a given* S*, where* q_u^S *is the active probability (or expectation equivalently) of the node* u *when* S *is selected as seed set under the IC propagation model. Denote* p_{vu} *the probability of node* v *can activate node* u *along a given path. Using these notations, we have*
$\sigma(S) = \sum_{u \in V}[p_{vu} + (1 - p_{vu})q_u^{S \setminus v}]$ *for any* S *and* $v \in S$*.*

Proof.

according to property above we have

$$\sigma(S) = \sigma(S\backslash v) + \Delta_v \sigma(S\backslash v)$$

Where $\sigma(S\backslash v) = \sum_{u \in V} q_u^{S\backslash v}$ and

$$\begin{aligned}\Delta_v \sigma(S\backslash v) &= \sum_{u \in V} [1 - (1 - q_u^{S\backslash v})(1 - p_{vu}) - q_u^{S\backslash v}] \\ &= \sum_{u \in V} (1 - q_u^{S\backslash v}) p_{vu}).\end{aligned}$$

Thus we have

$$\begin{aligned}\sigma(S) &= \sum_{u \in V} [q_u^{S\backslash v} + (1 - q_u^{S\backslash v}) p_{vu}] \\ &= \sum_{u \in V} [p_{vu} + (1 - p_{vu}) q_u^{S\backslash v}].\end{aligned}$$

GCMP Formulation

One way to boost content spread problem is to increase the number of connected edges between users. Some social network sites such as Facebook and Twitter provide the function for users that could recommend friends of your friends to you to make possible connections.

Vineet Chaoji et al.（2012）first 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.

Note: 1. This problem is NP-hard and the spread function is not submodular. Vineet Chaoji et al. construct a more restricted variant that is submodular and devised an approximation algorithm.

2. But their content spread function under IC and RMPP model has some limitations.

Limitations of Vineet Chaoji et al.'s work

1. Computing the expected number of nodes with specific content C is #P-hard. They derived a RMPP (Restricted Maximum Probability Path) model from a heuristic first proposed by Chen et al. which restricts influence propagation between a pair of nodes to be only along the MPP.

RMPP= MPP + at most one edge from X.

The spread function under RMPP model to be submodular but the content spread problem is still NP-hard.

2.The information propagation may not reflect the real flow on the network and the content spread node under RMPP model will have a large deviation from the actual value.

3. A predefined number of new links is added for each user, a case which not necessarily reflects to real-world application.

4. Higher computational cost, less scalable.

GCMP Formulation

From a marginal increment perspective to describe the content spread function value as accurate as possible. Our formulation based on (IC) model.

For the given acyclic directed social network $G(V,E,P)$, denote p_i is the probability with which node i shares content independently with each of its neighbors and $q_{v,S}^{cE}$ is the content spread of a content $c \in C$ contained at $v \in V$ under the topology of E with seed set S(strictly speaking, here S is S_c) and $q_S^{cE} = (\cdots, q_{v,S}^{cE}, \cdots)^T$ is the content spread vector under the topology of E with seed set S.

Then we have the following formula to calculate the marginal gain $\Delta_{e_{st}} q_{t,S}^{cE}$ of content spread of c at node t when an edge $e_{st} \in X$ is added to current topology of E.

GCMP Formulation

Theorem

The marginal gain $\Delta_{e_{st}} q_{v,S}^{cE}$ of content spread of c at node v when an edge $e_{st} \in X$ from a candidate set is added to current topology of E is calculated recursively as follows:

$$\Delta_{e_{st}} q_{t,S}^{cE} = (1 - q_{t,S}^{cE}) p_s q_{s,S}^{cE}$$

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

$$\Delta_{e_{st}} q_{v,S}^{cE} = \frac{1 - q_{v,S}^{cE}}{1 - p_t q_{t,S}^{cE}} p_t \Delta_{e_{st}} q_{t,S}^{cE} \tag{1}$$

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 $\Delta_{e_{st}} q_{v,S}^{cE} = 0$, for the vertex which is unreachable from vertex t during this process.

GCMP Formulation

Note to Th. During the process of updating marginal spread, if there are paths from vertex t reach to different in-neighbor nodes of node w, the marginal gain of spread of w should be updated according to equation 1 multi-times. But the overall marginal gain of content spread for w is independent of the update orders.

Definition

(Generalized Content Maximization Problem (GCMP):) Given a directed acyclic graph $G = (V, E, P)$, a constant K and content set C with given initial seed setsS_c for each $c \in C$, find an edge set $X \subseteq \bar{X} = \{e_{ij} : i, j \in V, i \in N_j, j \in N_i\}$ where N_i is the candidate node set of i such that: (1) At most K edges from X, (2)$f(X)$ is maximum.

Note: In this definition, total cardinality of the edge set is constraint.

Main Contributions

1. Content spread maximization problem is formulated in a marginal gain incremental way with almost no loss of the content spread.

2. The non-submodularity of the content spread function is given with analysis. Both submodular lower-bound and upper-bound of the original content spread function is presented and a Marginal Increment based Sandwich algorithm (MIS) that guarantees a data-dependent approximation factor is devised in the sandwich framework.

3. A novel heuristic scalable algorithm of boosting content spread in social networks IRFA is proposed.

Submodular bounds

The objective function of GCMP $f(X) = \sum_{c\in C}\sum_{v\in V}(q_{v,S}^{cE} + \sum_{e_{st}\in X}\Delta_{e_{st}} q_{v,S}^{c(E\cup X^{st})})$ is still non-submodular. In order to elaborate on the structure of $f(X)$, we can re-write it in marginal increment form. Denote $X = \{e_{s_1t_1}, e_{s_2t_2}, \cdots, e_{s_Kt_K}\}$, $X^k = \{e_{s_1t_1}, \cdots, e_{s_kt_k}\}$, $k = 0, 1, \cdots, K$ and $X^0 = \emptyset$ for convenience. Then we have $f(X) = f(X^0) + \sum_{k=1}^{K}\Delta_{e_{s_kt_k}} f(X^{k-1})$,here $f(X^0) = \sum_{c\in C}\sum_{v\in V} q_{v,S}^{cE}$ and $\Delta_{e_{s_kt_k}} f(X^{k-1}) = \sum_{c\in C}\sum_{v\in V}\Delta_{e_{s_kt_k}} q_{v,S}^{c(E\cup X^{k-1})}$,$k = 1, \cdots, K$. Fortunately, each term $\Delta_{e_{s_kt_k}} q_{v,S}^{c(E\cup X^{k-1})}$ in $\Delta_{e_{s_kt_k}} f(X^{k-1})$ is monotone decrease with $q_{v,S}^{c(E\cup X^{k-1})}$. Thus we have the following monotone decrease property of $f(X)$.

Submodular bounds

Property: Content spread function $f(X) = f(X^0) + \sum_{k=1}^{K} \Delta_{e_{s_k t_k}} f(X^{k-1})$ is monotone decrease with $q_{v,S}^{c(E \bigcup X^{k-1})}$, for $v \in V$ and $k = 1, \cdots, K$.

However, the monotone decrease property does not guarantee the sub-modularity of the objective function $f(X)$, this is just because the neighbor relationship will change during the new edges added.

To see this structure change clearly, we denote $N_E^+(v) = \{u \in V | (v,u) \in E\}$ the out-neighbor of vertex v for each $v \in V$. Obviously we have the inclusion relationship $N_E^+(v) \subseteq N_{E \bigcup X^1}^+(v) \subseteq N_{E \bigcup X^2}^+(v) \subseteq \cdots \subseteq N_{E \bigcup X^K}^+(v) \subseteq N_{E \bigcup \bar{X}}^+(v)$. With this notation,

$\Delta_{e_{s_k t_k}} f(X^{k-1})$ can be rewritten in a more detailed expression:

$$\begin{aligned} & \Delta_{e_{s_k t_k}} f(X^{k-1}) \\ = & \sum_{c \in C} \sum_{v \in V} \Delta_{e_{s_k t_k}} q_{v,S}^{c(E \bigcup X^{k-1})} \\ = & \sum_{c \in C} (\Delta_{e_{s_k t_k}} q_{t_k,S}^{c(E \bigcup X^{k-1})} \end{aligned}$$

Submodular bounds

The reason for $f(X)$ is not sub-modular lies in that the out-neighbor set of a vertex $v \in V$ may become larger and larger with new edges added in the boost content spread update process. However, for a given GCMP, E, P and S that contains content c are all fixed. When we further fix the in-neighbor relationship of all vertexes $v \in V$ during the whole recommendations boost update process, the number of marginal gain terms will remain unchanged during the whole procedure. Due to the monotone increase property of $q_{v,S}^{c(E \bigcup X^{k-1})}$ with respect to edge set X^{k-1} , for each newly added edge $e_{s_k t_k}$, the resultant marginal gain term of content spread $\Delta_{e_{s_k t_k}} q_{v,S}^{c(E \bigcup X^{k-1})}$ becomes monotone decrease with $q_{v,S}^{c(E \bigcup X^{k-1})}$, for $v \in V$ and thus further make it possible to guarantee that the associate content spread function is submodular. Next we construct submodular bounds objective functions by reasonably imposing restriction on the neighborhood structure of each vertex $v \in V$.

Submodular bounds

The lower bound of objective function is constructed as follow: $\underline{f}(X) = f(X^0) + \sum_{k=1}^{K} \Delta_{e_{s_k t_k}} \underline{f}(X^{k-1})$, where

$$
\begin{aligned}
& \Delta_{e_{s_k t_k}} \underline{f}(X^{k-1}) \\
= & \sum_{c \in C} \sum_{v \in V} \Delta_{e_{s_k t_k}} q_{v,S}^{c(E \cup X^{k-1})} \\
= & \sum_{c \in C} (\Delta_{e_{s_k t_k}} q_{t_k,S}^{c(E \cup X^{k-1})} \\
& + \sum_{v_1 \in N_E^+(t_k)} \Delta_{e_{s_k t_k}} q_{v_1,S}^{c(E \cup X^{k-1})} \\
& + \sum_{v_2 \in N_E^+(v_1), v_1 \in N_E^+(t_k)} \Delta_{e_{s_k t_k}} q_{v_2,S}^{c(E \cup X^{k-1})} \\
& + \cdots + \sum_{v_D \in N_E^+(v_{D-1}), v_{D-1} \in N_E^+(v_{D-2})} \Delta_{e_{s_k t_k}} q_{v_D,S}^{c(E \cup X^{k-1})})
\end{aligned}
$$

$k = 1, 2, \cdots, K.$

Submodular bounds

$\underline{f}(X)$ defined above is lower bound of $f(X)$ because that all the term $\Delta_{e_{s_k t_k}} q_{v,S}^{c(E \cup X^{k-1})}$ in $\underline{f}(X)$ is nonnegative and must be included in $f(X)$ due to the inclusion relationship
$N_E^+(v) \subseteq N_{E \cup X^1}^+(v) \subseteq N_{E \cup X^2}^+(v) \subseteq \cdots \subseteq N_{E \cup X^K}^+(v) \subseteq N_{E \cup \bar{X}}^+(v)$.
Furthermore, $\underline{f}(X)$ have the following nice submodular property.

Theorem

The lower bound of objective function $\underline{f}(X) = f(X^0) + \sum_{k=1}^{K} \Delta_{e_{s_k t_k}} \underline{f}(X^{k-1})$ *defined above is submodular with respect to* X.

Submodular bounds

The upper bound is $\overline{f}(X) = f(\overline{X}) - \sum_{e_{st} \in \overline{X} \backslash X} \Delta_{e_{st}} f(\overline{X})$, where

$$f(\overline{X}) = \sum_{c \in C} \sum_{v \in V} q_{v,S}^{c(E \cup \overline{X})}$$

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

$$\begin{aligned}
& \Delta_{e_{st}} f(\overline{X}) \\
= & \sum_{c \in C} \sum_{v \in V} \Delta_{e_{st}} q_{v,S}^{c(E \cup \overline{X})} \\
= & \sum_{c \in C} (\Delta_{e_{st}} q_{t_k,S}^{c(E \cup \overline{X})} \\
& + \sum_{v_1 \in N_E^+(t_k)} \Delta_{e_{st}} q_{v_1,S}^{c(E \cup \overline{X})} \\
& + \sum_{v_2 \in N_E^+(v_1), v_1 \in N_E^+(t_k)} \Delta_{e_{st}} q_{v_2,S}^{c(E \cup \overline{X})} \\
& + \cdots + \sum_{v_D \in N_E^+(v_{D-1}), v_{D-1} \in N_E^+(v_{D-2})} \Delta_{e_{st}} q_{v_D,S}^{c(E \cup \overline{X})})
\end{aligned}$$

Submodular bounds

$\overline{f}(X)$ defined above is upper bound of $f(X)$ because that $\overline{f}(X) - f(X) = (f(\overline{X}) - \sum_{e_{st} \in \overline{X} \setminus X} \Delta_{e_{st}} f(\overline{X})) - (f(X^0) + \sum_{k=1}^{K} \Delta_{e_{s_k t_k}} f(X^{k-1})) \geq \sum_{e_{st} \in \overline{X} \setminus X} (\Delta_{e_{st}} f(X^{l_{st}-1}) - \Delta_{e_{st}} f(\overline{X})) \geq 0$, here l_{st} denote the edge e_{st} is the l_{st}-th edge added into the recommendations boost network among all edges in $\overline{X}$. The last inequality holds because $\Delta_{e_{st}} f(X^{l_{st}-1})$ has at least the same number of items as $\Delta_{e_{st}} f(\overline{X})$ has and $\Delta_{e_{st}} q_{v,S}^{c(E \cup X^{l_{st}-1})} \geq \Delta_{e_{st}} q_{v,S}^{c(E \cup \overline{X})}$ due to the monotone decrease property. Similarly, $\overline{f}(X)$ have the following nice sub-modular property.

Theorem

The upper bound of objective function $\overline{f}(X) = f(\overline{X}) - \sum_{e_{st} \in \overline{X} \setminus X} \Delta_{e_{st}} f(\overline{X})$ defined above is submodular with respect to X.

MIS algorithm

Generally speaking, there is no effective way to optimize or approximate a non-submodular function. Lu et al.(2015) proposed a sandwich approximation strategy, which approximates the non-submodular objective function by approximating its submodular lower-bound and upper-bound.

For GCMP, although the original contend spread function $f(X)$ is non-submodular, we have obtained the submodular lower $\underline{f}(X)$ and upper bound $\overline{f}(X)$.

Therefore the Sandwich framework can be applied and we devise MIS that guarantees a data dependent approximation factor.

MIS Algorithm

The sandwich approximation strategy works as follows:

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 $f(X)$.
4: $X = \arg\max_{X_0 \in \{X_U, X_L, X_A\}} f(X_0)$.
5: **return** X.

Approximation Ratio

A data dependent approximation factor is

Theorem

Let X^* be the seed set returned by MIS and X_A^* is the optimal solutions to maximizing the original spread, then we have

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

$\alpha = \beta = 1 - \frac{1}{e}$ for both submodular bounds, thus we have

Corollary

Let X^* be the seed set returned by MIS and X_A^* is the optimal solutions to maximizing the original spread, then we have

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

IRFA algorithm

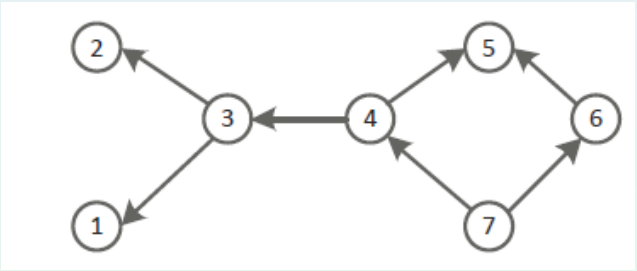
How to obtain high quality solution to the original problem $f(X)$?

We present IRFA: IRFA=influence ranking of single node + fast adjustment according to the recommendations selected.

Intuitively it is beneficial to set up the connection between the seed node and the strong influence node to boost content spread. Then how to rank the influence of single node? In order to obtain influence information, denote σ_v^E 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 $\sigma^E = (\cdots, \sigma_v^E, \cdots)^T$ 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 $e = (1, 1, \cdots, 1)_n$ is n-dimensional vectors with all components of 1 and $A = (a_{ij})_{n \times n}$ is the adjacency propagation matrix of G with $a_{ij} = p_{ij} = p_i$, if $e_{ij} \in E$ and 0 otherwise. D is the diameter of the network G. Usually $\sigma^l, l = 0, 1, \cdots, D$ reflects the l-th hop propagation spread of all vertex in the network.

An Example

Figure 1 gives an example to demonstrate how we calculate σ^E.



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

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

Marginal Gain

Using this spread influence information we know how to establish connections to maximize the spread.

Suppose that node 7 is the only node that have content c, if we add another connection among 7's two-hop neighbors, node 3 ($\sigma^{E_3} = 2$) is better than node 5 ($\sigma^{E_5} = 1$) because 3 has higher spread capability or we can say 3 is more influential than 5.

As for the spread vector, we have the following properties.

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

Marginal Gain

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: 1)$\|\sigma^l\|_1 = n(\overline{d}\overline{p})^l, l = 0, 1, \cdots, D$, where $\|\sigma^l\|_1 = \sum_{u=1}^{n} \sigma^l(u)$ denotes the 1-norm of n-dimension vector σ^l.
Especially,$\|\sigma^0\|_1 = n, \|\sigma^1\|_1 = n\overline{d}\overline{p}$.

2)$\|\sigma\|_1 = \sum_{l=0}^{D} \|\sigma^l\|_1 \leq \frac{n}{1-\overline{d}\overline{p}}$, if $\overline{d}\overline{p} < 1$; $\|\sigma\|_1 = \sum_{l=0}^{D} \|\sigma^l\|_1 = (D+1)n$, if $\overline{d}\overline{p} = 1$; $\|\sigma\|_1 = \sum_{l=0}^{D} \|\sigma^l\|_1 \leq \frac{n(\overline{d}\overline{p})^{D+1}}{\overline{d}\overline{p}-1}$, if$\overline{d}\overline{p} > 1$.

3) If $\overline{d}\overline{p} << 1$, every node at most have $\frac{1}{1-\overline{d}\overline{p}}$ influence in average sense, which can be reach a good approximation by only using 0-step and 1-step influence vector.

IR algorithm of single node

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

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.

1: initialize $\sigma^E = e$
2: **for** $d = 0$ to D **do**
3: Compute $\sigma^E = +A\sigma^E$
4: **end for**
5: **return** σ^E.

IRFA

For the GCMP, what we really care about is how much the spread increment caused by new edges added, not content spread itself.

So another important factor that influences the spread increment should be considered.

$q_{v,S}^{cE}$: the content spread of $c \in C$ 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.

IRFA

The main idea of IRFA is to select the node with maximum of weighted influence in the sense of $(1-q_{t,S}^{cE})\sigma^{E}(t)$ from the candidate set of vertex s, and add the connection (s,t) into edge set.

That is, add edge (s,t) in X such that $t=\arg\max_{v:(s,v)\in\overline{X}_s}(1-q_{v,S}^{c(E\bigcup X)})\sigma^{E}(v)$, where $q_{v,S}^{c(E\bigcup X)}$ is the probability that node v becomes activated after the diffusion process when the edge set is $E\bigcup X$ and the seed set that contains c is S.

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

IRFA

The influence fast update adjustment procedure (FA(s,t)) is as follow.

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

IRFA

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

Algorithm 3 Algorithm IRFA $(G, K, \overline{X})$.

Input: the social network $G = (V, E, P)$, candidate edge set $\overline{X}$, positive number K and content set C.

Output: X satisfying constraint of no larger than 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 $\sigma^{E \cup X}(v) = \sigma^E(v)$
3: **for** $c \in C$ **do**
4: Compute $q_{v,S}^{cE}$ and let $q_{v,S}^{c(E \cup X)} = q_{v,S}^{cE}$ for each $v \in V$
5: **end for**
6: **for** $k = 1$ to K **do**
7: Select Edge $(s,t) = \arg\max_{v:(u,v) \in \overline{X}_u, u \in S_c, c \in C} (1 - q_{v,S}^{c(E \cup X)}) \sigma^{E \cup X}(v)$
8: Run influence update procedure FA(s,t) to obtain $\sigma^{E \cup (X \cup \{e_{st}\})}(v)$
9: Compute $\Delta_{e_{st}} q_{v,S}^{c(E \cup X)}$ and Update $q_{v,S}^{c(E \cup X \cup \{e_{st}\})}$
10: Update $X = X \bigcup \{e_{st}\}$
11: **end for**
12: **return** X as the solution to the GCMP.

Complexity analysis

Initialization: $O(|D|m + \sum_{c \in C} |S_c|m)$.
Total K iterations: $O(k(|\overline{X}| + 2m))$.
Time complexity: $O(|D|m + (\sum_{c \in C} |S_c|m) + k(|\overline{X}| + 2m))$.

PERFORMANCE EVALUATION

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

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

Wikipedia: The Wikipedia dataset 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 103689 edges. Each node in the graph represents a user attend the voting procedure. Each directed edge denotes who vote for whom.

HEP-TH: This is a citation graph which from the e-print arXiv and covers all the citations within a dataset 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]

PERFORMANCE EVALUATION

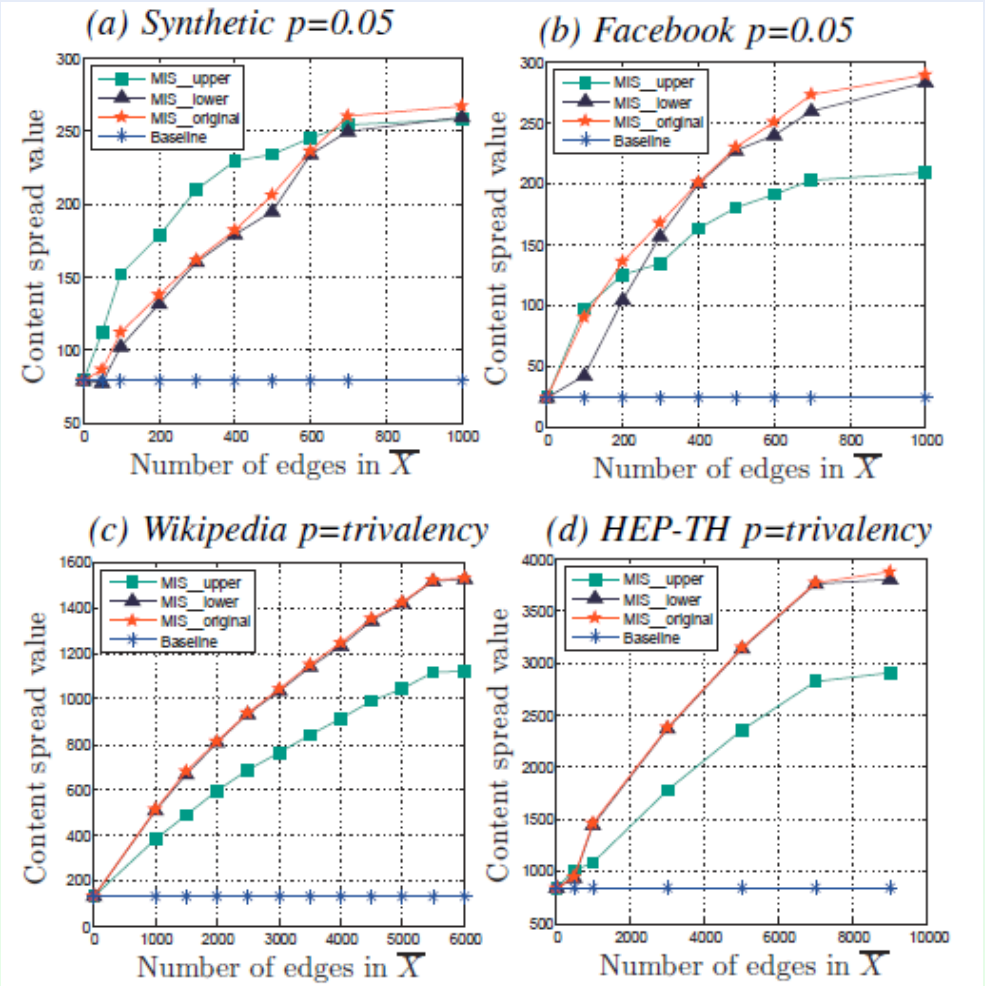
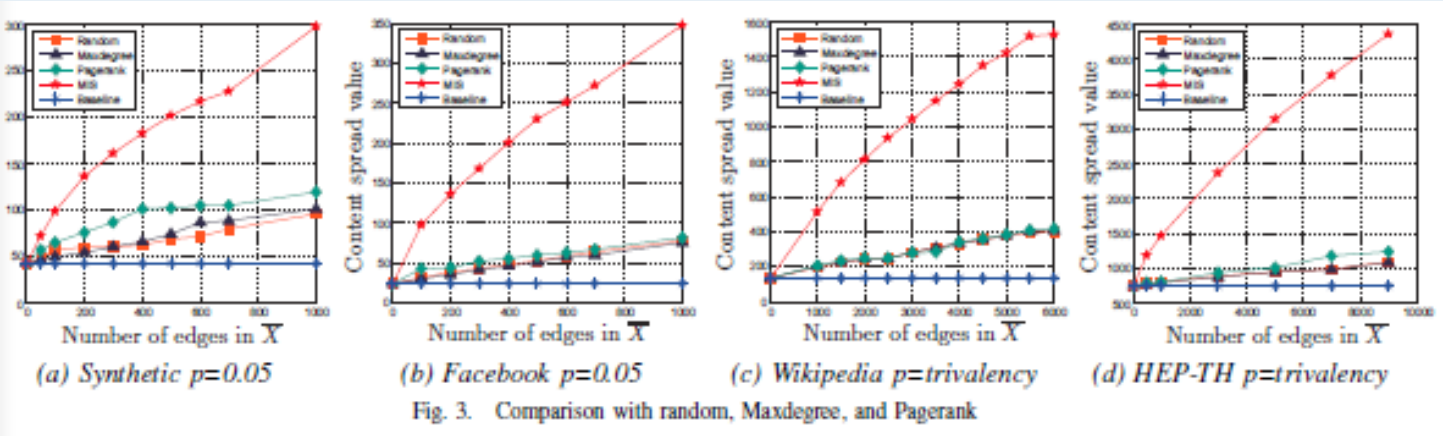


Fig. 2. Content spread value vs. incremental edges under uniform and trivalency propagation

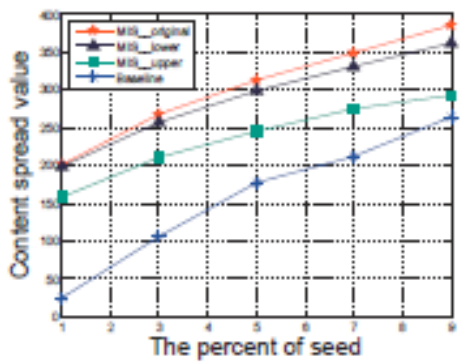
PERFORMANCE EVALUATION

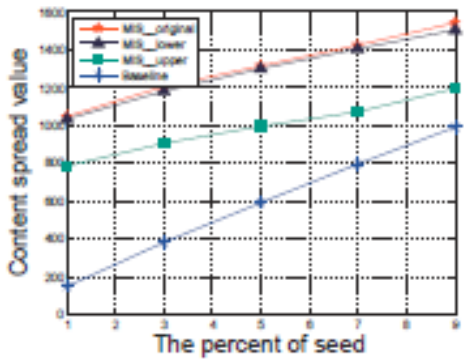


(a) Synthetic p=0.05 *(b) Facebook p=0.05* *(c) Wikipedia p=trivalency* *(d) HEP-TH p=trivalency*

Fig. 3. Comparison with random, Maxdegree, and Pagerank

PERFORMANCE EVALUATION

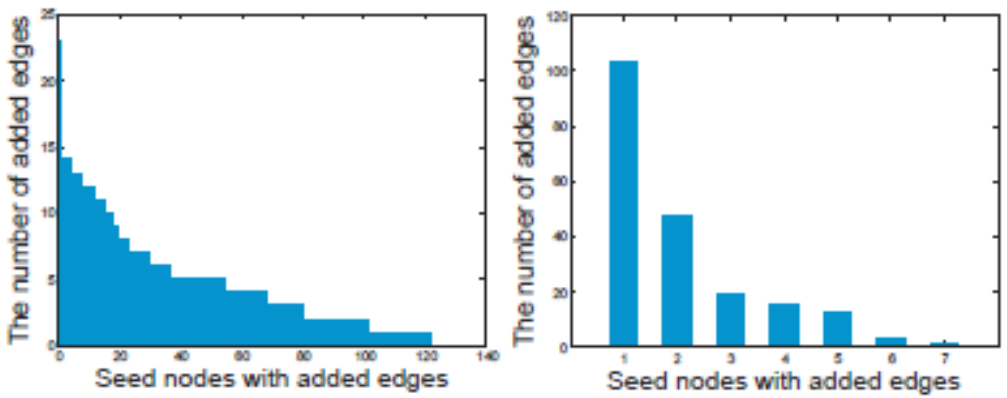




(a) Facebook p=trivalency *(b) Wikipedia p=0.05*

Fig. 4. Content spread value vs. incremental seed set size

PERFORMANCE EVALUATION



(a) HEP-TH p=trivalency *(b) Facebook p=0.05*

Fig. 5. Number of edges added on each seed node

MAP Formulation

In above social network, every node represents a user. If an activity requires a group of users to participate, then this activity can be represented by this group of users. Therefore, all activities form a hyper-edge set of a hypergraph with the node set V, denoted by $\mathscr{A}$. The profit $c : \mathscr{A} \to R^+$ is a non-negative function. For any seed set S, denote by $I(S)$ the set of active nodes at end of the diffusion process with seed set S initially. The expected profit of activities generated by active nodes would be defined as

$$f(S) = E\left[\sum_{A \subseteq I(S), A \in \mathscr{A}} c(A)\right].$$

Definition (Maximizing Activity Profit)

Given a social network $G = (V, E)$ with the IC model, a collection of activities, $\mathscr{A}$, a profit function $c : \mathscr{A} \to R^+$, and a positive integer k, find a set S of k seeds to maximize the expected total profit of activities consisting of users activated by S through influence.

MAP Formulation

From the definition of MAP, we can see that it is a generation of many classical IM problems:

1. Influence maximization problem. When $\mathscr{A}$ is defined as $\mathscr{A} = V$ only including all the single vertex activity and $c : \mathscr{A} \to R^+$ is defined as $c(v) = |v| = 1$, Influence maximization problem is a special case of Maximizing Activity Profit.

2. Activity maximization problem. When $\mathscr{A}$ is defined as $\mathscr{A} = E$ only including all the edge set activity and $c : \mathscr{A} \to R^+$ is defined as $c(e) = w_e$, Activity maximization problem is a special case of Maximizing Activity Profit.

Main Contributions

1. We propose a novel MAP problem to maximize the expected total profit in a social network.

2. The objective function for the MAP is neither submodular nor supermodular. We present an approximate algorithm that yields an approximation ratio of $\frac{1}{\Delta+2}$ provided that the supermodular degree is bounded with Δ.

3. We also develop an exchange-based algorithm to further improve the quality of the solution.

Complexity

The complexity results of Maximizing Activity Profit problem.

Theorem

(NP-hardness). The Maximizing Activity Profit problem is NP-hard.

Theorem

(#P-hardness). For a given set S, the computation of

$$f(S) = E\left[\sum_{A \subseteq I(S), A \in \mathscr{A}} c(A)\right]$$

is #P-hard.

Modularity

The properties of the objective function of MAP is neither submodular nor supermodular.

Definition (supermodular function)

Suppose that $f : 2^V \rightarrow R^+$ is non-negative set value function, where V is ground set. f is called super-modular if for any subset S_1 , S_2 of V with $S_1 \subseteq S_2 \subseteq V$, and any $v \in V \backslash S_2$, we have $\Delta_v f(S_1) \leq \Delta_v f(S_2)$.

Modularity

Theorem

(non-submodularity). The objective function

$$f(S) = E\left[\sum_{A \subseteq I(S), A \in \mathscr{A}} c(A)\right]$$

of Maximizing Activity Profit is non-submodular.

Similarly, we have

Theorem

(non-supermodularity). The objective function

$$f(S) = E\left[\sum_{A \subseteq I(S), A \in \mathscr{A}} c(A)\right]$$

of Maximizing Activity Profit is non-supermodular.

supermodular degree algorithm

We use the supermodular degree to measure to what degree the non-submodular version violates the submodularity inspired by M. Feldman et al.(2014). When the supermodular degree is bounded, denoted by Δ, an algorithm Improved Extendible System Greedy with consistant approximation ratio $\frac{1}{(\Delta+2)}$ is proposed.

Definition

(supermodular set) Given a monotone set value objective function $f(\cdot)$, the supermodular set of a node $v \in V$ is $D_f^+(v) = \{v' \in V | \Delta_v f(S \cup \{v'\}) > \Delta_v f(S), \exists S \subseteq V\}$, which includes all nodes that might increase the marginal gain of $v \in V$.

Definition

(supermodular degree). The supermodular degree, denoted by Δ, is defined as the maximum cardinality among all supermodular sets, i.e., $\Delta = \max_{v \in V} \mid D_f^+(v) \mid$.

It is obvious that when Δ=0 the function $f(\cdot)$ is submodular. As for the nonsubmodular case of MAP with bounded Δ, we design improved greedy with average marginal gain algorithm similarly to the naive Extendible System Greedy as follow.

IMEG Algorithm

Algorithm 1 Improved Extendible System Greedy (IESG).

Input: a hyper-graph,G, and a constant, k.

Output: a set of seed nodes,S.

1: Initialize $i = 0$ and $S_0 = \emptyset$.

2: While $|S_i| < k$ do

3: Find out
$\arg\max_{v \in V, D'_v \subseteq D^+_f(v)} [f(S_i \cup \{v\} \cup D'_v) - f(S_i \cup D'_v)]$,
constrained by $|S_i \cup \{v\} \cup D'_v| < k$.

4: Update $S_{i+1} = S_i \cup \{v\} \cup D'_v$.

5: Update i to be $i + 1$.

6: Return $S = S_i$ as the set of seed nodes.

New Rule

In the above Improved Extendible System Greedy algorithm (IMEG), we use a new rule to select the nodes which maximizes the average marginal gain of nodes selected instead of total incremental. This rule is more beneficial due to the monotone property of the objective function always select the nodes in modular set as large as possible.

Approximation Ratio

The IESG algorithm designed above have $\frac{1}{(\Delta+2)}$ approximation ratio.

Theorem

Algorithm IESG has an approximation ratio of $\frac{1}{(\Delta+2)}$ to the optimal algorithm.

Optimization Condition

Although IESG returns $\frac{1}{(\Delta+2)}$-approximate solution to MAPP, it is still a big gap between the approximate solution obtained and the optimal one in theory. How to determine whether an approximate solution can be improved or not is still a fundamental question.

Theorem

(optimization criterion). Suppose S^* is the optimum solution of the MAP, then

$$\min_{S\subseteq S^*} \Delta_S f(S^*\backslash S) \geq \max_{S\subseteq V\backslash S^*} \Delta_S f(S^*\backslash S_R) \tag{2}$$

where $S_R = \arg\min_{S\subseteq S^*} \Delta_S f(S^*\backslash S)$ and

$$f(S) = E\left[\sum_{A\subseteq I(S), A\in\mathscr{A}} c(A)\right].$$

Optimization Condition

Proof.

Proof: Suppose $\min_{S\subseteq S^*} \Delta_S f(S^*\backslash S) \geq \max_{S\subseteq V\backslash S^*} \Delta_S f(S^*\backslash S_R)$ is not satisfied, then there exists $S_A \subseteq V\backslash S^*$ such that $\Delta_{S_R} f(S^*\backslash S_R) < \Delta_{S_A} f(S^*\backslash S_R)$.
On the other hand

$$f(S^*) = f(S^*\backslash S_R) + \Delta_{S_R} f(S^*\backslash S_R) \tag{3}$$

$$f(S^* - S_R + S_A) = f(S^*\backslash S_R) + \Delta_{S_A} f(S^*\backslash S_R) \tag{4}$$

Therefore $f(S^* - S_R + S_A) > f(S^*)$. This is conflict with the optimality of S^* . □

Optimization Condition

When just the singleton subset of S^* is considered, a corollary of the above optimization criterion is immediately obtained as follow.

(Corollary:necessary condition for optimality) Suppose S^* is the optimum solution of the Maximizing Activity Profit problem, then

$$\min_{1\leq i\leq k} \Delta_{v_i} f(S^*\backslash\{v_i\}) \geq \max_{v\in V\backslash S^*} \Delta_v f(S^*\backslash\{v_R\}) \tag{5}$$

where $v_R = \min_{1\leq i\leq k} \Delta_{v_i} f(S^*\backslash\{v_i\})$.

M-convexity of feasible region

Now we turn to the property of the feasible region of MAP. The feasible region of MAPP is an M-convex set.

Theorem

(M-convexity of feasible region) The feasible region of Maximizing Activity Profit problem $F = \{S | S \in \{0,1\}^V \wedge \| S \|_1 = k\}$ *is M-convex set.*

Proof.

For $S_1, S_2 \in F$ and $u \in supp^+(S_1 - S_2)$, there exists $v \in supp^-(S_1 - S_2)$ such that $S_1 - \chi_u + \chi_v \in F$ and $S_2 + \chi_u - \chi_v \in F$, where χ_u is the indicator vector of singleton set $\{u\}$. In fact, any $v \in S_2 \backslash S_1$ is candidate that meets the requirements. By the definition of M-convex set, it follows the theorem. □

Exchange Improvement Property

By the M-convexity of feasible region of MAP, a feasible solution is reserved unchanged under exchange operations. So we have the following:

Theorem

(exchange improvement property) For any feasible solution S of Maximizing Activity Profit problem, if necessary condition for optimality is not satisfied, then S can be improved through exchange operations.

Proof.

According to the optimization criterion and M-convexity of feasible region, $S-\chi_{v_R}+\chi_{v_A} \in F$ outperforms S, that is, $f(S-\chi_{v_R}+\chi_{v_A}) = f(S)+\Delta_{v_A} f(S\backslash\{v_R\}) - \Delta_{v_R} f(S\backslash\{v_R\}) > f(S)$, if necessary condition for optimality is not satisfied. □

Exchange Improvement Property

Now the definition of non-improvable solution of MAP is presented.

Definition

(non-improvable solution). S is said to be non-improvable solution of MAP, if the necessary condition for optimality (Corollary) is satisfied.

Based on the exchange improvement property mentioned above, we can design exchange improvement algorithm (EIA) as follow.

The basic idea behind the EIA is replace the point with the minimum marginal gain in the current solution with the maximum marginal gain point in $V \backslash S$.

Exchange Improvement Algorithm

Algorithm 2 Exchange Improvement Algorithm for MAPP (EIA):

Input: a solution S_0 of influence function maximization problem with $|S_0| = k$.

Output: a set of non-improvable solution $\widehat{S}$.

1: Initialize $i = 0$ and $S = S_0$

2: Find out $v_R = \arg\min_{v \in S} \Delta_v f(S \backslash \{v\})$, let $\Delta(v_R) = \Delta_{v_R} f(S \backslash \{v_R\})$

3: Find out $v_A = \arg\max_{v \in V \backslash S} \Delta_v f(S \backslash \{v_R\})$ and let $\Delta(v_A) = \Delta_{v_A} f(S \backslash \{v_R\})$

4: If $\Delta(v_R) \geq \Delta(v_A)$, then S is non-improvable, stop; otherwise $S := S - v_R + v_A$ and go to step 2.

NP-harness of Supermodular Set Determine

Insurmountable difficulties to solve the MAP in practice: 1) the NP-haedness of the problem and 2) the $\#P$-hardness of the objective function value computation. Furthermore 3) compute the supermodular set is still NP-hardness.

In improved greedy algorithm (IESG), we should determine supermodular set $D_f^+(v)$ for each node $v \in V$ at the initial phase. Unfortunately, decide whether a node u in supermodular set $D_f^+(v)$ of given v is NP-hard.

Theorem

(NP-harness of supermodular set determine). To determine whether a node u is in supermodular set $D_f^+(v)$ of given v is NP-hard.

Upper Bound

Thus we have an upper bound of

$$f(S) = E\left[\sum_{A\subseteq I(S), A\in\mathscr{A}} c(A)\right]$$

which can be reformulated as $f(S) \leq \sum_{A\subseteq I(S),A\in\mathscr{A}}(c(A)\min_{v\in A}\{q_v^S\})$. The upper bound can be frther reformulated as $\overline{f}(S) = \sum_{A\subseteq I(S),A\in\mathscr{A}}(c(A)\frac{1}{|A|}\sum_{v\in A}\{q_v^S\})$ due to the inequality $\min_{v\in A}\{q_v^S\} \leq \frac{1}{|A|}\sum_{v\in A}\{q_v^S\}$.

Submodularity of Upper Bound

Theorem

(submodularity of upper bound). The upper bound objective function $\overline{f}(S)$ *of Maximizing Activity Profit is submodular.*

Proof.

This is because the upper bound of objective function of MAPP $\overline{f}(S) = \sum_{A \subseteq I(S), A \in \mathscr{A}} (c(A) \frac{1}{|A|} \sum_{v \in A} \{q_v^S\}) = \sum_{v \in V} q_v^S \overline{c}(v)$, where $\overline{c}(v) = \sum_{A \in \mathscr{A}} \frac{1}{|A|} c(A)$. This is a weight version of influence maximization problem which is submodular. □

PERFORMANCE EVALUATION

TABLE I
CHARACTERISTICS OF SOCIAL NETWORK DATASETS.

Name	#Nodes	#Edges	Avg. degree
Facebook	899	72, 821	165
arXiv	16, 726	66, 759	11.9
Epinions	22, 166	353, 546	33.5

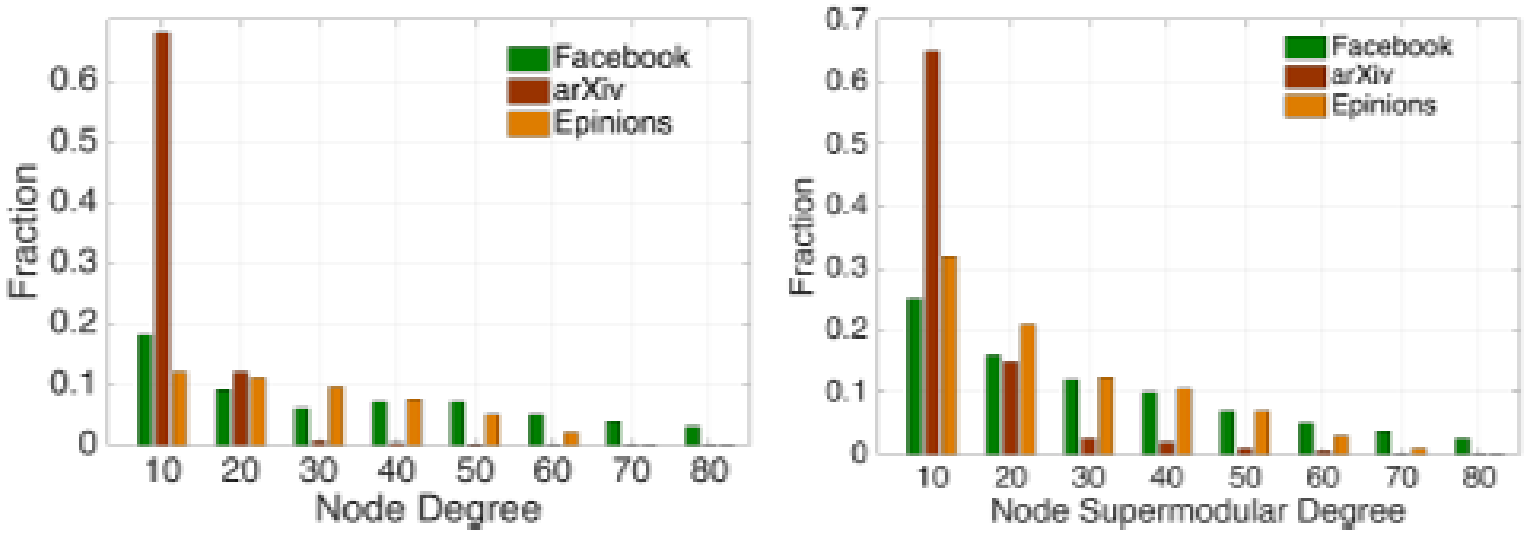


Fig. 1. Node degree and supermodular degree distribution

PERFORMANCE EVALUATION

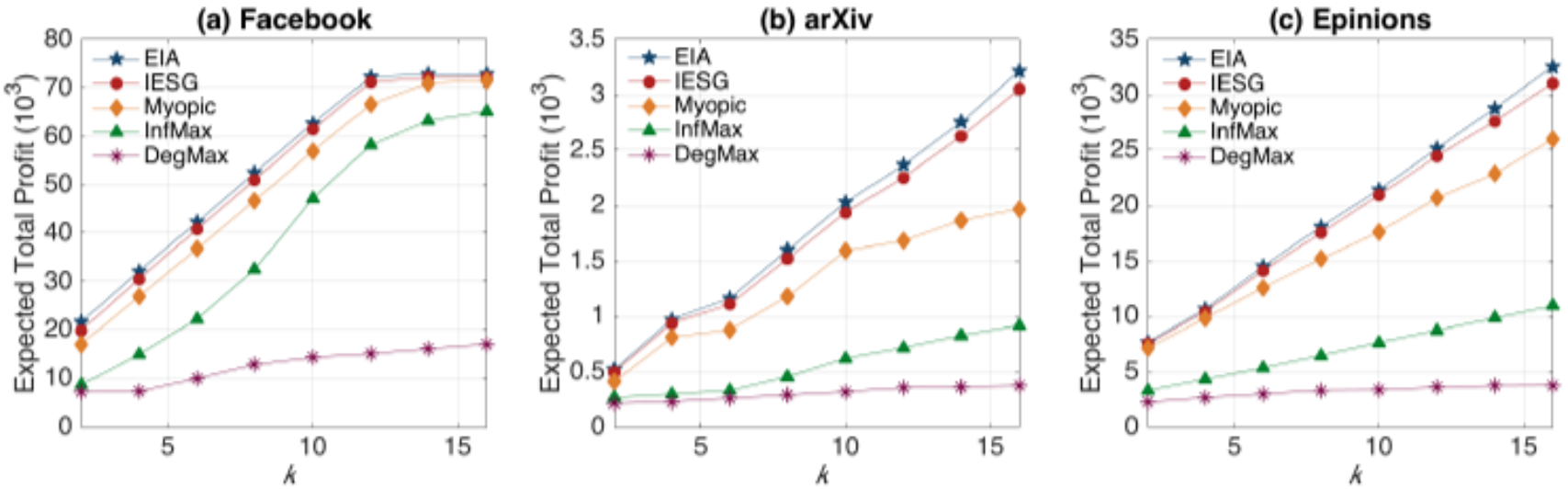


Fig. 2. Expected total profit vs. seed set size produced by various algorithms under uniform profit setting

PERFORMANCE EVALUATION

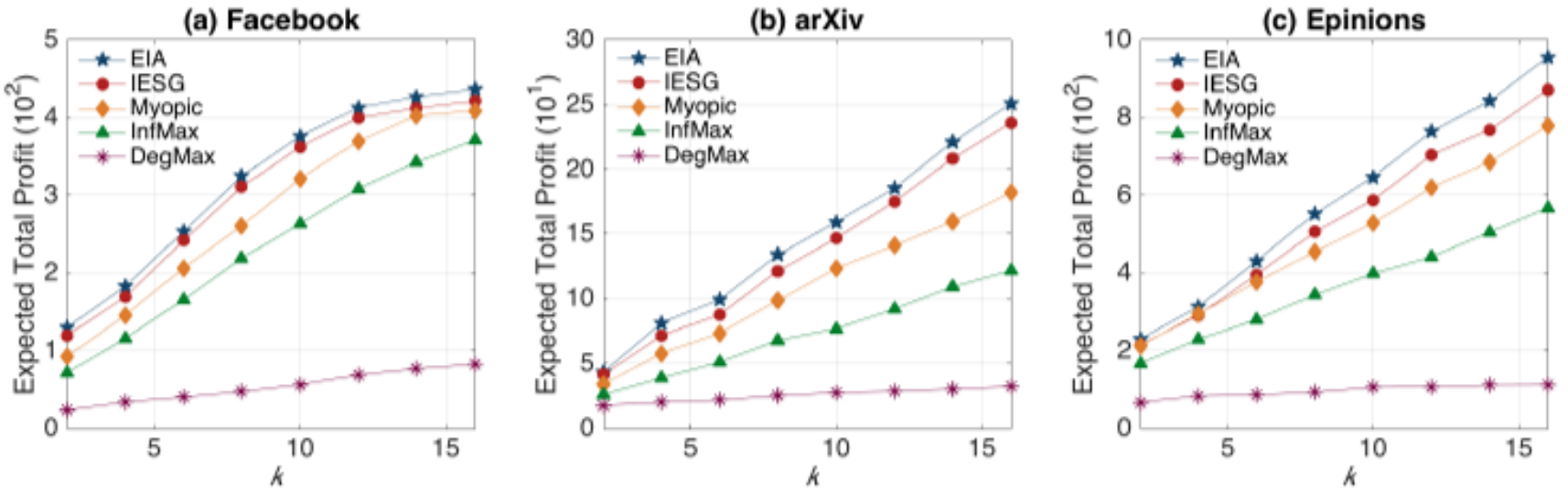


Fig. 3. Expected total profit vs. seed set size produced by various algorithms under influence profit setting

PERFORMANCE EVALUATION

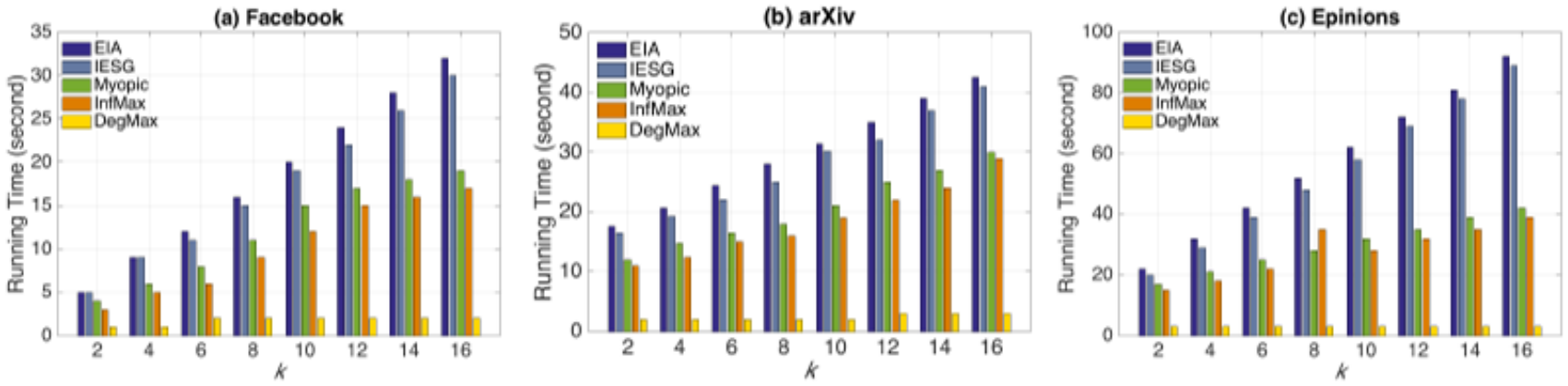


Fig. 4. Running time vs. seed set size produced by various algorithms

PERFORMANCE EVALUATION

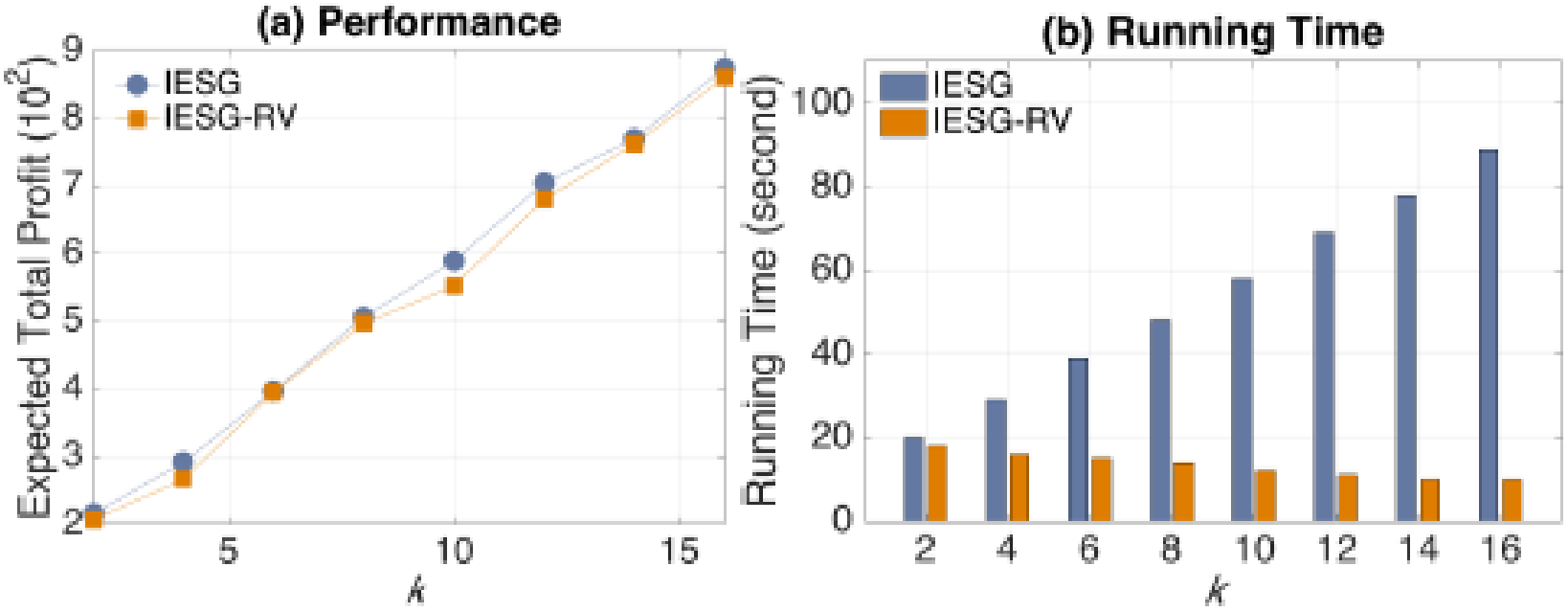


Fig. 5. Performance vs. running time for random variation

PAPER

Yang Wenguo; Yuan Jing; Wu Weili; Ma Jianmin; Du Dingzhu. Maximizing Activity Profit in Social Networks. IEEE Transactions on Computational Social Systems. ISSN:2329-924X Volume:6 Issue:1 Page:117-126.

Yang Wenguo; Ma Jianmin; Li Yi; Yan Ruidong; Yuan Jing; etal. Marginal Gains to Maximize Content Spread in Social Networks. IEEE Transactions on Computational Social Systems. ISSN:2329-924X Volume:6 Issue:3 Page:479-490.

Zhang Yapu; Yang Xianliang; Gao Suixiang; Yang Wenguo. Budgeted Profit Maximization Under the Multiple Products Independent Cascade Model. IEEE Access. ISSN:2169-3536 Volume:7 Page:20040-20049.

REFERENCE

D.Kempel, J.Kleinberg, E.Tardos, Maximizing the spread of influence through a social network, in: Proc. ACM SIGKDD' 03, 2003, pp. 127-136.

W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In KDD, pages 1029 – 1038, 2010.

W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rinc′ on, X. Sun, Y. Wang, W. Wei, and Y. Yuan, Influence maximization in social networks when negative opinions may emerge and propagate. In SDM, pages 379 – 390, 2011.

W. Chen, W. Lu, and N. Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In AAAI, 2012.

C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier.Maximizing social influence in nearly optimal time. In SODA, pages 946 – 957, 2014.

REFERENCE

Y. Tang, X. Xiao, and Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. In SIGMOD, pages 75 – 86, 2014.

Y. Tang, Y. Shi, and X. Xiao. Influence Maximization in Near-Linear Time: A Martingale Approach. In SIGMOD, pages 1539 – 1554, 2015.

Zhefeng Wang, Yu Yang, Jian Pei, and Enhong Chen, Activity maximization by effective information diffusion in social networks.IEEE Transactions on Knowledge and Data Engineering,2017.

V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt. Recommendations to boost content spread in social networks. In WWW, pages 529 – 538, 2012.

W. Lu, W. Chen, and L. V. Lakshmanan, "From competition to complementarity: comparative influence diffusion and maximization," Proc. of the VLDB Endowment, vol. 9, no. 2, pp. 60 – 71, 2015.

Thank You for your attention!