Two-Step Optimal Advertisement Strategy on Social Network System

Jing Wu jw3233@columbia.edu

1 Abstract

The main purpose of this project is to give an advertisement strategy on social network system. In order to advertise a new product effectively, we want to find a set of customers maximizing the total influence in the entire network. Under the Independent Cascade Model, I give a close form of information maximization problem by a linear approximation. In order to give an efficient and stable algorithm to this optimization problem, I screen down the network to a smaller-scale one by using Greedy Algorithm. Working on the smaller subset, using L_1 relaxation, the problem can finally be solved by linear programming.

2 Introduction

Social networks have been studied extensively by social scientists for decades. Enabled by the Internet and sparked by the recent advent of online social networking sites such as Facebook and LinkedIn, research on social networks is witnessing an unprecedented growth due to the ready availability of large scale social network data.

One important study in this literature is viral marketing. That is, a company may want to promote a new product or innovation through word-of-mouth effects in social networks. Research shows that people trust the information obtained from their close social circle far more than the information obtained from general advertisement channels such as TV, newspaper and online advertisement.[1]

My project is motivated by a variety of existing and future applications in which advertisement strategy is carried out based on network effect. There is one specific research direction regarding this problem, called *Influence Maximization*(IM). Large body of related works discussed:

1. How to model in detail the influence diffusion process in the network;
2. How to maximize the total influence in the network.

Influence maximization as an algorithmic technique for viral marketing was first proposed by Domingos and Richardson [2]. Given a social graph $G = (V, \mathbf{E})$ and a stochastic diffusion model on G, we want to find a subset $S \subset V$, such that the influence spread of S, $\sigma(S)$, under the given diffusion model is maximized. That is, compute $S \subset V$ such that

$$S = \arg\max_{S \subset V} \sigma(S) \tag{1}$$

For modeling stochastic diffusion process, there are two widely used model: the *Independent Cascade Model*(ICM) by Goldenberg [3] and the *Linear Threshold* (LT) model by Granovetter.[4] In this project, I'm aiming at maximization part, so the modeling problem is beyond my discussion. I will just adopt the most widely used model, ICM, in my work, which will be described in the Section 3.

To maximize the diffusion of influence in the network, many methods are proposed based on set cover problem which has been well studied.[5] However, due to the various structure of network and different marketing purposes, different algorithms are needed for different models.

In reality, to advertise a new product, it is reasonable to put a budget constraint and must-cover constraint in this information maximization problem. Hence, in this project, I consider the following optimization problem: first, I will use Independent Cascade Model to compute influence spread of each person. To maximize the total influence faster, then, I select a featured subset of customers by a reasonable criterion, which considered both the relevance and diversity of customers. Under the constraints of *(1)* total number of targeted customers and *(2)* necessity of covering the featured subset, I use two algorithms to address this optimization problem: Greedy algorithm and Linear programming after some relaxation.

The organization of the rest of the project report is as follows:

- Section 3 discusses the information maximization problem in social network in detail. To address the goal, mathematical formulation of the model is constructed.Some tradition methods to slove information maximization problem are studied. Shortages of those methods in regard of previous setting are discussed.
- Section 4 To deal with those shortages, two-step algorithm inspired by screening method in variable selection is proposed. The rationality of using Greedy algorithm and variation of linear programming is presented.
- Section 5 shows the simulation result....
- Section 6 presents a final discussion of my model and method: the contribution and what could be done in the future.

3 Model Formulation

3.1 Viral Marketing through Social Networks

Humans behave in a viral fashion and have a natural inclination to share information so as to gain reputation, trustworthiness, or money. This *word-of-mouth* (WOM) dissemination of information through social networks is of paramount importance in our everyday life. Basically, the flow of information or influence through a large-scale network can be thought of as unfolding with the dynamics of an epidemic. As individuals become aware of new ideas, technologies, fads, rumors, or gossip, they have the potential to pass them on to their friends and colleagues, causing the resulting behavior to cascade through the network.

One strong motivation for studying information and influence diffusion models and mechanisms is viral marketing. That is, a company may want to promote a new product through word-of-mouth effects in social networks. A cost-effective way to do it could be to target influencers in the network, investing resources in getting them to adopt the product, for instance by emailing adertisement to them or by giving them product samples for free. The hope is that these influencers will be able to drive other people in the network to adopt the product, generating a potentially large cascade in the network.

A classical example of the viral marketing is the Hotmail free email service, which grew from zero to 12 million users in 18 months on a minuscule advertising budget, thanks to the inclusion of a promotional message with the services's URL in every email sent using it.[6]

While viral marketing is an important application of influence maximization, many other applications may also benefit from the study of it, such as expert finding, trendy topic monitoring, disease outbreak detection, etc.[7]

3.2 Mathematical Formulation

For better introducing the problem I'm considering and my approch, I'll first show the general notations regarding social networks and issues related to the traditional social influence modeling and maximization problem.

3.2.1 Problem Formulation

Let $G = (V, \mathbf{E}, \mathbf{T})$ be an unweighted directed network (as shown in Figure.1), where $V = \{1, 2, ..., n\}$ is node (costomer) set. $\mathbf{E} = [e_{ij}]_{t \times t}$ represents the connections between nodes. If i costumer can influence j (or you can say, j trusts i with non-zero probability), then $e_{ji} = 1$, otherwise, $e_{ji} = 0$. $\mathbf{T} = [t_{ij}]_{n \times n}$ is a transmission matrix for influence propagation. t_{ij}represents the propagation probability ability from i to j. If $e_{ji} = 1$ (i.e., j trust i), then $t_{ij} > 0$, otherwise $t_{ij} = 0$. In this project, I assume $\mathbf{T}$ is known and usually $\sum_{i=1}^{n} t_{ij} \leq 1$.[8] The reason why G is assumed to be directed is that influence propagation is specific to direction in the most general case.

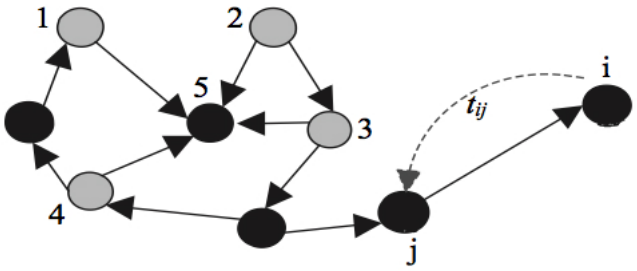

Figure 1: Network $G = (V, \mathbf{E}, \mathbf{T})$

3.2.2 The Independent Cascade Model (ICM)

Modern models of social influence have been augmented with various features allowing for arbitrary network structure, non-uniform interactions, probabilistic events and other aspects. Here, I'm going to use the basic stochastic model of social influence, the *Independent Cascade Model* (ICM). Because the main point of my project is to develop algorithm in order to deal with optimization problem with constraint, so the modeling part is beyond discussion.

The ICM was introduced by Goldenberg et.al in 2001[3] to model the dynamics of viral marketing and is inspired from the field of interacting particle systems. In this model, we start with an initial set of active individuals. Each active individual i has a single chance to activate each non-active neighbour j of his/her. However, the process of activation is deemed stochastic and succeeds with probability t_{ij} independently for each attempt. Therefore, from an initial population of active individuals, the activation process spreads in a cascading manner as newly activated individuals may activate new nodes that either previous attempts failed to activate or were not before accessible. This iterative propagation process will not stop until there is no newly influenced node. (Figure.2)

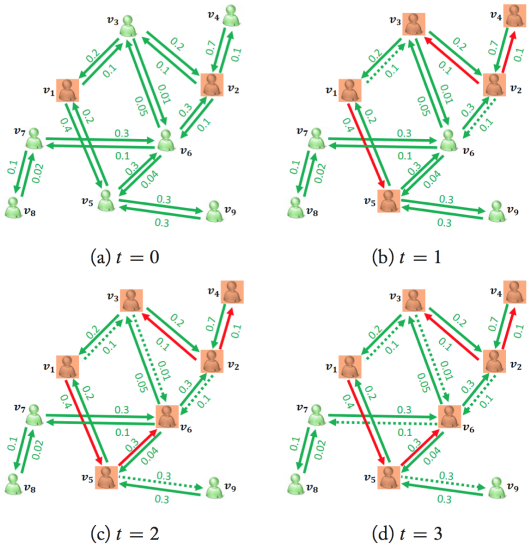

Figure 2: The Independent Cascade Model (ICM)

Denote the influence from i to j by $f_{i\to j}$. From the influence process we described above, the model is proposed as[9]

$$f_{i\to i} = \alpha_i,\ \ \alpha_i > 0 \tag{2}$$

$$f_{i\to j} = \frac{1}{1+\lambda_j}\sum_{k\in N_j} t_{kj} f_{i\to k},\ \ for j \neq i \tag{3}$$

where $N_j = \{j_1, j_2, ..., j_m\}$ is j's trust-friends set, i.e., for $\forall k \in N_j, e_{jk} = 1$. From the definition of $\mathbf{E}$, we know that, if $k \notin N_j, e_{jk} = 0$. Then Equation(3) is equivalent to

$$f_{i\to j} = \frac{1}{1+\lambda_j}\sum_{k\in V} t_{kj} f_{i\to k},\ \ for j \neq i \tag{4}$$

In Equation(2), we assign each node i a prior probability value α_i. If i has a full probability to spread the information, this value should be the maximum (e.g., 1). Conversely, if i has no interest at all, it will be 0. Equation(3) is formulated as we assume the influence from i to j is proportional to the linear combination of the influence from i to js neighbors $k \in N_j$. In the ICM, we know that the influence process succeeds with probability t_{kj}. That's why we combine the total influence from i to j in this way: one part is from i itself, another part is from i's inderect influence, i.e., from j's other neighbors, who are influenced by i and will pass those information to j with non-zero probability. The parameter λ_j is the damping coefficient of j for the influence propagation. It is in the range $(0, +\infty)$, and the smaller λ_j is, the less influence will be blocked by node j. For simplicity, we choose the same λ_0 for each node, and name $\lambda_0\mathbf{I}$ as the damping matrix. We denote total influence spread from i to the entire network as $f_{i\to V} = \sum_{j\in V} f_{i\to j}$.

Under the above model, then, we want to compute the influence spread vector $\vec{f_i} = [f_{i\to 1}, f_{i\to 2}, ..., f_{i\to n}]^T$, which is the influence from i to each node in the network. We can rewrite Equation (2) and

(4) in the vector form,

$$\begin{aligned}\vec{f_i} &= \begin{pmatrix} f_{i\to 1} \\ \vdots \\ f_{i\to i} \\ \vdots \\ f_{i\to n} \end{pmatrix} = \frac{1}{1+\lambda_0} \begin{pmatrix} \sum_{k\in V} t_{k1} f_{i\to k} \\ \vdots \\ \alpha_i \\ \vdots \\ \sum_{k\in V} t_{kn} f_{i\to k} \end{pmatrix} \\ &= \frac{1}{1+\lambda_0} \left[\begin{pmatrix} t_{11} & \cdots & t_{n1} \\ \vdots & & \vdots \\ t_{1i} & \cdots & t_{ni} \\ \vdots & & \vdots \\ t_{1n} & \cdots & t_{nn} \end{pmatrix} \begin{pmatrix} f_{i\to 1} \\ \vdots \\ f_{i\to i} \\ \vdots \\ f_{i\to n} \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ v_{ii} \\ \vdots \\ 0 \end{pmatrix} \right] \\ &= (\mathbf{I} + \lambda_0 \mathbf{I})^{-1} (\mathbf{T}^T \vec{f_i} + \vec{v_i}) \end{aligned} \tag{5}$$

where $\vec{v_i} = [0, 0, ..., v_{ii}, ..., 0]^T$ is a vector with only the i-th entry v_{ii} is nonzero, s.t.,

$$\sum_{k\in V} t_{ki} f_{i\to k} + v_{ii} = \alpha_i \tag{6}$$

which guarantees $f_{i\to i} = \alpha_i$ as described in Equation (2).
From Equation(5), after some calculation, we got

$$\vec{f_i} = (\mathbf{I} + \lambda_0 \mathbf{I} - \mathbf{T}^T)^{-1} \vec{v_i} \triangleq \mathbf{P} \vec{v_i} \tag{7}$$

where $\mathbf{I} + \lambda_0 \mathbf{I} - \mathbf{T}^T$ is invertible, so denote $\mathbf{P} = (\mathbf{I} + \lambda_0 \mathbf{I} - \mathbf{T}^T)^{-1}$, which is a $n \times n$ matrix. The inverse matrix here is not hard to caculate, because it statisfies the convergence condition for Gaussian Siedle method: positive-definite and diagonally dominant [10].

As $\vec{v_i}$ is a vector with only v_{ii} is nonzero, Equation (7) can be rewritten as $\vec{f_i} = v_{ii} \mathbf{P}_{\cdot i}$. Specifically, with Equation (2) and (7), we could get $v_{ii} = \frac{\alpha_i}{p_{ii}}$. Thus,

$$\vec{f_i} = \frac{\alpha_i}{p_{ii}} \mathbf{P}_{\cdot i} \tag{8}$$

Then, the total influence from node i to the entire network G should be

$$f_{i\to V} = \vec{f_i}^T \vec{e} = \sum_{j=1}^{n} f_{i\to j} = \frac{\alpha_i}{p_{ii}} \sum_{j=1}^{n} p_{ji} \tag{9}$$

Actually, this model is a linear approximation for the Independent Cascade Model[8]. In general IC model, computing influence spread $\sigma(S)$ is $\#P$-hard problem, which occurs in counting problem and corresponds to NP-hard problem in decision problem [11]. Using this linear approximation, although losing some accuracy, we gain the efficiency for computing influence spread. Another advatage of using this model will be further discussed in Section 3.2.3.

3.2.3 Influence Maximization (IM)

As I stated in the Section 2, I want to find a subset $S \subset V$ achieves the maximization of the total influence spread over S. The stochastic nature necessitates taking expectation of the influence spread. However, one trivial solution of this optimization problem is that S is exactly V. In reality, due to the time or budget limit, we always put a contraint on the cardinality of S, i.e., $|S| \leq K$, where K is an ineger smaller than n. The problem, therefore, is to find S, s.t.,

$$S = \underset{S \subset V, |S| \leq K}{\arg\max} \mathbb{E}\Big[\sigma(S)\Big] \tag{10}$$

In most general models, the information diffusion process is usually untractable. To deal with this stochastic optimization problem, Kempe proposed originally a method to get around the hardness of influence computation [12]. The basic idea is similar to *Sample Average Approximation* (SAA): use Monte Carlo simulations of the diffusion process to estimate the influence spread $\sigma(S)$. Given seed set S, we can simulate the randomized diffusion process with seed set S for R times. Each time we count the number of active nodes after the diffusion ends, and then take the average of these counts over the R times. Although this method can be highly accurate by increasing R, it is so inefficient, because in each search iteration, we need to go through all nodes $\omega \in V \setminus S$, and for each node ω we need to run R simulations. A large number of Monte Carlo simulations are needed, which is almost computational imfessible for large-scale networks.

To deal with the computational inefficiency, our model could avoid the problem of doing simulations exhaustedly. By using linear approximation, although the accuracy, to some extent, is sacrified , we gain the efficiency of computing information spread. Now, under our model, the optimization problem is

$$S = \underset{S \subset V, |S| \leq K}{\arg\max} \sum_{i \in S} f_{i \to V} \triangleq \underset{S \subset V, |S| \leq K}{\arg\max} f_{S \to V} \tag{11}$$

However, there is another limitation of taking expectation as in the traditional method: the set with the maximum influence spread may not result in the maximized information awareness on the network. To illustrate this phenomenon, an example is shown in Figure 3. Suppose that we just want to select the most influential node, i.e., $K = 1$. Node 1 and 2 are highlighted in black respectively in Figure 3 (a) and (b). From the simulated influence spread results for these two nodes, we may choose node 1, since it successfully made node 4, 5 and 6 active, and node 2 only influenced node 4 and 5. However, after closely observation, we can find that the influence spread distribution of node 2 is much more balanced than node 1. Therefore, it has stronger ability to influence more people than node 1. In this example, it may be better to choose node 2 rather than node 1, which is selected inappropriately by this method. The underlying reason of this drawback is that, the traditional method only optimizes over the expectation of total influence spread. Even if the influence is maximized, it doesn't mean that more people will become aware of the information. Thus, to deal with this problem, we could measure the information awareness in the entire network by directly using the probability of selected subset being influeced, i.e., $f_{i \to j}$. Treating this measurment as a relaxed definition of being influenced, we can interpret it as a measurement about really making reaction after being influenced (e.g. buying the new product). It makes sense in reality: the company cares more about the number of people who will purchase the new product after advertised, rather than the expected number of people who are just influenced . Details of how to measure will be discussed in the Section 3.

Under the consideration of computation issue and the purpose of this optimization problem in reality, another drawback in above way of optimization should not be ignored. We can find that

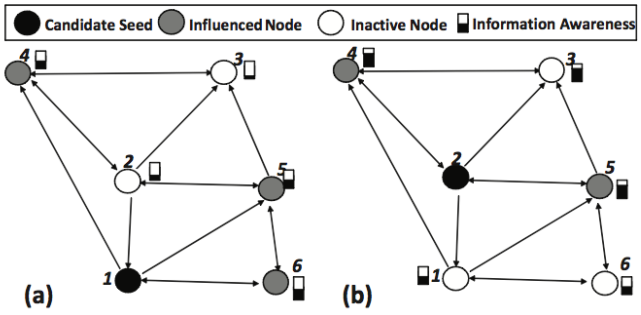

Figure 3: Motivation Example

some important nodes may not be selected under the criterion (11) (e.g. the customer who will buy the product with high probability). Thus, it is necessary to consider the subset including important nodes. I will also address this issue in the Section 3.

4 Two-step Algorithm

4.1 Problem Reformulation

As I stated in the Section 3, there are two main issues I want to deal with:*(1)*some important nodes should be included in S in order to achieve specific goals both in reality and computation, *(2)* new measurement of influence spread need to be proposed in order to avoid selecting unbalancedly distributed nodes.

For the first issue, inspired by the *Screening* method in variable selection [13]: firstly, screen the variable down to a smaller scale, then employ LASSO or SCAD to further select a subset of variable. Actually, this two-step idea is more and more common now, due to the explosion of data. Hence, I apply this idea to the network model: first, use some criterion to screening nodes down to some extent, then, do the optimization over that subset of nodes. The advantages are: *(1)* we can use specific criterion to measure the importance of the nodes. (Here, I will consider the marketing setting specificlly); *(2)* to cover the selected subset, it seems that more constraints are added, however, in fact, this makes optimization much easier, because we only need to explore over a smaller-scale subset.

For the second issue, intuitively, if we want to avoid taking expectation, we should operated over the influence spread directly. The influence spread $f_{i\to j}$ is, in fact, measuring how large the influence from i to j will probably be. If we put a thredshold on it, we can interpret the qualified $f_{i\to j}$ as the influence will actually make marketing effectively in the future. This is the real goal of marketing, which makes this step reasonable. Details will be discussed in Section 4.3.

4.2 First Step — Screening

4.2.1 Objective Function

Recall two drawbacks lying in former method: it cannot *(1)* cover important nodes *(2)* cover as more as possible nodes. In the setting of marketing, hence, we want to achieve these two goals: *(1)* relevance – we need a customer-level indicator to measure whether this person is worthy of propagating or not; *(2)* diversity – a set-level indicator to measure the selected subset covering

most segmentations of the importation customers. Then, the goal is to find a subset $U \subset V$, which can maximize the criterion

$$g(U) = \lambda \sum_{u \in U} r(u) + (1 - \lambda) H_0(U) \tag{12}$$

where $r(u)$ measures the relevance of customer u, $H_0(U)$ measures the diversity over entire set U, λ is the weight to balance these two aspects.

Noting that the problem of selecting a relevent as well as diverse subset occurs often in different settings. For example, in social media message selection problem, in order to enhance readers' reading experience, it is meaningful to select articles containing informative as well as diverse messages. Inspired by an efficient method in this setting [14], I will formulate similar measurements in the view of marketing in Section 4.2.2 and 4.2.3, and then, use the result in [13] to verify the optimization algorithm in Section 4.2.4.

4.2.2 Relevance

Define $r(u)$ To select the most relevant nodes in regarding of one specific product a, I adopt the Jaccard coefficient, which is widely used for measuring similarity. To define the extent of interest of a customer u in the product a, naively, $r(u)$ is

$$r(u) = \frac{|I_u \cap I_a|}{|I_u \cup I_a|} \tag{13}$$

where I_u denotes the set of u's interesting items, I_a the set of items similar to a. Hence, $r(u)$ can measure the importance of making propagation of a to u properly.

Submodularity of $R(U)$ In order to propose an effient algorithm regarding the objective function, analogy to the concavity in regular real-valued function, submodularity of the set-function is an critical property to optimize it globally. That is, a local optimal solution will result in global optimization, just as concavity. The strict definition is[7], a set-function $f : 2^V \to \mathbb{R}$ is called submodular if for all subsets $S \subset U \subset V$ and $u \in V \setminus U$ the following inequality holds:

$$f(S \cup \{u\}) - f(S) \geq f(U \cup \{u\}) - f(U) \tag{14}$$

From (14), we know that $R(U) \triangleq \sum_{u \in U} r(u)$ is submodular, since

$$R(S \cup \{u\}) - R(S) = r(u) = R(U \cup \{u\}) - R(U) \tag{15}$$

4.2.3 Diversity

Define $H_0(U)$ Using the similar definition of normalized entropy of the U in [16], we define our measurement of diversity. Denote f_i as binary feature variable: if constomer in U purchased i-th item, then $f_i = 1$, otherwise, $f_i = 0$, where $i \in I_a$. Define $H_0(U)$ as

$$H_0(U) = -\frac{\sum_{i=1}^{|I_a|} p(f_i = 1) \log p(f_i = 1)}{\log |I_a|} \tag{16}$$

The intuition is that, to improve the diversity of costomers, we'd like to add the users who haven't purchase similar items before, i.e., with smaller $p(f_i = 1)$.

Submodularity of $H_0(U)$ From[15], we know that the entropy of a sample is submodular. (The normalization doesn't effect the submodularity property.)

Setting λ λ in Equation (12) is used for balancing the effect between relevance and diversity: if the marketing strategy emphasizes on effectiveness, the larger λ is, the more accurate of targeting will be; on the other hand, if emphasizing on exploring new market, the smaller λ is, the more new customer will included.

4.2.4 Optimization Algorithm

Submodularity of $g(U)$ In order to find an optimal solution for the maximization problem (12), we could exhaustively search the space of all possible subsets. However this proves computationally prohibitive. From the submodularity of both $R(U)$ and $H_0(U)$, the linear combination of submodular function is still submodular function. This key property allows us to use a greedy algorithm [15] that provides efficient linear solutions within a deterministic error bound from the best possible solution.

Greedy Algorithm Under the submodularity of this optimization problem, we can apply general greedy algorithm directly. Initially, $U = \varnothing$. By computing $r(u)$ for each $u \in V$, we select a subset of nodes with the highest $r(u)$ and set this subset as U. Then, iterativelly, add a new node u into U, which could provide the biggest boost into objective function $g(U)$. This iterative process will keep running until the size of U is K_U, a given upper bound. (Attached gif. shows the screening process when $n = 5, \tilde{n} = 3$)

4.3 Second Step — L_1 relaxation

4.3.1 Reformulate the problem after screening

As I stated in the Section 4.1, the first issue is solved by first-step. Given a selected subset U, all we need to do is to set a new criterion to deal with unbalanced distributed situation and select a subset $S \subset U$. Denote $|U| \triangleq \tilde{n}$.

Instead of taking expectation of influence spread, to achieve a balance condition, I put a threshold t on $f_{i\to j}$. Introduce a new $\tilde{n} \times \tilde{n}$ matrix $\mathbf{R} = [R_{ji}]_{\tilde{n}\times\tilde{n}}$, where $R_{ji} = 1$ if $f_{i\to j} > t$, otherwise $R_{ji} = 0$. We can interpret $\mathbf{R}$ as a spread coverage matrix: for large enough $f_{i\to j}$, node j is highly probably influenced by i, which implies $R_{ji} = 1$. To address the second issue, we want to maximize the number of influenced people. Let $\vec{p}$ be a $\tilde{n} \times 1$ vector, where $p_i = 1$ if node i is selected, $p_i = 0$ if not. For node j, as long as there is one $R_{ji} = 1$ as well as i is selected, $\mathbf{R}_{j\cdot}\vec{p} \neq 0$, which means that j is influenced anyway. In sense of this, combining constraint on the cardinality of selected subset, the problem can be formulated as

$$\begin{aligned} \vec{p} &= \arg\max_{\vec{p}\in\{0,1\}^{\tilde{n}}} \|\mathbf{R}\vec{p}\|_0 \\ \text{s.t. } & \mathbf{1}^T\vec{p} = K \leq \tilde{n} \end{aligned} \tag{17}$$

4.3.2 Linear programming after relaxation

The optimization of L_0 norm is generally NP-hard, we can replace the L_0 norm with L_1 norm, as done in Lasso [17] . Therefore, we would like to maximize $\|\mathbf{R}\vec{p}\|_1$. Realizing that, the L_1 norm

optimization problem can be solved easily by linear programming. At last, our problem formulated as

$$\begin{aligned} \vec{p} = \arg\max_{\vec{p}} & \|\mathbf{R}\vec{p}\|_1 \\ \text{s.t. } & \mathbf{1}^T\vec{p} = K \leq \tilde{n} \\ & \mathbf{0} \leq \vec{p} \leq \mathbf{1} \\ & \vec{p} \in \mathbb{Z}^{\tilde{n}} \end{aligned} \tag{18}$$

5 Numerical Results

5.1 Comparision with traditional method

To address problem(10), the most famous tradtional method is to combine Monte Carlo and Greedy Algorithm [7]. Given seed set S, simulate the randomized diffusion process with S for R times; each time we take the average over the number of activated nodes to estime σS. Then, use Greedy Algorithm through the estimation.

Here, I compare my method with tradition one when $n = 5, 10, 15, 20$. The corresponding $K = 3, 6, 9, 12$. Generating the data, where t_{ij} is uniformly distributed and the distribution of rating score is Beta(2,2). For each n, I generated 100 networks, comparing the result of these two methods, it shows that, as n grows, my algorithm works better and better(Figure 4.). However, notice that, for small n, the screening step makes the selection less accurate.

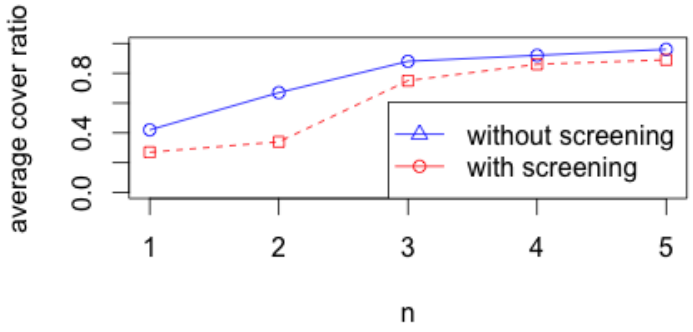

Figure 4: Accuracy for small n

5.2 Performance on large-scale network

When n is too large, the above traditional algorithm is almost computational infeasible. However, it shows that my algorithm can optimize the influence over entire network efficiently and stable. In the Figure 5, we can see that, even when $n = 500$, our algorithm still performs well.

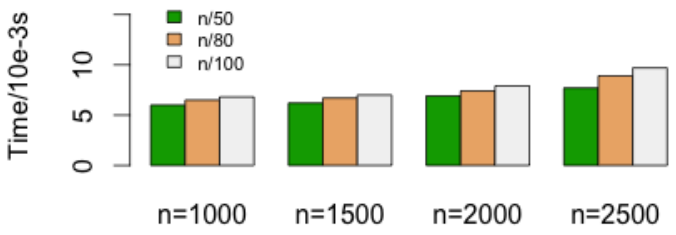

Figure 5: Efficiency for large n

6 Future Work

1. The numerical result shows that, the two-step algorithm is accurate for small-scale network, and efficient for large-scale network. To test whether this method works well on large-scale real-world data, further work is needed once we have required data.

2. To make linear approximation of the expected influence, theoretical prove can be done to show some properties, e.g., asymptotic unbiasedness.

References

[1] Nail, J. (2004). The consumer advertising backlash. Forrester research and intelliseek market research report.

[2] Domingos, P., & Richardson, M. (2001). Mining the network value of customers. ACM.

[3] Goldenberg, J., & Libai, B.,& Muller, E.(2001).Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing letters.

[4] Granovetter, M. (1978). Threshold models of collective behavior. American journal of sociology.

[5] Vazirani, V.(2001). Approximation algorithms. Springer.

[6] Jurvetson, S.(2000). What exactly is viral marketing? Red Herring.

[7] Chen, W., & Lakshmanan, L., & Castillo, C. (2013). Information and influence propagation in social networks. Morgan& Claypool.

[8] Yang, Y., & Chen, E. (2012). On approximation of real-world influence spread. Machine Learning and Knowledge Discovery in Databases.

[9] Xiang, B., & Liu, Q. (2013). PageRank with Priors: An Influence Propagation Perspective. ACM.

[10] Golub, G., & Van Loan, C. (1996). Matrix Computations. JHU Press.

[11] Wang, C., & Chen, W. (2012). Scalable influence maximization for independent cascade model in large-scale social networks. Data Mining and Knowledge Discovery.

[12] Kempe, D., & Kleinberg, J. (2003). Maximizing the spread of influence through a social network. ACM.

[13] Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica.

[14] Stajner, T., & Thomee, A. (2013). Automatic selection of social media responses to news. ACM.

[15] Nemhauser, L., & Wolsey, L. (1978). An analysis of approximations for maximizing submodular set functions. Mathematical Programming.

[16] Ko, C., & Lee, J. (1995). An exact algorithm for maximum entropy sampling. Operations Research.

[17] Tibshirani,R. (1996). Regression shrinkage and selection via the lasso. JRSSB.