Robust Nodes Selection for Influence Maximization

Yucheng Ding
517021910416
lecklas.d@sjtu.edu.cn

Abstract

We consider a practical phenomenon in the influencce maximization problem, where a fraction of the initial nodes are not successfully activated. Such a phenomenon would deteriorate the final influence of the selected initial nodes. To overcome the mentioned problem, we study the objective of robust nodes selection for the influence maximization, i.e, how to choose a set of initial nodes to maximize the expected final influence when some of them may be unactivated. We then propose a robust initial node selection (RNS) algorithm with a fast and accurate influence estimation method (IEM). In particular, RNS use IEM, which computes the activation probability of each node and estimates the expected influence, to ouput a robust seed by selecting some influencial nodes and a node seed with greate total influence. We extensively evaluate our algorithms over the facebook social network. Evaluation results demonstrate that IEM can accurately estimate the influence and RNS significantly outperforms the conventional greedy algorithm in terms of final influence when a fraction of initial nodes are not activated.

1 Introduction

Nowadays, a social network, which denotes the relationships and interactions within a group of individuals, plays a fundamental role as a medium for the spread of information, ideas, and influence among its members [Kempe et al.(2015)Kempe, Kleinberg, and Tardos]. Applications such as Facebook, Twitter, and Wechat allow people to connect and communication with each other at anytime and anywhere. As a consequence, viral marketing [Bass(2004), Steffes and Burgee(2009), Domingos and Richardson(2001), Mahajan et al.(1993)Mahajan, Muller, and Bass, Richardson and Domingos(2002)] via social network becomes increasing popular and significant for the industrial circle.

To handle the problem, [Kempe et al.(2015)Kempe, Kleinberg, and Tardos] formulates the influence maximization (IM) problem and adopts two basic models, independent casade (IC) and linear threshold (LT), to model the influence diffussion in the social network. In particular, the problem is, given a social network, select k most influencial nodes to maximize the final influence size under the diffussion model, where k denotes the size of the initial node seed.

However, in the real viral marketing, we have to consider the following practical characteristics: (1) *Unsuccessful initialization*: when some of the initial clients in the viral marketing have poor experience with the new products, they may not be willing to spread the information. That is, a fraction of the initial nodes may not be successfully activated. (2) *Decreasing diffusion efficiency*: as the information diffusion progresses, the diffusion efficiency decreases. For example, client A is an initial node, B is an out-neighbour of A, and C is an out-neighbour of B. For client B, the final source of information about the new product is A, who is his friend. But for C, the final source is a friend of his friend, whose trust-level is less than C's "direct" friends.

Such two characteristics are not including in the existing works. In what follows, we propose a new max-min objective fucnction to model the unsuccessful initialization and make a new assumption

SJTU EE447 Final Project.

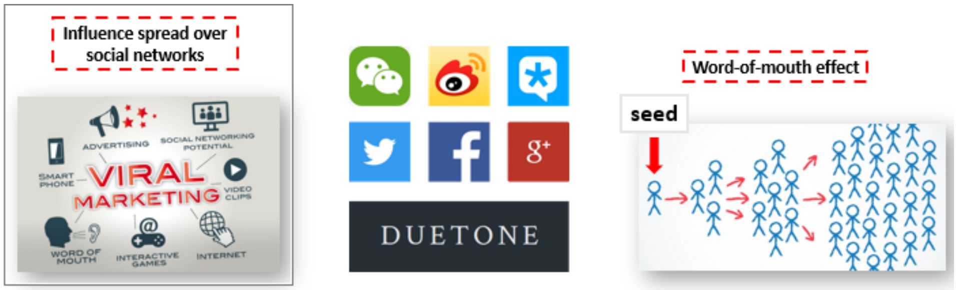


Figure 1: Influence Maximization Problem (Copied from the course PPT).

about information diffussion to model the decreasing diffusion efficiency in IC problem. The max-min objective aims to find such a *robust* initial seed: the successfully activated nodes have great influence even in the "worst" case, i.e, the most influencial nodes are not activated. And the new information diffussion assumption has covered the decreasing diffusion efficiency by introducing an decreasing coefficient.

Under our objective and diffussion model, we propose Robust Node Selection (RNS) algorithm, which adopts greedy strategy and a new influence estimation method (IEM). We first prove that the influence that our approach estimation is quite close to the real influence expectation. We then analyse the approximation ratio of RNS. Finally, we evaluate IEM and RNS over the facebook dataset. Evaluation results demonstrate the accuracy of IEM and the effectiveness of RNS with a remarkable performance improvement compared with the conventional greedy algorithm.

2 Problem Formulation

We consider an influence maximization problem based on the independent casade (IC) model. Given a social network $G(V, E)$, we need to choose a robust initial seed $\mathcal{S}$ ($|\mathcal{S}| \leq k$) to maximize the expected number of influenced nodes even in the case that at most m ($m \leq k-1$) nodes refuse to diffuse positive information.

We can formalize it into a max-min optimization problem:

$$\max_{\mathcal{S}\subseteq V, |\mathcal{S}|\leq k} \; \min_{\mathcal{H}\subseteq \mathcal{S}, |\mathcal{H}|\leq m} F\left(\mathcal{S} \setminus \mathcal{H}\right), \tag{1}$$

where $F(\mathcal{X})$ denotes the expected number of influnced nodes when the initially activative seed is $\mathcal{X}$.

Namely, we call the optimization problem robust nodes selection. To better simulate the information diffusion, solve the problem, and make approximation analysis, we make the following assumptions.

Assumption 1 (Independent Casade)**.** *The information diffuses based on the independent casade model. Starting with a set of inital active nodes, the information diffusses in descrete steps. At each step t, the newly activated node (say, node u) independently activates its out-neighbor (say, node v) with some probability $p(u, v)_t$.*

Assumption 2 (Decreasing Diffusion Efficiency)**.** *As the information diffusion progresses, the probability of successful transmission decreases exponentially. That is, $p(u, v)_t$ is not equal to $p(u, v)_{t-1}$, but satisfies $p(u, v)_t = \delta * p(u, v)_{t-1}$, where $\delta < 1$.*

Assumption 1 is based on the independent casade model, and Assumption 2 is based on the fact that the diffusion efficiency decreases as the information diffusion progresses, as introduced in Section 1.

We then formulise the information diffusion process. In the r-th step, each newly activated node u independently activates its out-neighbor v with probability $\delta^{r-1}p(u, v)_0$, where $p(u, v)_0 = p(u, v)$ is the original probability of u succesfully activating v, which is given as the information of the social network.

3 Algorithm Design

In this section, we propose a initial nodes selection algorithm for the robust nodes selection problem. The basic idea is that we separate the optimization into two parts: the first part is to select the

Algorithm 1 Robust Node Selection (RNS)

1: **Input:** Social influence graph $G(V, E)$, parameter k, m, influence estimation function $\hat{F}(\mathcal{X})$.
2: **Output:** Initial seed $\mathcal{S}$.
3: $\mathcal{S}_1, \mathcal{S}_2 \leftarrow \varnothing$;
4: **while** $|\mathcal{S}_1| < m$ **do**
5: $\quad s \leftarrow \arg\max_{s \in \mathcal{V} - \mathcal{S}_1} F(\{s\})$;
6: $\quad \mathcal{S}_1 \leftarrow \mathcal{S}_1 \cup \{s\}$;
7: **end while**
8: **while** $|\mathcal{S}_2| < k - m$ **do**
9: $\quad s \leftarrow \arg\max_{s \in \mathcal{S} - \mathcal{S}_1 - \mathcal{S}_2} F(\mathcal{S}_2 \cup \{s\})$;
10: $\quad \mathcal{S}_2 \leftarrow \mathcal{S}_2 \cup \{s\}$;
11: **end while**
12: $S = S_1 \cup S_2$;

influencial nodes, denoted as S_1, and the second part is to select a node seed with great total influence, denoted as S_2. The key idea is that if some of the nodes in S_2 are unactivated, the influencial nodes in S_1 will replace them, which makes up for the influence loss. In particular, we adopt a greedy strategy in each part.

The algorithm is presented in details in Algorithm 1. In this part (from lines 4 to 7), we continuesly select an element with the largest influence until m elements have been selected. And set $\mathcal{S}_2$ approximate the best set $\mathcal{S} - \mathcal{S}_1$ with $\mathcal{S}_1$ removed from $\mathcal{S}$. In this part (from lines 8 to 11), we continuesly select an element with the largest marginal influence until $k - m$ elements have been selected. And the selction of S_2 is the conventional greegy algorithm, which will obtian a node seed with great total influence. Finally, after $\mathcal{S}_1$ and $\mathcal{S}_2$ being selected, the algorithm will outputs a robust initial node seed $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$.

4 A Fast Influence Spread Estimation

To implement Algorithm 1, the main challenge is how to design the influence estimation function $\hat{F}(\mathcal{X})$. Many existing works estimates the influence using Monto Carlo simulation, which may repeat the indepedent casade process for thousands of time. To make a fast and accurate estimation, in this section, we propose a new influence estimation method (IEM). Instead of making Monto Carlo simulation, we estimate the probability that each nodes being activated. The final expected influence is the sum of the activation probability over all nodes. We first show the accurate expected influence calculation method, and then show that we could only focus on the first few steps to give a fast estimation of the expected influence. To speed up the estimation, we only calculate the first r_0 steps. And in Theorem 1, we shows that for each node, the difference between the accurate activation probability and the estimated activation probability cound be bounded.

4.1 Expected Influence

In this section, we introduce the methods of calculating the accurate probability of each node being activated. Let A_0 denote the initially active node seed, and I_v^r denote the probability of node v being activated at *exactly* the r-th step. First we initialize all I_v^0:

$$I_v^0 = \begin{cases} 1, & v \in A_0 \\ 0, & v \notin A_0 \end{cases} \tag{2}$$

When $r \geq 1$, for each node $v \in V - S$, we have

$$I_v^r = \left(1 - \sum_{i=0}^{r-1} I_v^i\right) \left(1 - \prod_{u \in \Gamma(v)} \left(1 - \delta^{r-1} I_u^{r-1} p\left(u, v\right)\right)\right), \tag{3}$$

where $\Gamma(v)$ denotes the in-neighbours of v. The first part calculates the probability that v has not been activated at the first $(r-1)$ steps, and the second part calculates the probability that v is activated at the r-th step by its in-neighbours, which was activated at exactly the $(r-1)$-th step.

Algorithm 2 Influence Estimation Method (IEM)

```
1: Input: Social influence graph G(V, E), initial node seed S, decreasing coefficient δ.
2: Output: Estimated infleucne IF.
3: Initialize I_v^r with for each node v step r(r ≤ t);
4: for Each v ∈ S do
5:     I_v^0 ← 1;
6: end for
7: for r = 1, 2, ··· , t do
8:     for Each node v ∈ V do
9:         I_v^r = (1 − Σ_{i=0}^{r−1} I_v^i)(1 − Π_{u∈Γ(v)} (1 − δ^{r−1} I_u^{r−1} p(u, v)));
10:    end for
11: end for
12: IF = Σ_{r=1}^{t} Σ_{v∈V} I_v^r;
```

So the total probability of node v being activated is

$$\text{Pr}(v) = \sum_{r=0}^{+\infty} I_v^r. \tag{4}$$

which is the sum of the probability of v being activated at all steps.

4.2 Influence Estimation Method

In this section, we propose a new influence estimation method with a error bound based on the influence calculation method mentioned in the above section. Instead of calculating the probabilities of all steps, we focus on the results coming from the first t steps, and show that the estimated probability is close to the accurate probability. The algorithm is presented in details in Algorithm 2. We calculate the probability that each node v is activated at the r-th step, where $r = 1, 2, \cdots, t$, according to equation (3) (from lines 3 to 11), and estimate the final influence (line 12).

We then prove that the difference between the estimated probability and the accurate probability of each node (say, node v) being activated is bounded.

Theorem 1. *When we only calculate the first t steps and $D\delta^{t/2-1} \leq 1$, where D is the maximum input degree in $G(V, E)$, the difference between the real probability and estimated probability is bounded by:*

$$Pr(v) - Pr_t(v) \leq \frac{\delta^{t+1}}{1-\delta},$$

Proof of Theorem 1. When $v \in \mathcal{S}$, it is trivial that $I_v^1 \leq \delta^0 = 1$. Then we focus on the relationship between the activated probability of the neighbouring two steps.

$$\begin{aligned} I_v^r &= \left(1 - \sum_{i=0}^{r-1} I_v^i\right)\left(1 - \prod_{u\in\Gamma(v)} \left(1 - \delta^{r-1} I_u^{r-1} p(u, v)\right)\right) \\ &\leq 1 - \prod_{u\in\Gamma(v)} \left(1 - \delta^{r-1} I_u^{r-1} p(u, v)\right) \\ &\leq \sum_{u\in\Gamma(u)} \delta^{r-1} I_u^{r-1} \leq D\delta^{r-1} I^{r-1}, \end{aligned}$$

where $I^{r-1} = \max_{u\in V}\{I_u^{r-1}\}$, so we have $I^r \leq D\delta^{r-1}I^{r-1}$. From the recurrence equation, we can bound the maximum probability that a node is activated at the r-th step: $I^r \leq D^r\delta^{r(r-1)/2}$. So the error, which is the probability that a node is activated after the t-th step is bounded:

$$\text{error(u)} \leq \sum_{r=t+1}^{\infty} D^r\delta^{r(r-1)/2} = \sum_{r=t+1}^{\infty} D^r\delta^{r^2/2}\delta^{-r/2}. \tag{5}$$

When $D\delta^{t/2-1} \leq 1$, we have $D^r \delta^{r^2/2} \delta^{-r/2} \leq \delta^r$, and

$$\text{error(u)} \leq \sum_{r=t+1}^{\infty} \delta^r = \frac{\delta^{t+1}}{1-\delta} \overset{t\to\infty}{\to} 0. \tag{6}$$

□

5 Approximation Analysis

In this section, we analyse the approximation ratio of Algorithm 1.

Theorem 2. *Algorithm 1 will output a set $\mathcal{A}$ satisfying*

$$\frac{\hat{F}(\mathcal{A} \setminus H^*(\mathcal{A}))}{\hat{F}(\mathcal{A}^* \setminus H^*(\mathcal{A}^*))} \geq (1 - \frac{1}{e})\frac{1}{k-m},$$

where $\hat{F}(\mathcal{X})$ is the influence estimation function, $H^*(\mathcal{A})$ is the worst removal of set $\mathcal{A}$, $\mathcal{A}^*$ is the optimal seed, and $H^*(\mathcal{A}^*)$ is the worst removal of set $\mathcal{A}^*$.

Proof Sketch of Theorem 2. We now give the outline of the proof, and details are given in the Appendix.

Lemma 1. *For an arbitrary instance of the information diffussion model under Assumptions 1 and 2, the resulting estimated influence function $\hat{F}$ is submodular.*

Lemma 2. *Having removed node set $\mathcal{A}_1$ from V, Algorithm 1 will choose a set $\mathcal{A}_2$, whose estimated influence is no less than $1 - \frac{1}{e}$ of the maximum estimated influence:*

$$\hat{F}(\mathcal{A}_2) \geq (1 - \frac{1}{e}) \max_{\mathcal{S} \subseteq V \setminus \mathcal{A}_1, |\mathcal{S}| \leq k-m} \hat{F}(\mathcal{S}).$$

Lemma 3. *We next prove that the maximum estimated influence after removing node set $\mathcal{A}_1$ is no less than the estimated influence of the optimal seed $\mathcal{A}^*$ with $H^*(\mathcal{A}^*)$ removed from it:*

$$\max_{\mathcal{S} \subseteq V - \mathcal{A}_1, |\mathcal{S}| \leq k-m} \hat{F}(\mathcal{S}) \geq \hat{F}(\mathcal{A}^* \setminus H^*(\mathcal{A}^*)).$$

Lemma 4. *The estimated influence of $\mathcal{A}$ returned by Algorithm 1 with $H^*(\mathcal{A})$ reomeved is no less than $1/(k-m)$ of the estimated influence of $\mathcal{A}_2$:*

$$\hat{F}(\mathcal{A} \setminus H^*(\mathcal{A})) \geq \frac{1}{k-m} \hat{F}(\mathcal{A}_2).$$

Based on the above lemmas, we can derive that

$$\begin{aligned}
\hat{F}(\mathcal{A} \setminus H^*(\mathcal{A})) \geq & \frac{1}{k-m} \hat{F}(\mathcal{A}_2) \\
\geq & \frac{1}{k-m} (1 - \frac{1}{e}) \max_{\mathcal{S} \subseteq V \setminus \mathcal{A}_1, |\mathcal{S}| \leq k-m} \hat{F}(\mathcal{S}) \\
\geq & (1 - \frac{1}{e}) \frac{1}{k-m} \hat{F}(\mathcal{A}^* \setminus H^*(\mathcal{A}^*)).
\end{aligned}$$

□

Theorem 2 shows that RNS obtains a set of initial nodes, whose influence is $(1 - \frac{1}{e})\frac{1}{k-m}$ times of influence of the optimal seed even in the worst removal case. So the seed is actually very robust, no matter which m nodes are not activated, it always has a great total influence.

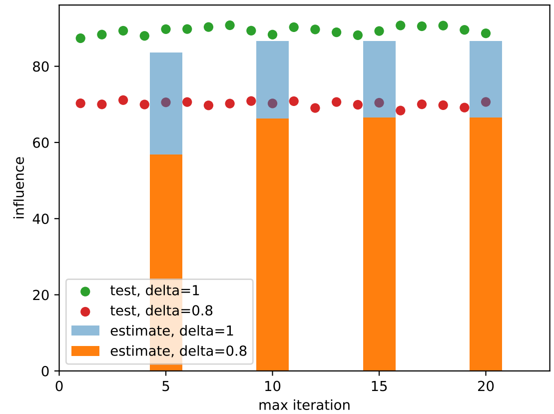
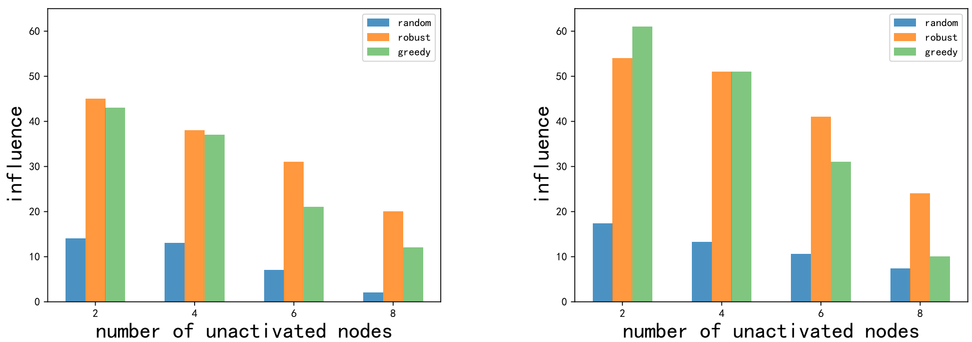


Figure 2: Influence Estimation Method Evaluation



(a) Robust Test with $\delta = 0.6$ (b) Robust Test with $\delta = 0.8$

Figure 3: Robust Influence Test with Different Decreasing Coefficient δ.

6 Experiments

In this section, we evaluate our algorithms. Due to the computation constraint of the computer, we use a small dataset, which is part of the facebook network and consists of 249 nodes and 407 edges. And we randomly allocate a weight $p \in (0, 1)$ to each edge.

Influence Estimatiion Method Evaluation. We first evaluate our influence estimation function. Under the setting that the number of steps $t = 5, 10, 15, 20$ and the decreasing coefficient $\delta = 0.8, 1.0$, we compare the estimation result with the experimental result. We use the result of RNS as the inital nodes and evaluate the influence estimation method. We first do the influence test for 20 times, for each experiment we repeatly execute the information diffusion process under our information difusion model for 100 times and calculate the average amount of the final influenced nodes. Each dot in Figure 6 denotes a experiment result. We then use IEM to estimate the result, which is displayed in Figure 6 in the form of bar. Figure 6 shows that IEM can well approximate the real influence when the max step t for estimation is greater than 10.

Robust Node Selection Evaluation. We then evaluate our robust seed selection algorithm by by comparing with the randomly chosen seed and conventional greedy algorithm. We set the size of the initial node seed k to 10, the number of unactivated initial nodes m to $2, 4, 6, 8$, and the decreasing coefficient δ to 0.8. Under each setting, we compare the test average influence for each algorithm. In particular, for each initial seed, we randomly remove m nodes from it and make the influence test, which is the average influence of 100 information diffusion experiment. We repeat the above operations for 5 times, and record the least test influence, which is to approximately find out the worst removal case. Figure 3 shows that RNS always outpeforms the random seed, and have better performance compared with the conventional greedy algorithm when the numer of unactivated initial nodes $m \geq 4$.

7 Conclusion

In this paper, we study the practical problem of robust nodes selection for the influence maximization. To handle this problem, we propose RNS, which outputs a robust node seed, and a new influence estimation method, IEM. We extensively evaluate our algorithms over a small facebook social network. Expirical studies over the facebook dataset demonstrate the accuracy of IEM and the effectiveness and robustness of RNS with a remarkable performance improvement compared with the conventional greedy algorithm.

References

[Kempe et al.(2015)Kempe, Kleinberg, and Tardos] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, 11: 105–147, 2015.

[Bass(2004)] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, 50(12-Supplement):1825–1832, 2004.

[Steffes and Burgee(2009)] Erin M. Steffes and Lawrence E. Burgee. Social ties and online word of mouth. *Internet Res.*, 19(1):42–59, 2009.

[Domingos and Richardson(2001)] Pedro M. Domingos and Matthew Richardson. Mining the network value of customers. In *KDD*, pages 57–66. ACM, 2001.

[Mahajan et al.(1993)Mahajan, Muller, and Bass] Vijay Mahajan, Eitan Muller, and Frank M. Bass. Chapter 8 new-product diffusion models. In *Marketing*, volume 5 of *Handbooks in operations research and management science*, pages 349–408. North-Holland, 1993.

[Richardson and Domingos(2002)] Matthew Richardson and Pedro M. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, pages 61–70. ACM, 2002.

A Proof of the Lemmas

Proof of Lemma 1. For an arbitrary node set T, a subset $S \subseteq T$, and any node v, we will prove that

$$\hat{F}(S \cup \{v\}) - \hat{F}(S) \geq \hat{F}(T \cup \{v\}) - \hat{F}(T), \tag{7}$$

where $\hat{F}$ is the expectation of the influence in the first t round. Let A_t denote the nodes activated at the first t steps. The diffussion is a random process, and we can view the process as two stage. At each step, the first stage is that each edge decides to exist or not based on the edge probability, the second step is that information diffusses according to the connection decided in the first stage. We will prove that $A_t(T \cup \{v\}) \setminus A_t(T) \subseteq A_t(S \cup \{v\}) \setminus A_t(S)$ when the connection of edges in the first t steps in the two graphs are the same.

For the t- step, we have

$$A_t(T \cup \{v\}) \setminus A_t(T) = A_t(\{v\}) \setminus A_t(T)),$$

and

$$A_t(S \cup \{v\}) \setminus A_t(S) = A_t(\{v\}) \setminus A_t(S))$$

where $A_t(S) \subseteq A_t(T)$. So we have

$$A_t(\{v\}) \setminus A_t(T)) \subseteq A_t(\{v\}) \setminus A_t(S)),$$

which indicates that

$$A_t(T \cup \{v\}) \setminus A_t(T) \subseteq A_t(S \cup \{v\}) \setminus A_t(S).$$

In the calculation of the expectation, the probability that two graphs reach the same connectio is equal and one-to-one, and for each connection,

$$|A_t(T \cup \{v\}) \setminus A_t(T)| \leq |A_t(S \cup \{v\}) \setminus A_t(S)|.$$

We take expectation on both sides and have equation (7), which indicates that the influence estimated function is submodular. □

Proof of Lemma 2. According to the property of submodular functions, we have the $(1 - \frac{1}{e})$ approximation ratio. That is:

$$\hat{F}(\mathcal{A}_2) \geq (1 - \frac{1}{e}) \max_{\mathcal{S} \subseteq V \setminus \mathcal{A}_1, |\mathcal{S}| \leq k-m} \hat{F}(\mathcal{S}).$$

□

Proof of Lemma 3. We first define $C_1 = A^* \cap A_1$, and then find another set $C_2 \subseteq A^* \setminus C_1$ where $|C_1| + |C_2| = m$. Since $H^*(A_*)$ minimize the influence $\hat{F}(A^* \setminus H^*(A^*))$, we have

$$\hat{F}(A^* \setminus H^*(A^*)) \leq \hat{F}(A^* \setminus (C_1 \cup C_2))$$

In addition, $A^* \setminus (C_1 \cup C_2) \subseteq V \setminus A_1$ and $|A^* \setminus (C_1 \cup C_2)| = k - m$, we have

$$\max_{\mathcal{S} \subseteq V - \mathcal{A}_1, |\mathcal{S}| \leq k-m} \hat{F}(\mathcal{S}) \geq \hat{F}(A^* \setminus (C_1 \cup C_2)) \geq \hat{F}(\mathcal{A}^* \setminus H^*(\mathcal{A}^*)).$$

□

Proof of Lemma 4. We will prove the lemma case by case:

1. If $H^*(A) = A_1$, we have $\hat{F}(A \setminus H^*(A)) = \hat{F}(A_2)$, so the lemma holds.
2. If $H^*(A) \neq A_1$, we have at least one node v in A_1 left. So we have

$$\hat{F}(\mathcal{A} \setminus H^*(\mathcal{A})) \geq \hat{F}(\{v\}) \geq \frac{1}{k-m} \sum_{u \in A_2} \hat{F}(\{u\}) \geq \frac{1}{k-m} \hat{F}(A_2).$$

□