

CS599: Structure and Dynamics of Networked Information (Spring 2005)
04/13/2005: Submodularity of $f(\cdot)$ in the Threshold Model
Scribes: Ramakrishna Gummadi and Shiva Kintali

In the last lecture, we investigated the existence or non-existence of node sets that would create a cascade of the entire graph in threshold models with fixed thresholds. From a practical (and algorithmic) viewpoint, the more interesting question is how to actually find them. We phrased this question as an optimization problem as follows: find an initial set of k nodes with the largest infection impact.

We were considering a simple greedy algorithm [1] for the problem of maximizing the expected number $f(S)$ of active nodes starting with set S . Our claim was that it is a $(1 - \frac{1}{e})$ -approximation. We proved the Theorem of Nemhauser and Wolsey, which says that the greedy algorithm is such an approximation for any monotone and submodular function. As monotonicity of f is straightforward, we now want to prove that $f(\cdot)$ is submodular. First, we will show this for a different (and easier to analyze) model of infection.

Independent Cascade Model: Given a complete directed graph G with edge probabilities p_e , we consider an infection model where once a node $u \in V(G)$ becomes active, it infects a neighboring node v with probability $p_{(u,v)}$. If the attempt succeeds, v becomes active; u , however, does not get to try infecting v again.

We can observe that for each edge e , we can decide ahead of time (randomly) if the activation attempt will succeed when/if it happens, with probability p_e . We observe that, in the graph G of “successful edges”, the nodes active in the end are exactly the ones reachable from S .

As a first step, we will show that the number of reachable nodes in a given graph G is a submodular function. We define $f_G(S)$ to be the number of nodes reachable from S in G , and prove

Claim 1 f_G is submodular for all G .

Proof. We need to show that $f_G(S+x) - f_G(S) \geq f_G(S'+x) - f_G(S')$ whenever $S \subseteq S'$. We write $R_G(S)$ for the set of all nodes reachable from S in G . Then, $f_G(S) = |R_G(S)|$, and because any node reachable from $S+x$, but not from S , must be reachable from x , we can observe that

$$|R_G(S+x)| - |R_G(S)| = |R_G(S+x) \setminus R_G(S)| = |R_G(\{x\}) \setminus R_G(S)|.$$

By monotonicity of R_G , we have that $R_G(\{x\}) \setminus R_G(S) \supseteq R_G(\{x\}) \setminus R_G(S')$, so $|R_G(\{x\}) \setminus R_G(S)| \geq |R_G(\{x\}) \setminus R_G(S')|$, which proves submodularity of f_G . ■

From the submodularity for any fixed graph, we would like to derive that the same holds for the function f , over randomly chosen graphs. To that end, we use the following useful way of rewriting a random variable's expectation $E[X]$: If $\{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_m\}$ is a partition of the probability space Ω , and X is a random variable on Ω , then:

$$E[X] = \sum_i \text{Prob}[\mathcal{E}_i] \cdot E[X | \mathcal{E}_i]$$

In our case, the random variable X is the number of nodes reached from S , and $f(S) = E[X]$ is its expectation. Let G_1, \dots, G_m be an enumeration of all graphs on n nodes (note that m is large). Notice that the events $\mathcal{E}_i = [G = G_i]$ form a partition of the probability space we are considering. Thus, the above identity implies that $f(S) = \sum_i \text{Prob}[G = G_i] \cdot f_{G_i}(S)$. Because all the coefficients $\text{Prob}[G = G_i]$ are non-negative, and each f_{G_i} is submodular, the submodularity of f now follows from the following

Fact 2 A non-negative linear combination of submodular functions is itself submodular.

Proof. Let $\alpha_i \geq 0$, $f = \sum_i \alpha_i f_i$. If each f_i is submodular, then for all $S \subseteq S'$,

$$\begin{aligned} f(S+x) - f(S) &= \sum_i \alpha_i f_i(S+x) - \sum_i \alpha_i f_i(S) \\ &= \sum_i \alpha_i (f_i(S+x) - f_i(S)) \\ &\geq \sum_i \alpha_i (f_i(S'+x) - f_i(S')) \\ &= f(S'+x) - f(S'). \end{aligned}$$

■

Notice that we did not use a lot of properties of the Independent Cascade Process. In fact, we proved the following stronger lemma:

Lemma 3 *If the activation process is such that we can define a distribution on graphs $\{G_i\}$ such that $f(S)$ equals the expected number of nodes reachable from S under the distribution, then f is submodular.*

We would like to use this lemma to prove submodularity for the Linear Threshold Model that we started out with. In order to do that, we need to come up with a distribution over graphs such that the requirements of the lemma are met. In this case, it is not as obvious which distribution to choose, but we will be able to show that the following model works:

Random Graph Model: Each node v has at most one incoming edge which emanates from u with probability $w_{(u,v)}$ (the weight of the influence of u on v — recall that $\sum_u w_{(u,v)} \leq 1$ for all v). With probability $1 - \sum_u w_{(u,v)}$, v has no incoming edge.

Claim 4 *This model is equivalent to the Threshold Model in the sense that the expected number of nodes reached is the same.*

Proof. We will prove by induction on t that for each time step $t \geq 0$, and any node sets $T \subseteq T'$, the probability that exactly T is active at time t and T' at time $t+1$ is the same in both the threshold and the random graph processes.

In the base case $t=0$, the probability is 1 for the pair (\emptyset, S) for both processes, and 0 for all other pairs, because both processes start with only the selected set S active.

For the inductive step, assume that $T \subseteq T'$ are the active sets at time $t-1$ and t , and consider some node $v \notin T'$. We investigate the probability that v becomes active at time $t+1$ in either process.

In the threshold model, because v was not active at time t , we know that $\theta_v \geq \sum_{u \in T} w_{(u,v)}$. However, subject to that, θ_v is still uniformly random, so by the Principle of Deferred Decisions, we can re-choose θ_v uniformly at random from the interval $(\sum_{u \in T} w_{(u,v)}, 1]$.

v becomes active at time $t+1$ iff $\theta_v \leq \sum_{u \in T'} w_{(u,v)}$. This happens with probability, $\frac{\sum_{u \in T' \setminus T} w_{(u,v)}}{1 - \sum_{u \in T} w_{(u,v)}}$.

Now, let us look at the probability that $v \notin T'$ becomes active at time $t+1$ in the random graph process. Because v is not active at time t , we know that v 's incoming edge, if any, does not come from T . Thus, v becomes active at time $t+1$ iff the edge comes from $T' \setminus T$.

The probability that v 's edge does not come from T is $1 - \sum_{u \in T} w_{(u,v)}$, and the probability that it comes from $T' \setminus T$ is $\sum_{u \in T' \setminus T} w_{(u,v)}$. Hence, the conditional probability of v becoming active is $\frac{\sum_{u \in T' \setminus T} w_{(u,v)}}{1 - \sum_{u \in T} w_{(u,v)}}$.

So for any individual nodes, the probability of activation at time $t+1$ is the same under both processes. As these decisions are independent in both processes (thresholds resp. edges are chosen independently), the probability that exactly all nodes from W become active is the same under both processes for any set W .

So,

$$\begin{aligned} & \text{Prob}[\langle T', T'' \rangle \text{ active at time } \langle t, t + 1 \rangle] \\ &= \sum_T \text{Prob}[\langle T, T' \rangle \text{ active at time } \langle t - 1, t \rangle] \cdot \text{Prob}[\text{exactly } T'' \setminus T' \text{ becomes active}] \end{aligned}$$

is the same for both. Hence, by induction, the processes behave the same, and in particular, the final expected number of activated nodes is the same. Hence, f is submodular. ■

References

- [1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence in a social network. In *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining*, pages 137–146, 2003.