Scribes: Shi Feng, Jakob de Raaij

Influence Maximization in Social Networks

Lecture 18

In the last lecture, we saw a model for the spread of a new idea or behavior in a social network starting from a set of *seeds*. We now study the *influence maximization problem*: Given a social network and a model of information spread, what seeds should we pick to make the idea or behavior spread as much as possible?

As a motivating example, consider a firm preparing to launch a new product. The firm wants to directly advertise the product to a small group of early adopters. The hope is that these early adopters (the seeds) will spark a cascade of adoptions throughout the network, leading to widespread use of the product. Given data on its customers' social connections, which individuals should initially be chosen to maximize the total spread?

1 The Influence Maximization Problem

Definition 1 (Influence Maximization Problem)**.** Let the finite graph $G = (V, E)$ represent a social network and consider some progressive[1] diffusion process on this graph which determines which nodes get activated. For any set of seeds $S \subseteq V$, let $P(S)$ be the set of nodes that are active at the end of the process, starting with seeds S — if the process is random, $P(S)$ is a random variable. Let the *influence function* $f(S)$ be the expected number of active nodes at the end of the process when starting at seeds S,

$$f(S) = \mathbb{E}[|P(S)|].$$

Given such a social network, diffusion process, and an integer k, the *influence maximization problem* is to find

$$S^* \in \arg\max_{S \subseteq V,\ |S|=k} f(S),$$

that is, to find a set of k seeds that maximizes the expected final number of adopters.

As one may suspect (and we will see later), for many diffusion processes the influence maximization problem is **NP**-hard. However, it turns out that the natural greedy algorithm has good theoretical guarantees — and is frequently almost optimal in practice — when the influence function f satisfies some fairly general properties. We will explore this now.

Definition 2 (Submodularity and Monotonicity)**.** Let $2^V = \{X : X \subseteq V\}$ be the *powerset* of V, i.e., the set containing all subsets of V. A *set function* $f : 2^V \to \mathbb{R}$ assigns a real $f(X)$ to every $X \subseteq V$.

- A set function f is *monotone* if for all $X, Y \subseteq V$, such that $X \subseteq Y$, it holds that

  $$f(X) \leq f(Y).$$

  In other words, adding an element to a set X never decreases the value of $f(X)$.

- A set function f is *submodular* if for all $X, Y \subseteq V$, such that $X \subseteq Y$, and any element $z \notin Y$,

  $$f(X \cup \{z\}) - f(X) \geq f(Y \cup \{z\}) - f(Y).$$

  In other words, the marginal increase of f from adding an element (z) decreases as more elements get added to the set (i.e., $X \to Y$).

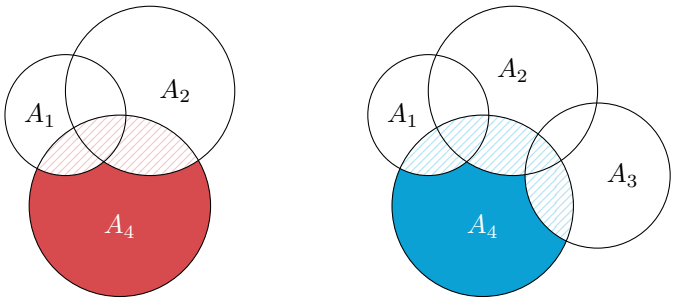
[1] Recall, progressive means that vertices only switch from being inactive to being active, and will stay active once they are activated once.

The definition of a monotone and submodular set function is more general than the influence maximization problem. Let's consider a different example:

Example 1 (Monotone Submodular Functions)**.** Let U be a universe and $A_1, \ldots, A_n \subseteq U$. Let $[n] = \{1, ..., n\}$. The *coverage function* $f : 2^{[n]} \to \mathbb{R}^+$ for each $S \subseteq [n]$ is the number of elements in U that are *covered*, i.e., included in at least one of A_i for $i \in S$,

$$f(S) = \Big|\bigcup_{i \in S} A_i\Big|.$$

The coverage function is monotone, since $\bigcup_{i \in X} A_i \subseteq \bigcup_{i \in Y} A_i$ whenever $X \subseteq Y$ — adding subsets cannot decrease the number of covered elements. The coverage function is also submodular since the number of additional elements covered due to a new set A_z cannot increase as other sets get added; that is, $A_z \setminus (\bigcup_{i \in X} A_i) \supseteq A_z \setminus (\bigcup_{i \in Y} A_i)$ whenever $X \subseteq Y$. To illustrate this, consider the figure below. Let $X = \{1, 2\}$ correspond to having selected A_1 and A_2 and let $Y = \{1, 2, 3\}$ correspond to additionally selecting A_3. Let's compare the impact that adding A_4 has on $f(X)$ and $f(Y)$:



$f(\{1,2\} \cup \{4\}) - f(\{1,2\})$ $\qquad$ $f(\{1,2,3\} \cup \{4\}) - f(\{1,2,3\})$

Adding 4 to $X = \{1, 2\}$ leads to the marginal gain shown in red; adding 4 to $Y = \{1, 2, 3\}$ leads to the marginal gain shown in blue. Since the area $A_3 \cap A_4$ is already part of the coverage of Y but not a part of the coverage of X, the additional new coverage due to A_4 is larger for X (red) than for Y (blue). Thus, $f(X \cup \{4\}) - f(X) > f(Y \cup \{4\}) - f(Y)$.

We'll now consider two additional coverage-based set functions on $2^{[n]}$,

$$f_1(S) = \mathbb{1}[1 \in S] \cdot \Big|\bigcup_{i \in S} A_i\Big| \qquad\qquad f_2(S) = \mathbb{1}[1 \in S] \cdot |A_1| \; + \; \Big|\bigcup_{i \in S} A_i\Big|.$$

First, note that both functions are monotone: Adding elements can only increase the number of covered elements, $\big|\bigcup_{i \in S} A_i\big|$, and set the indicator variable $\mathbb{1}[1 \in S]$ (being 1 if and only if $1 \in S$, else 0) from 0 to 1.

It turns out that f_1 is *not* submodular. As a counterexample, consider $X = \emptyset$, $Y = \{1\}$, and any $z \neq 1$. Then $X \subseteq Y$, $z \notin Y$, and

$$f_1(X \cup \{z\}) - f_1(X) = 0 - 0 < |A_1 \cup A_z| - 0 = f_1(Y \cup \{z\}) - f_1(Y),$$

violating the submodularity inequality. Hence f_1 is not submodular.

In contrast, f_2 is submodular. We can write $f_2(S) = g(S) + f(S)$ where

$$g(S) = \mathbb{1}[1 \in S] \cdot |A_1|, \qquad\qquad f(S) = \Big|\bigcup_{i \in S} A_i\Big|.$$

We already know that f, the coverage function, is submodular. We can check that function g is submodular since its value depends only on whether $1 \in S$: If $1 \in Y$, then for all $z \notin Y$ (so $z \neq 1$), $g(X \cup \{z\}) - g(X) = 0 = g(Y \cup \{z\}) - g(Y)$. If $1 \notin Y$, thus $1 \notin X$, then for all $z \notin Y$, $g(X \cup \{z\}) - g(X) = \mathbb{1}[z = 1] \cdot |A_1| = g(Y \cup \{z\}) - g(Y)$. Now, it is not too difficult to verify (and we will in fact prove this formally later) that the sum of two submodular functions is itself submodular.

The reason we defined monotone, submodular functions is that they can be maximized well with a greedy algorithm.

Theorem 1. *Let f be a monotone and submodular set function. The greedy algorithm that iteratively adds the element with largest marginal gain until it obtains a k-element set is a $(1 - 1/e)$-approximation algorithm for f. That is, the set S returned by the greedy algorithm satisfies*

$$f(S) \geq \left(1 - \frac{1}{e}\right) f(S^*),$$

where $S^* \in \arg\max_{S' \subseteq V,\, |S'|=k} f(S')$ is any optimal k-element set.

We will now consider two possible diffusion processes for the influence maximization problem. In both cases, solving for the optimal set of k seeds exactly is **NP**-hard. However, we will see that in both cases the influence function f is monotone and submodular, which allows us to use Theorem 1 to find that the greedy algorithm is a good approximation to the optimum.

2 The Independent Cascade Model

Definition 3 (Independent Cascade Model)**.** In the *independent cascade model*, there is a directed graph $G = (V, E)$ and for each edge $(i, j) \in E$ a weight $w_{ij} \in [0, 1]$. We may assume $w_{ij} = 0$ for all $(i, j) \notin E$. Given a set of seeds $S \subseteq V$, vertices get activated according to the following progressive process:

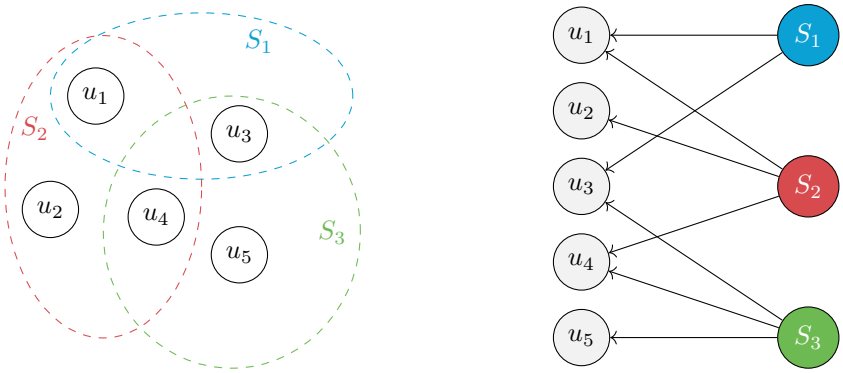
1. Initially all seeds are active and all other vertices are inactive.
2. In each round, every node i that was activated in the previous round tries to activate each currently inactive neighbor j, succeeding independently with probability w_{ij}. This node i will not (try to) activate any vertices in any subsequent round.
3. Once activated, a node remains active forever. The process terminates when no node is activated in a round.

Theorem 2. *Under the independent cascade model, the influence maximization problem is **NP**-hard.*

Proof. We show **NP**-hardness by reducing from the set cover problem, known to be **NP**-complete:

> *Set Cover.* Given a universe $U = \{u_1, \ldots, u_t\}$ and $S_1, ..., S_m \subseteq U$, does there exist a cover of U of size k, i.e., $C \subseteq [m], |C| = k$ so that $\bigcup_{j \in C} S_j = U$?

Given an instance of set cover, we construct a bipartite directed graph with a node u_i on the left side for each $i \in [t]$ and a node S_j on the right side for each $j \in [m]$. We draw a directed edge of weight 1 from S_j to u_i whenever $u_i \in S_j$.



There exists a cover of size k in the set cover instance if and only if there exists a set of k seeds S that leads to $f(S) = t + k$ active vertices in the influence maximization instance. We prove both directions.

If there is a cover C of size k, then choose the corresponding vertices on the right side as seeds, $S = \{S_j : j \in C\}$. Since for every element of the universe $u_i \in U$, it holds that $u_i \in S_j$ for some $j \in C$, any node on the left side will be attempted to be activated by at least one seed. Weight 1 edges guarantee activation, so this leads to all t universe vertices being activated, yielding $t + k$ total active vertices. The process ends right after, so $f(S) = t + k$.

Conversely, if there is a set of k seeds S such that $f(S) = t + k$, then all vertices in S are from the right side of the graph: Since all m vertices on the right side have no incoming edges, they will never be activated if they aren't a seed. Since there are only t vertices on the left, all k seeds need to be right vertices for there to be $t + k$ active vertices at the end of the process. By the same argument, all t vertices on the left must be active at the end of the process. Since any vertex on the left can only get activated by a seed on the right, this implies that any vertex u_i on the left has an incoming edge from a vertex $S_j \in S$. Thus, if we set $C = \{j : S_j \in S\}$, we know that for any $u_i \in U$, it holds that $u_i \in S_j$ for some $S_j \in C$. Therefore, C is a cover of size k.

Since this is a polynomial time reduction from the known **NP**-complete set cover problem to the influence maximization problem, we can conclude that the influence maximization problem is **NP**-hard. □

Theorem 3. *Under the independent cascade model, the influence function f is monotone and submodular.*

In order to prove this theorem, we will first formally show that the sum of submodular functions is still submodular, as already claimed earlier.

Lemma 1. *Let $f_1, \dots, f_r : 2^V \to \mathbb{R}$ be submodular set functions and let $c_1, \dots, c_r \geq 0$. Then, the set function*

$$f(S) = \sum_{i=1}^{r} c_i f_i(S)$$

is also submodular.

Proof. Consider any $X, Y \subseteq V$ with $X \subseteq Y$, and an element $z \notin Y$. Then

$$\begin{aligned}&\big[f(X \cup \{z\}) - f(X)\big] - \big[f(Y \cup \{z\}) - f(Y)\big] \\ &\quad = \sum_{i=1}^{r} c_i \Big[\big(f_i(X \cup \{z\}) - f_i(X)\big) - \big(f_i(Y \cup \{z\}) - f_i(Y)\big)\Big].\end{aligned}$$

Since each f_i is submodular, each bracketed term in the second line is ≥ 0. Since also $c_i \geq 0$, the entire sum is nonnegative. It follows that $f(X \cup \{z\}) - f(X) \geq f(Y \cup \{z\}) - f(Y)$, so f is submodular. □

Proof of Theorem 3. To prove that under the independent cascade model the influence function $f(S) = \mathbb{E}[|P(S)|]$ satisfies diminishing returns, we use that $\mathbb{E}[|P(S)|]$ can be decomposed into the possible realizations of the randomness in the model.

In particular, for any $R \subseteq E$, let

$$\Pr[R] = \prod_{(i,j) \in R} w_{i,j} \prod_{(i,j) \in E \setminus R} (1 - w_{i,j})$$

denote the probability that an attempt at activation works for all edges in R and does not work for all edges in $E \setminus R$. Define

$$f_R(S) = \big|\{\, v : v \text{ is reachable from } S \text{ via edges in } R\}\big|.$$

Equivalently, v counts as 1 towards $f_R(S)$ if and only if v is activated from set of seeds S if activation is successful exactly along edges in R.

For each fixed R, f_R counts the nodes reachable from S. If for each $v \in V$, we define $A_v \subseteq V$ to be the set of vertices reachable from v via edges in R, then we can equivalently define $f_R(S) = \big|\bigcup_{v \in S} A_v\big|$. This is exactly a coverage function, which we know is monotone and submodular from Example 1.

To illustrate this, assume the edges in R are as shown in the graphs below. For $X = \{v\}$, the vertices in blue in the left graph are A_v, i.e., the set of vertices that gets activated if v is a seed. Similarly, the vertices in orange on the left are A_z. Thus, if we add z to X, the coverage function f increases by the number of

orange vertices, 3. Now, for $Y = \{v, u\} \supseteq X$, the vertices in $A_v \cup A_u$ that will be activated if u and v are seeds are shown in blue on the right. If we add z to Y, the coverage function f only increases by 2, since one vertex in A_z is already in $A_v \cup A_u$.

Since it holds that

$$f(S) = E[|P(S)|] = \sum_{R \subseteq E} \Pr[R] \cdot f_R(S),$$

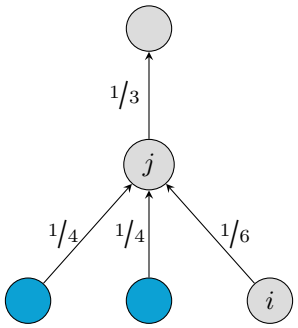
f is a nonnegative weighted sum of monotone, submodular functions. It immediately follows that f is monotone. Lemma 1 tells us that f is submodular. □

3 The Linear Threshold Model

We now consider a random diffusion process that is more similar to the cascade model we saw in the last lecture, in which each vertex is activated if a certain number of their neighbors are active.

Definition 4 (Linear Threshold Model)**.** In the *linear threshold model*, there is a directed graph $G = (V, E)$ and for each edge $(i, j) \in E$ a weight $w_{ij} \in [0, 1]$ so that for any vertex $j \in V$, the sum of weights on incoming edges is at most 1. That is, for all $j \in V$, $\sum_{i \in V} w_{ij} \leq 1$, where we again set $w_{ij} = 0$ for all $(i, j) \notin E$. Each vertex $j \in V$ has a *threshold* θ_j drawn uniformly at random from $[0, 1]$. A node $j \in V$ is activated if the total weight of its active neighbors is at least its threshold, so when $\sum_{i \in V} w_{ij} \cdot \mathbb{1}[i \text{ is activated}] \geq \theta_j$.

Example 2 (Linear Threshold Model)**.** Consider the following graph with edge weights written next to the corresponding edge. Assume the set of seeds S is shown in blue.



Let's try to calculate $f(S)$, the expected number of nodes that will be active at the end of the process. First, note that the two blue seeds are always active. The total weight of the active (incoming) neighbors of j is $1/4 + 1/4 = 1/2$, so

$$\Pr[j \text{ activated}] = \Pr[\theta_j \leq 1/2] = 1/2.$$

If j is activated, its only outgoing neighbor, the top vertex, is activated with probability $1/3$. Thus, we get that

$$f(S) = 2 + \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{3} = \frac{8}{3}.$$

Next, let's calculate the probability that j is activated when i is added to the set of seeds, if we know that it is inactive when only the blue vertices are seeds. Since we know that j was inactive initially, we know

that its threshold must exceed $1/2$. If i is active as well, j's total incoming weight becomes $1/4 + 1/4 + 1/6 = 2/3$. Thus,

$$\Pr[j \text{ activated} \mid j \text{ inactive initially}] = \Pr[\theta_j \leq \tfrac{2}{3} \mid \theta_j > \tfrac{1}{2}] = \frac{\frac{2}{3} - \frac{1}{2}}{1 - \frac{1}{2}} = \tfrac{1}{3}.$$

Similarly to the independent cascade model, the influence maximization problem is computationally hard while the influence function f is monotone and submodular, implying that the greedy algorithm is a $(1 - 1/e)$-approximation to f:

Theorem 4. *Under the linear threshold model, the influence maximization problem is **NP**-hard.*

Theorem 5. *Under the linear threshold model, the influence function f is monotone and submodular.*

The difficulty in proving Theorem 5 is that, unlike in the independent cascade model, the set function f_Θ for a realization $\Theta = (\theta_i)_{i \in V}$ of the randomness in the process is not necessarily submodular itself. For example, consider vertex j in Example 2 and assume its threshold realized as $\theta_j = 2/3$. Let X be the set of one of the two blue vertices and let Y be the set containing both. The marginal benefit of adding an element i to X is 0 since j will not be activated either way. However, the marginal benefit of adding i to Y is (at least) 1 since now vertex j is activated. Thus, f_Θ for $\theta_j = 2/3$ is not submodular. We now sketch the proof of how we can show that f is submodular despite this.

Proof sketch of Theorem 5. The key idea of this proof is to define a different random diffusion process and show that it is equivalent to the linear threshold model. We can write f as the non-negative weighted sum of functions that arise from the possible realizations of randomness in this equivalent process and show that each of those functions is monotone and submodular.

Let's start by defining the alternative process. At the beginning of the process, each vertex j selects one or none of its incoming edges, where it chooses edge (i, j) with probability w_{ij} and chooses no edge with probability $1 - \sum_{i \in V} w_{ij}$. An inactive vertex j that selected edge (i, j) is activated in round t if and only if i is active at the beginning of round t.

We'll now show that this process is equivalent to the linear threshold model. Let A_t denote the set of active nodes at the end of round t. Under the linear threshold model, the probability that a vertex $j \in V$ is activated in round $t + 1$ (and thus inactive in round t) is

$$\Pr[j \in A_{t+1} \mid j \notin A_t] = \frac{\Pr[j \in A_{t+1} \wedge j \notin A_t]}{\Pr[j \notin A_t]} = \frac{\Pr[\sum_{i \in A_t} w_{ij} \geq \theta_j > \sum_{i \in A_{t-1}} w_{ij}]}{\Pr[\theta_j > \sum_{i \in A_{t-1}} w_{ij}]} = \frac{\sum_{i \in A_t \setminus A_{t-1}} w_{ij}}{1 - \sum_{i \in A_{t-1}} w_{ij}}.$$

Under the alternate process, j is activated in round $t + 1$ exactly if its chosen edge (i, j) comes from one of the nodes i that were activated at round t. The probability that j picked an edge from a node that was activated in a round before t is $\sum_{i \in A_{t-1}} w_{ij}$. The probability that j picked an edge from a vertex i that was activated in round t is $\sum_{i \in A_t \setminus A_{t-1} w_{ij}}$. Thus, we get that also under the alternate process,

$$\Pr[j \in A_{t+1} \mid j \notin A_t] = \frac{\sum_{i \in A_t \setminus A_{t-1}} w_{ij}}{1 - \sum_{i \in A_{t-1}} w_{ij}}.$$

Hence, the two processes are equivalent and lead to the same distribution over final sets of active nodes.

For any realization $R \subseteq E$ of chosen edges in the alternative process, the set of nodes reachable from a set of seeds S, the function $f_R(S)$, is a coverage function, so we know it is monotone and submodular. We can write

$$f(S) = \mathbb{E}[|P(S)|] = \mathbb{E}_R[f_R(S)] = \sum_R \Pr[R] \cdot f_R(S)$$

where each f_R is monotone and submodular and each $\Pr[R]$ is non-negative. Thus, f is a non-negative weighted sum of monotone, submodular functions. Again, we immediately get that f itself is monotone, and we can apply Lemma 1 to get that f itself is also submodular. □