Harvard CS 1360, Spring 2025
Instructor: Ariel Procaccia
Scribe: Shi Feng

Social Networks 2: Influence Maximization

Lecture 18

In this lecture, we study the *influence-maximization problem*: given a social network and a diffusion model, choose a seed set S of size k to maximize the expected number of activated nodes, denoted $f(S)$. Although finding the optimal S is NP-hard under both the Independent Cascade and Linear Threshold models, f is monotone and submodular. Hence the simple greedy algorithm achieves a $(1-1/e)$-approximation. In this lecture we define f, prove its hardness and submodularity, and discuss the resulting approximation guarantee for influence maximization.

1 Influence Functions

Imagine a firm preparing to launch a new product. Armed with data on its customers' social connections, the firm can directly market only to a small group of early adopters. The hope is that these seeds will spark a cascade of adoptions throughout the network — but which individuals should be chosen to maximize total uptake?

Formally, let $G = (V, E)$ be a finite social network and assume a progressive diffusion process (nodes switch only from inactive to active). For any seed set $S \subseteq V$, define the influence function

$$f(S) \;=\; \mathbb{E}\big[\,|A_\infty(S)|\,\big],$$

where $A_\infty(S)$ denotes the set of active nodes once the cascade terminates starting from seeds S. The *influence-maximization problem* asks:

$$\max_{S\subseteq V,\ |S|=k} f(S),$$

i.e., select k seeds to maximize the expected final number of adopters.

2 Submodularity

A set function $f : 2^V \to \mathbb{R}$ is *submodular* if for all $X \subseteq Y \subseteq V$ and any element $z \notin Y$,

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

Intuitively, adding an element yields diminishing marginal returns as the set grows. Moreover, f is *monotone* if $X \subseteq Y$ implies $f(X) \leq f(Y)$.

Theorem 1. *If f is monotone and submodular, then the greedy algorithm that iteratively adds the element with largest marginal gain produces a k-element set S satisfying*

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

where S^* is any optimal k-element seed set.

2.1 Example: Coverage Functions

Let U be a universe and $A_1, \ldots, A_n \subseteq U$. The coverage function $f : 2^{[n]} \to \mathbb{R}^+$ such that

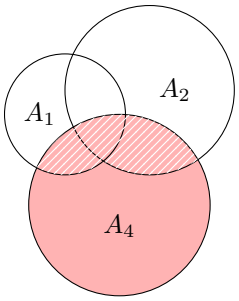
$$f(S) = \Big|\bigcup_{i\in S} A_i\Big|$$

is monotone and submodular. In the figure below, adding A_4 to $S = \{1, 2\}$ yields a larger marginal gain than adding it to $T = \{1, 2, 3\}$:

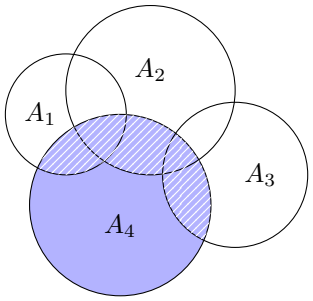
$$f(\{1,2\} \cup \{4\}) - f(\{1,2\}) > f(\{1,2,3\} \cup \{4\}) - f(\{1,2,3\}).$$

Define two set-functions on $S \subseteq [n]$:

$$f_1(S) = \mathbf{1}_{1\in S} \cdot \Big|\bigcup_{i\in S} A_i\Big| \quad f_2(S) = \mathbf{1}_{1\in S} \cdot |A_1| \;+\; \Big|\bigcup_{i\in S} A_i\Big|.$$



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



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

Poll 1: Which function is monotone submodular?

Answer: Only f_2 is monotone submodular.

Proof. For monotonicity, if $X \subseteq Y$, then adding elements can only enlarge unions or trigger the indicator, so $f_1(X) \leq f_1(Y)$ and $f_2(X) \leq f_2(Y)$.

For submodularity, we prove a counterexample for f_1. Let $X = \{2\}$, $Y = \{2, 3\}$, and $z = 1$. Then $X \subseteq Y$, $z \notin Y$, and

$$f_1(X \cup \{z\}) - f_1(X) = |A_1 \cup A_2| < |A_1 \cup A_2 \cup A_3| = f_1(Y \cup \{z\}) - f_1(Y),$$

violating the submodularity inequality. Hence f_1 is not submodular.

Next we prove f_2 is submodular. Write $f_2(S) = g(S) + h(S)$ where

$$g(S) = \mathbf{1}_{1 \in S} |A_1|, \quad h(S) = \Big|\bigcup_{i \in S} A_i\Big|.$$

It is straightforward to verify by case analysis that g is monotone and satisfies diminishing returns, and h is the standard coverage function (monotone submodular). Therefore their sum f_2 is also monotone submodular (see Lemma 1 below). □

3 Independent Cascade Model

Let $G = (V, E)$ be a directed graph where each edge $(i, j) \in E$ carries a weight $w_{ij} \in [0, 1]$, and we denote $w_{ij} = 0$ for all $(i, j) \notin E$. The diffusion proceeds in discrete rounds according to the following progressive process:

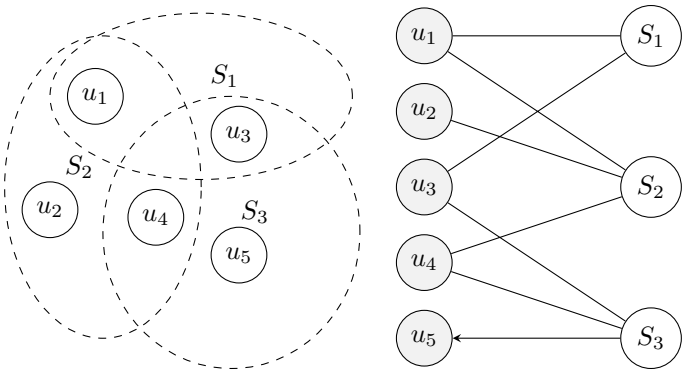
1. Start with a seed set $S \subseteq V$ of active nodes; all others are inactive.
2. In each round, every node i that became active in the previous round gets exactly one chance to activate each currently inactive neighbor j, succeeding independently with probability w_{ij}.
3. Once activated, a node remains active forever. The process terminates when no further activations occur.

Theorem 2. *Under the Independent Cascade model:*

- *Influence maximization* ($\max_{|S|=k} f(S)$) *is NP-hard.*
- *The influence function f is monotone and submodular.*

Proof of NP-hardness. We reduce from Set Cover. Given universe $U = \{u_1, \ldots, u_t\}$ and collection $\{S_1, \ldots, S_m\}$, construct a bipartite directed graph:

$$\text{Left side: } \{S_1, \ldots, S_m\}, \quad \text{Right side: } \{u_1, \ldots, u_t\}.$$



Place a directed edge $(S_j \to u_i)$ of weight 1 whenever $u_i \in S_j$.

Choosing a seed set of size k corresponds exactly to selecting k subsets in Set Cover. If there is a cover of size k, then activating those k seeds immediately activates all t element-nodes (weight-1 edges guarantee activation), yielding $t + k$ total active nodes. Conversely, if no cover of size k exists, any seed set of size k leaves at least one element inactive, so strictly fewer than $t + k$ nodes become active.

Hence influence maximization solves Set Cover and is NP-hard. □

Proof of submodularity. We prove a lemma at first.

Lemma 1. If $f_1, \ldots, f_r$ are submodular set-functions and $c_1, \ldots, c_r \geq 0$, then

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

is also submodular.

Proof. Fix any $X \subseteq Y \subseteq V$ and 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 is ≥ 0, and $c_i \geq 0$. Hence the entire sum is nonnegative, proving that f satisfies the diminishing-returns inequality and is submodular. □

We prove that under the Independent Cascade model the influence function $f(S) = \mathbb{E}[|A_\infty(S)|]$ satisfies diminishing returns by first fixing all randomness in advance.

For each directed edge (i, j) flip a biased coin (with success probability w_{ij}) once and record whether the edge is "live". Let α denote one fixed outcome of all $2^{|E|}$ coin flips. Define

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

Equivalently, $v \in f_\alpha(S)$ if and only if there exists a directed path of live edges from some seed in S to v.

For each fixed α, f_α counts the nodes covered (reachable) by seeds S. This is exactly a coverage function, which is known to be monotone and submodular. See Figure 1 for an illustration.

Since

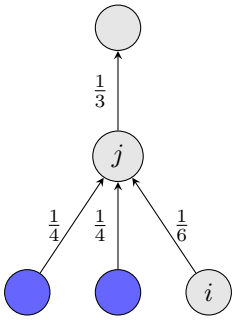
$$f(S) \;=\; \sum_{\alpha} \Pr[\alpha] \cdot f_\alpha(S),$$

f is a nonnegative weighted sum of submodular functions. By the standard lemma, f itself is monotone submodular. □

Figure 1: Under a fixed live-edge realization α, adding the orange seed alone (left) activates three nodes, whereas adding it on top of the blue seed set (right) activates only two. This illustrates both monotonicity (more seeds never decrease reach) and diminishing returns (marginal gain shrinks), proving that each f_α is submodular.

4 Linear Threshold Model

In the linear threshold model, each node j has incoming edge-weights $w_{ij} \in [0, 1]$ satisfying $\sum_i w_{ij} \leq 1$. Each node draws a threshold $\theta_j \sim \text{Uniform}(0, 1)$. An inactive node j becomes active as soon as the total weight of its active neighbors meets or exceeds θ_j. As a warm-up, consider the following example.



Poll 2: What is $f(S)$?

Answer: $f(S) = \frac{8}{3}$.

Proof. In the example, S consists of the two blue seeds. They are always active (+2). Node j has total incoming weight $1/4 + 1/4 = 1/2$, so $\Pr[j \text{ activates}] = \Pr[\theta_j \leq \frac{1}{2}] = \frac{1}{2}$. If j activates, its only outgoing neighbor (top) becomes active with probability $1/3$. Hence $f(S) = 2 + \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{3} = \frac{8}{3}$. □

Poll 3: Given that j is inactive, probability it becomes active after i does?

Answer: The probability is $\frac{1}{3}$.

Proof. Given j was inactive initially, its threshold must exceed $1/2$. After the orange seed activates, j's total incoming weight becomes $1/4 + 1/4 + 1/6 = 2/3$. Thus

$$\Pr[j \text{ activates} \mid j \text{ inactive initially}] = \Pr\left[\theta_j \leq \tfrac{2}{3} \,\middle|\, \theta_j > \tfrac{1}{2}\right] = \frac{\frac{2}{3} - \frac{1}{2}}{1 - \frac{1}{2}} = \frac{1}{3}.$$

□

As in the Independent Cascade case, the influence-maximization problem remains NP-hard, yet its influence function f is monotone and submodular.

Theorem 3. *Under the Linear Threshold model:*

- *Influence maximization ($\max_{|S|=k} f(S)$) is NP-hard.*
- *The influence function f is monotone and submodular.*

A difficulty is that, unlike the idependent cascade model, the influence function f_α is not submodular under the linear threshold model for fixed coin flips α. For example, consider a node 4 with threshold $3/4$ and three incoming edges from nodes $1, 2, 3$ with weight $1/3$ each. The marginal benefit of 3 to $\{1\}$ is 1, but the marginal benefit of 3 to $\{1, 2\}$ is 2 (as node 4 is activated). Nevertheless, f is submodular, as we sketch below.

Proof sketch of submodularity. Equivalently, each node j may pre-select exactly one "live" incoming edge (i, j) with probability w_{ij} (or no edge with probability $1 - \sum_i w_{ij}$). In each round t, if i becomes active for the first time, then any j whose chosen live edge is (i, j) immediately becomes a candidate to activate in round $t + 1$.

Let A_t denote the set of active nodes at the end of round t. Under the threshold-based process,

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

Under the live-edge process, j activates in round $t + 1$ exactly if its chosen edge (i, j) comes from one of the nodes i that became newly active at round t. The probability of this event matches the threshold-based expression. Hence both processes induce identical distributions over final active sets.

Once all edges are chosen or not chosen (fixing one "realization" of the random selection), the set of nodes reachable from S is effectively a coverage function, which is known to be monotone and submodular. Writing

$$f(S) \;=\; \mathbb{E}_\alpha[f_\alpha(S)],$$

where each f_α is submodular, we conclude f is a nonnegative weighted sum of submodular functions. Therefore f itself is monotone submodular. □