# CMSC5706 Topics in Theoretical Computer Science

## Week 11: Influence Maximization on Social Networks

Instructor: Shengyu Zhang

# Location change for the final 2 classes

- Nov 17: YIA 404 (Yasumoto International Academic Park 康本國際學術園)
- Nov 24: No class.
  - Conference leave.
- Dec 1: YIA 508 (Yasumoto International Academic Park 康本國際學術園)

# Social network

- Extensively studied by social scientists for decades.
  - Usually small datasets.
- Social networks on Internet are gigantic
  - Facebook, Twitter, LinkedIn, WeChat, Weibo, …
- A large class of tasks/studies are about the *influence and information propagation*.
- A typical task: *select some seed customers and let them influence others*.

# Motivating examples

- Adoption of smart phones.
  - Good: easy access to Internet, many cool apps, etc.
  - Bad: expensive, absorbing too much time, …
- Once you start to use smart phones, it's hard to go back.
  - There are not even many choices of traditional phones.
- Similar adoption: Religion, new idea, virus, …
- This lecture focuses on progressive models: once a node becomes active, it stays active.
  - There are also non-progressive models.

# Popular models

- Social network: a directed graph $G = (V, E)$.
- Note that the edges are directed:
  - How much an individual $u$ can influence another individual $v$ is generally different than how much $v$ can influence $u$. --- Just think of stars and fans.
- We consider the scenario where the diffusion proceeds in discrete time steps.

- Each node $v$ has two states: *inactive* and *active*.
  - *inactive*: the node hasn't adopted smart phones.
  - *active*: the node has adopted smart phones.
- Start from $S_0$, a *seed set*.
  - All nodes in $S_0$ are active.
- Nodes in $S_0$ influence some of their neighbors, who then become active.
  - Who are exactly the influenced ones depends on the variant of the model.

# model

- These new active nodes further influence some of their neighbors, and so on,
- until no more nodes are influenced, reaching a set $S_{\text{final}}$.
  - "Final active set".
- For a social graph $G = (V, E)$, a *stochastic diffusion model* specifies how active sets $S_t$, for all $t \geq 1$, is generated, given the initial seed set $S_0$.

# Model 1: IC

- *Independent cascade* (IC) model.

- Every edge $(u, v) \in E$ has an associated *influence probability* $p(u, v) \in [0,1]$•
  - Specifying the extent to which node $u$ can influence node $v$.

- For each time step $t \geq 1$, the set $S_t$ is generated as follows.
  - For each node $v \in S_{t-1} \backslash S_{t-2}$, for each edge $(v, u) \in E$ where $u$ is inactive, $u$ becomes active with probability $p(v, u)$.
  - This $u$ is then put in set $S_t$ of active nodes in time $t$.
  - Different edges influence independently.
- For each inactive node $u$, if it has many neighbors $v \in S_{t-1} \backslash S_{t-2}$: as long as one such $v$ successfully influences $u$, $u$ becomes active.

# An equivalent model

- Given a graph $G = (V, E)$, we mark each edge $(u, v)$ of $G$ as either live or blocked.
  - $\Pr[live] = p(u, v)$.
- The subgraph $G_L = (V, E_L)$ where $E_L$ contains all the live edges.
- The step-$t$ active set is $R^t_{G_L}(S_0) = \{v\text{: reachable from } S_0 \text{ within } t \text{ steps}\}$
- The final active set is defined as

$$R_{G_L}(S_0) = R^{n-1}_{G_L}(S_0) = \{v\text{:} \mathit{reachable\ from}\ S_0\}$$

- This model is equivalent to the IC model.
- In IC, each edge $(u, v)$ is "used" only once.
  - Flip a coin to decide whether the edge "works".
  - Success with probability $p(u, v)$.
- Thus we can just flip all the coins at the beginning, and then later follow the outcomes.

# Model 2: LT

- *Linear threshold* (LT) model.
- In many situations, multiple and independent sources are needed for an individual to be convinced to adopt some idea.
- E.g. Seeing 1/3 of your friends using smart phone, you made the decision.

- In linear threshold model, every edge $(u, v) \in E$ has a influence weight $w(u, v) \in [0,1]$,
- indicating the importance of $u$ on influencing $v$.
- The weights are normalized s.t. $\forall v$, the sum of weights of all incoming edges is at most 1

$$\sum_{u:(u,v)\in E} w(u, v) \leq 1$$

- Each node $v$ has a threshold $\theta_v$,
  - model the likelihood that $v$ is influenced by its active neighbors.
  - A large value of $\theta_v$ means that many active neighbors are required in order to activate $v$.
- Specifically, $v$ is activated if

$$\sum_{\substack{u:\, u \text{ active}, \\ (u,v)\in E}} w(u,v) \geq \theta_v$$

- Recall $\sum_{u:(u,v)\in E} w(u,v) \leq 1$.
- Since $\sum_{u:(u,v)\in E} w(u,v)$ may be smaller than $1$, it's possible that $v$ can't be activated even if all its in-neighbors are active.
- Corresponding to the people who just don't want smart phones.

# Diffusion process in LT model

- First each node $v$ independently selects a threshold $\theta_v$ uniformly at random in $[0,1]$.
- At each time step $t \geq 1$,
  - Set $S_t = S_{t-1}$
  - For each inactive node $v$, if the total weight of the edges from its active in-neighbors is at least $\theta_v$, i.e. $\sum_{u:(u,v)\in E} w(u,v) \geq \theta_v$, then $v$ becomes active (and is added into $S_t$).
- Note: all the randomness is in the threshold selection. Once this is done, the rest diffusion process is all deterministic.

# An equivalent model: LT'

- Similar to IC model, LT also has an equivalent model, in which the live edges are selected at the beginning.
- For each $v \in V$, among all incoming edges $(u, v) \in E$, we will select at most one to be live.
- $(u, v)$ is the one with probability $w(u, v)$.
- The set $E_L$ of live edges gives a subgraph $G_L = (V, G_L)$.

- If $\sum_{u:(u,v)\in E} w(u,v) < 1$, there is a chance that no incoming edge is live.
  - which happens with probability
$$1 - \sum_{u:(u,v)\in E} w(u,v)$$
- For any $t \geq 1$, the active set $S_t$ is set to be $R_{G_L}^{t}(S_0)$.
  - The set reachable from $S_0$ within $t$ steps in $G_L$.
- The final active set is $R_{G_L}(S_0) = R_{G_L}^{n-1}(S_0)$.

- This new model is equivalent to the LT model:
- Suppose that at time $t$, the current active set is $S_{t-1}$, and we want to show that any node $v$ is activated at the same probability as in LT model.
- Suppose $A$ is the set of active incoming neighbors. ($A = S_{t-1} \cap N^{in}(v)$)
- In LT: $v$ is activated with probability $\sum_{u \in A} w(u, v)$.
- In LT': $v$ is reached from $A$ if some $u \in A$ is selected to be the live incoming neighbor of $v$, which happens with probability $\sum_{u \in A} w(u, v)$.

# task

- Suppose we have a budget $k$ for seeds.
  - That is, $|S_0| = k$.
- The main task is to find a seed set $S_0$ so that it influence as many other nodes as possible.
- Since the influence propagation is a random process, we like to maximize

$$\sigma(S_0) = \mathbf{E}[|S_{\text{final}}|]$$

  - the expectation of size of final active set.

# Monotonicity

- $\sigma(S_0)$ as a function of set $S_0$ has two important properties.
- Definition. A function $f$ on subsets of $V$ is *monotone* if

  for any subsets $S \subseteq T \subseteq V$, $f(S) \leq f(T)$.
- *Theorem*. $\sigma(S_0) = \mathbf{E}[|S_{\text{final}}|]$ *(as a function of set $S_0$) is monotone*.
- This is pretty intuitive: More seeds generate more active nodes (in expectation).

# Submodularity

- Definition. A function on subsets of $V$ is *submodular* if $\forall S \subseteq T \subseteq V$ and $\forall v \in V \setminus T$,
  $$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$
- *Diminishing marginal returns:*
  marginal contribution of $v$ when added to $T$
  $\leq$ marginal contribution of $v$ to a smaller $S \subseteq T$.
- A dollar to a millionaire counts less than a dollar to a beggar.
- *Theorem*. $\sigma(S_0) = \mathbf{E}[|S_{\text{final}}|]$ *(as a function of set* $S_0$*) is submodular*.

- Fact: linear combination of submodular functions with non-negative coefficients is also submodular.
  - Set of submodular functions is closed under non-negative linear combination.
- $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$ --- (1)
- $g(S \cup \{v\}) - g(S) \geq g(T \cup \{v\}) - g(T)$ --- (2)
- Consider $h = af + bg$ where $a, b \geq 0$.
- $(1) * a + (2) * b$ gives

$$h(S \cup \{v\}) - h(S) \geq h(T \cup \{v\}) - h(T)$$

- *Theorem*. $\sigma$ *is submodular*.
  - $\sigma(S_0) = \mathbf{E}[|S_{\text{final}}|]$.
- *Proof*. Consider the equivalent model of selecting subgraph $G_L$ at the beginning.
- Since $\sigma(S_0) = \sum_{G_L:\text{subgraph of } G} \Pr[G_L] \left|R_{G_L}(S_0)\right|$, a non-negative linear combination of $\left|R_{G_L}(S_0)\right|$ for different subgraphs $G_L$ of $G$.
- It's enough to prove submodularity for $\left|R_{G_L}(S_0)\right|$, for each fixed $G_L$.

- Recall that $R_{G_L}(S_0)$ contains vertices reachable from $S_0$.
- We'll show that for any $S \subseteq T$, it holds that $R_{G_L}(T \cup \{v\}) \backslash R_{G_L}(T) \subseteq R_{G_L}(S \cup \{v\}) \backslash R_{G_L}(S)$ (*)
  - which implies $\left|R_{G_L}(T \cup \{v\})\right| - \left|R_{G_L}(T)\right| \leq \left|R_{G_L}(S \cup\right.$

# Influence maximization

- Our task of influence maximization:
Given a social graph $G = (V, E)$, a stochastic diffusion model, a budget $k$, find a seed set $S_0 \subseteq V$ with $|S_0| \leq k$ to maximize $\sigma(S_0)$.
- Namely, find an $S_0 \in argmax_{S_0 \subseteq V, |S_0| \leq k} \sigma(S_0)$.
- A related problem of influence spread computation: Given $G$, a diffusion model, and a seed set $S_0$, compute $\sigma(S_0)$.

# Bad news: #P-hard

- *Theorem. Both influence maximization and influence spread computation problems are #P-hard.*
  - In both IC and LT models.
  - #P-hard even if $|S_0| = k = 1$.
- Recall NP-complete problem SAT: decide whether a given CNF formula has a satisfying assignment.
- #P-complete problem #SAT: decide how many satisfying assignments does a given CNF formula have.
- Clearly #P is harder than NP: If one can count the number of solutions, then it's trivial to see whether it is 0 or not.

# Good news

- The hardness comes from two sources
  - Combinatorial nature.
  - Influence computation.
- The first can be partly overcome by a greedy approximation algorithm.
- The second can be overcome by Monte Carlo simulation.

# Greedy algorithm

- **Input**: $(k, f)$, where $k \in \mathbb{N}$, and $f$ is a monotone and submodular set function
- **Output**: a subset $S$

*Algorithm:*

- $S = \emptyset$
- **for** $i = 1$ **to** $k$ **do**
  - Take any $u \in \operatorname{argmax}_{w \in V-S}\big(f(S \cup \{w\}) - f(S)\big)$
  - $S = S \cup \{u\}$
- **return** $S$

- Basically, in each round $i$, we take an element $u$ with a largest marginal contribution to $f$ with respect to the current $S$.
- Repeat this until we select $k$ elements.
- *Theorem. The algorithm outputs a set $S$ with*

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

*where $f(S^*)$ is the optimal value*

  - $f(S^*) = max_{|S_0| \leq k} f(S_0)$.

- *Proof*. Suppose the algorithm selects the elements $s_1, s_2, \dots, s_k$ in that order,
- and an optimal solution is $S^* = \{s_1^*, s_2^*, \dots, s_k^*\}$.
- Let $S_i = \{s_1, s_2, \dots, s_i\}$ and $S_i^* = \{s_1^*, s_2^*, \dots, s_i^*\}$.
- Since $f$ is monotone, we have $f(S^*) \leq f(S_i \cup S^*) = f(S_i \cup S_{k-1}^* \cup \{s_k^*\})$.
  - Note that by our notation, $S^* = S_{k-1}^* \cup \{s_k^*\}$.

- $f(S^*) \le f(S_i \cup S^*) = f(S_i \cup S^*_{k-1} \cup \{s^*_k\})$.
- Recall submodularity:
  $f(S \cup \{v\}) - f(S) \ge f(T \cup \{v\}) - f(T)$.
- Take $S = S_i \subseteq T = S_i \cup S^*_{k-1}$, and $v = s^*_k$, we have

$$f(S_i \cup S^*_{k-1} \cup \{s^*_k\}) \le f(S_i \cup \{s^*_k\}) - f(S_i) + f(S_i \cup S^*_{k-1})$$

- Greedy algorithm selects the max marginal contribution, so

$$f(S_i \cup \{s^*_k\}) - f(S_i) \le f(S_i \cup \{s_{i+1}\}) - f(S_i)$$

- $S_i \cup \{s_{i+1}\}$ is just $S_{i+1}$. Thus

$$\begin{aligned} f(S^*) &\le f(S_i \cup \{s^*_k\}) - f(S_i) + f(S_i \cup S^*_{k-1}) \\ &\le f(S_{i+1}) - f(S_i) + f(S_i \cup S^*_{k-1}). \end{aligned}$$

- $f(S_i \cup S^*) \leq f(S_{i+1}) - f(S_i) + f(S_i \cup S^*_{k-1})$.
- Applying this argument on $f(S_i \cup S^*_{k-1})$, we have

$$f(S_i \cup S^*_{k-1}) \leq f(S_{i+1}) - f(S_i) + f(S_i \cup S^*_{k-2})$$

- Repeat $k$ times, we have

$$f(S^*) \leq f(S_i \cup S^*) \leq k\big(f(S_{i+1}) - f(S_i)\big) + f(S_i).$$

- Rearranging it yields

$$f(S_{i+1}) \geq \left(1 - \frac{1}{k}\right) f(S_i) + \frac{f(S^*)}{k}$$

- $f(S_{i+1}) \geq \left(1 - \frac{1}{k}\right) f(S_i) + \frac{f(S^*)}{k}$
- Multiply both sides by $(1 - 1/k)^{k-i-1}$, list the inequality for all $i$, and sum them up. We get

$$
\begin{aligned}
f(S) = f(S_k)& \\
&\geq \sum_{i=0}^{k-1} \left(1 - \frac{1}{k}\right)^{k-i-1} \frac{f(S^*)}{k} \\
&= \left(1 - \left(1 - \frac{1}{k}\right)^k\right) f(S^*) \\
&\geq \left(1 - \frac{1}{e}\right) f(S^*),
\end{aligned}
$$

as claimed.

- Recall that $\sigma(\cdot)$ as a set function is monotone and submodular.

- Apply this greedy algorithm enables us to find a seed set $S_0$ with $\sigma(S) \geq \left(1 - \frac{1}{e}\right)\sigma(S^*)$,
  - *where* $\sigma(S^*)$ *is the optimal value* $max_{S_0 \subseteq V, |S_0| \leq k} \sigma(S_0)$.

- But there is one catch.
- In the algorithm we used this step:
  - Take any $u \in \mathrm{argmax}_{w \in V-S}\left(\sigma(S \cup \{w\}) - \sigma(S)\right)$
- But $\sigma$ is hard to compute!
- Solution: Monte Carlo simulation.
- For any seed set $S_0$, run the diffusion process starting from $S_0$ enough number of times to get a good estimate to $\sigma(S_0)$.

# Putting everything together

- With details omitted, here is the final result.

- *Theorem. We have an algorithm with parameters* $(n, k)$ *achieving* $(1 - \frac{1}{e} - \epsilon)$*-approximation ratio in time* $O(\epsilon^{-2} k^3 n^3 m \log n)$*, for both IC and LT models.*

# Summary

- Two diffusion models: IC and LT.
- Influence maximization and influence spread computation problems are both #P-hard.
  - In IC and LT models.
- There exist $\left(1-\frac{1}{e}-\epsilon\right)$-approximation algorithms with polynomial running time.
  - In IC and LT models.