Maximizing the Spread of Influence through a Social Network

Han Wang
Department of Computer Science
ETH Zürich
Problem Example 1: Spread of Rumor

- 2012 = end!
Problem Example 2: Viral Marketing

- ezPad 1 beats iPad 3
Problem Definition

- G: a social network (n nodes)
- Model: spread process
- S: initially active subset (k seeds)
- $\sigma(S)$: #final active nodes (achievement)

- Task: Choose $S^*$
- Goal: $\sigma(S^*) = \max \sigma(S)$ NP-Hard

Realistic Goal:
Approximate the maximum with a guarantee
Choose S: $\sigma(S) \geq r \cdot \sigma(S^*)$
Contents in This Talk

- **G**: a social network (n nodes)
- **Model**: spread process
- **S**: initially active subset (k seeds)
- **\( \sigma(S) \)**: #final active nodes (achievement)

**Task**: Choose \( S^* \)

**Goal**: \( \sigma(S^*) = \max \sigma(S) \)

Realistic Goal:
Approximate the maximum with a guarantee
Choose \( S \): \( \sigma(S) \geq r \cdot \sigma(S^*) \)

**Prove**: Two Models

**Prove**: NP-Hard
Model 1: Independent Cascade Model
Model 1: Cascade Model

- Each active node try to activate his neighbors
  - $p_{u,v}$
  - $1 - p_{u,v}$
  - Only a single chance

![Diagram showing activation probabilities](image)
Model 1: Cascade Model
Model 1: Cascade Model

- $S = \{A, C\}$, $\sigma(S) = 5$
Model 2:
Linear Threshold Model
Model 2: Threshold Model

- Each inactive node picks a random $\theta_v \in [0,1]$
  - Active condition: $\sum_{u: \text{active neighbor of } v} b_{u,v} \geq \theta_v$

```
Iteration 2: 0.2 < 0.3
b_{C,D} = 0.2

Iteration 4: E → active
b_{E,D} = 0.7

Iteration 5: 0.2+0.7 > 0.3
D → active
```

\[
\theta_D = 0.3
\]
Model 2: Threshold Model

Iteration: 2

A connected to B: $\theta = 0.5$

B connected to C: $\theta = 0.5$

C connected to D: $\theta = 0.3$

C connected to E: $\theta = 0.6$

C connected to F: $\theta = 0.9$

D connected to E: $\theta = 0.6$

E connected to F: $\theta = 0.9$
Model 2: Threshold Model

- $S = \{A, C\}, \quad \sigma(S) = 4$
How to Prove the Guarantee?

Given a spread model

find $S$, s.t. \( \sigma(S) \geq r \cdot \sigma(S^*) \)

find $S$, s.t. \[ f(S) \geq (1 - \frac{1}{e}) \cdot f(S^*) \]

Nemhauser

\( f(S) \): Non-negative monotone submodular
Submodularity

- $U$: a finite ground set
- $P(U)$: power set of $U$
- $f(\cdot): P(U) \to R^*$

Submodularity: \( \forall \text{ node } v, \forall S \subseteq T \)

\[
f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)
\]
Example: Submodularity

- \( f(S) \): number of vertexes reachable from vertexes in S
How to Prove the Guarantee?

Given a spread model

find $S$, s.t.
$\sigma(S) \geq r \cdot \sigma(S^*)$

$\sigma(S)$ is Submodular

Prove:

Nemhauser

find $S$, s.t.
$f(S) \geq (1 - \frac{1}{e}) \cdot f(S^*)$

$f(S)$: Non-negative monotone submodular
We Want to Prove…

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma(S)$ is Submodular</th>
<th>NP-hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Cascade</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Linear Threshold</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Prove:
Submodularity

Cascade Model
Submodularity (Cascade Model)

- Recall: flip coin
Submodularity (Cascade Model)

- Why not flip all the coins in the beginning?
Submodularity (Cascade Model)

- Live edges $\rightarrow$ live paths
- Blocked edges

![Graph Diagram]

- Edge weights:
  - AB: 0.4
  - BC: 0.3
  - CD: 0.2
  - DE: 0.7
  - CF: 0.6
  - BE: 0.8
Simplify Cascade Model

Node $v$ ends up active

A live path: some seed $\rightarrow v$
Achievement (Simplified Model)

- **X**: coin flipping outcome
  - e.g. X1, X2

- **R_X(ν)**
  - R_{X1}(A) = \{A, B\}
  - R_{X1}(C) = \{C, D, E\}

- **σ_X(S) = | \bigcup_{ν∈S} R_X(ν) |**
  - σ_{X1}(\{A, C\}) = |\{A, B, C, D, E\}| = 5
Submodularity (Cascade Model)

- Fix $x$, $\sigma_x(S)$ is submodular

- Linear combination of submodular functions is still submodular

\[
\sigma(S) = \sum_x \text{Prob}[X] \cdot \sigma_x(S)
\]
Summary of the proof

Active = Has a live path

\( \sigma_x(S) \) is submodular

\( \sigma(S) \) is submodular
Prove: NP-hard

Simplified Cascade Model
NP-Hard (Cascade Model)

- Set Cover Problem: k subsets cover all?
  - K=1: No
  - K=2: No
  - K=3: Yes
  - K=4: ...
NP-Hard (Cascade Model)

- Solve Set Cover
  Q: 2 subsets cover all?

- Influence maximization
  Q: $|S| = 2, \sigma(S) \geq 2 + 5$?
Influence Maximization Problem

is at least as difficult as

Set Cover Problem
Prove:
Submodularity

Linear Threshold Model
Recall: Threshold Model

\[ \theta = 0.5 \]

\[ \theta = 0.3 \]

\[ \theta = 0.6 \]

\[ \theta = 0.9 \]
Gamble: Roulette
Gamble: Roulette

\[ \theta = 0.4 \]
Submodularity (Threshold Model)

\[ \theta = 0.5 \]

\[ \theta = 0.3 \]

\[ \theta = 0.6 \]

\[ \theta = 0.9 \]
Submodularity (Threshold Model)

- Live edges $\rightarrow$ live paths
Correctness of Simplification

For node $v$:

$$P(\text{active in Iteration } t + 1 \mid \text{inactive in Iterations } \leq t) = \frac{P(\text{active in Iteration } t + 1)}{P(\text{inactive in Iterations } \leq t)}$$
Simplified Model

- **Active before iteration 5**
- **becomes active in iteration 5**

Diagram:
- Node **V** is connected to nodes N1, N2, N3, N4, N5, and N6.
- Node V is active before iteration 5 and becomes active in iteration 5.
- The diagram shows the probabilities of transitioning to each node from node V:
  - N1: 0.2
  - N2: 0.15
  - N3: 0.07
  - N4: 0.23
  - N5: 0.1
  - N6: 0.14

Pie chart:
- The chart shows the distribution of nodes:
  - N1: 36%
  - N2: 24%
  - N3: 7%
  - N4: 23%
  - N5: 10%
  - N6: 14%

Legend:
- Green: Node active before iteration 5.
- Yellow: Node becomes active in iteration 5.
- Gray: None.
Simplified Model

\[ A_t: \text{Nodes becoming active in iteration } t \]

\[
\frac{\sum_{u \in A_t} b_{u,v}}{1 - \sum_{u \in A_1 \cup A_2 \cup \ldots \cup A_{t-1}} b_{u,v}}
\]
Original Model

<table>
<thead>
<tr>
<th>N2</th>
<th>N6</th>
<th>N4</th>
<th>N3</th>
<th>N1</th>
<th>N5</th>
<th>None</th>
</tr>
</thead>
</table>

Graph:
- N6 to V: 0.14
- N1 to V: 0.2
- N2 to V: 0.15
- N5 to V: 0.07
- N3 to V: 0.23
- N4 to V: 0.1
$A_t$: Nodes becoming active in iteration $t$

$$
\frac{\sum_{u \in A_t} b_{u,v}}{1 - \sum_{u \in A_1 \cup A_2 \cup \cdots \cup A_{t-1}} b_{u,v}}
$$
Simplify Threshold Model

Node v ends up active

A live path: some seed $\rightarrow v$
Similarly, we have...

Active = Has a live path

$\sigma_X(S)$ is submodular

$\sigma(S)$ is submodular
Prove:
NP-hard

Linear Threshold Model
NP-Hard (Threshold Model)

- Vertex Cover Problem
  - $k$ vertexes ($S$)

  each edge
  is incident to
  at least one vertex in $S$
NP-Hard (Threshold Model)

- Vertex Set Cover
  Q: 3 vertexes cover all?

- Influence maximization
  Q: $|S| = 3, \sigma(S) = 6$?
Influence Maximization

Q: $|S| = 3, \sigma(S) = 6$?

YES

Q: $|S| = 2, \sigma(S) = 6$?

NO
NP-Hard (Threshold Model)

Influence Maximization Problem

is at least as difficult as

Vertex Cover Problem
End of Proofs

- Influence Maximization Problem

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma(S)$ is Submodular</th>
<th>NP-hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Cascade</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Linear Threshold</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
**Initial Problem**

Given a spread model

Prove: $\sigma(S)$ is Submodular

$f(S)$: Non-negative monotone submodular

Greedy Hill Climbing

$\text{MAX}_v f(S \cup \{v\}) - f(S)$

(Maximize Marginal Gain)

**Problem**

find $S$, s.t.

$\sigma(S) \geq (1 - \frac{1}{e} - \epsilon) \cdot \sigma(S^*)$

find $S$, s.t.

$f(S) \geq (1 - \frac{1}{e}) \cdot f(S^*)$
Summary

- Problem Description
- Two Models
  - Independent Cascade Model
  - Linear Threshold Model
- Submodular Functions
- Proof of Approximation Guarantee
- Proof of NP-Hardness