Online Influence Maximization in Graphs

CPT S 553: Graph Theory Final Project

Reet Barik [1]

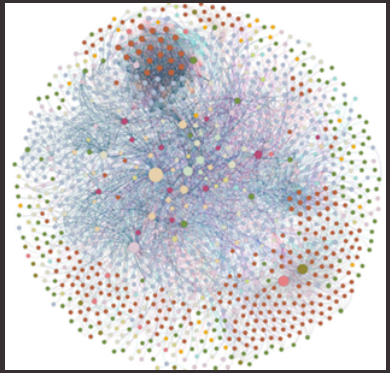
[1]School of Electrical Engineering and Computer Science, Washington State University

December 17, 2020

Outline

Motivation

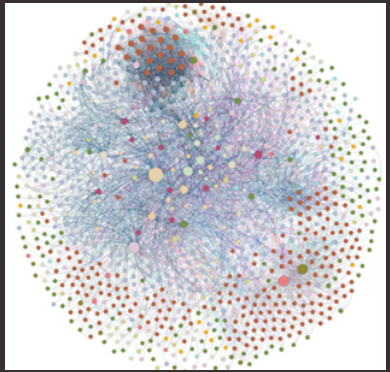
Influence Maximization (Offline)



National and Cybersecurity network (source credit: Madelyn Dunning)

- What is Influence Maximization?
- Why do we need Influence Maximization?

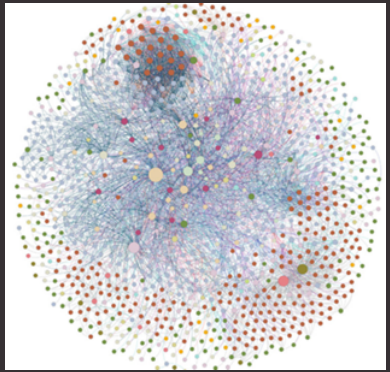
Influence Maximization (Offline)



National and Cybersecurity network (source credit: Madelyn Dunning)

- What is Influence Maximization?
- Why do we need Influence Maximization?

Influence Maximization (Offline)

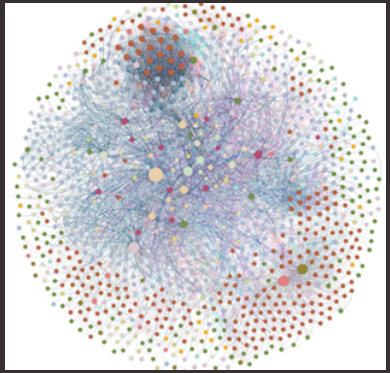


National and Cybersecurity network (source credit: Madelyn Dunning)

- What is Influence Maximization?

Given a graph $G(V, E, p)$, and a budget k, find a set of nodes S where $|S| \leq k$, such that those **seed** nodes on activation, results in the maximum number of nodes in the graph getting activated.
Kempe et. al -> NP Hard

Influence Maximization (Offline)



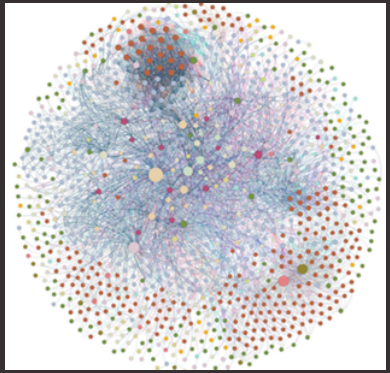
National and Cybersecurity network (source credit: Madelyn Dunning)

- Why do we need Influence Maximization?

Some use cases are:

- Marketing: Brand advertisement on social networks
- Epidemiology

Influence Maximization (Offline)



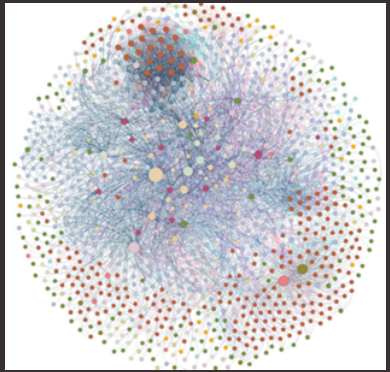
National and Cybersecurity network (source credit: Madelyn Dunning)

- Why do we need Influence Maximization?

Some use cases are:

- Marketing: Brand advertisement on social networks
- Epidemiology

Influence Maximization: Online vs Offline



National and Cybersecurity network (source credit: Madelyn Dunning)

- When the influence probabilities on the edges are unknown, the problem is known as Online Influence Maximization (OIM)

Problem formulation: Given a graph $G(V, E, p)$ with unknown probabilities $p_{u,v}$ on each edge, a budget of N trials with $1 \leq k \leq |V|$ activated nodes per trial, find a set of nodes S where $|S| \leq k$, such that those **seed** nodes on activation, results in the maximum number of nodes in the graph getting activated.

Definitions and Technical Background

Types of Diffusion Models

Linear Threshold Model

- Every node v has a threshold value θ_v which is sampled from $U[0, 1]$.
- A node v is influenced by each neighbor u according to a weight $w_{v,u}$ such that

$$\sum_{w \in neighbor(v)} w_{v,u} \leq 1$$

- If θ_v fraction of node v's neighbors are active, then v itself becomes active.

Independent Cascade Model

- Initially, a set of nodes S are *activated* or *influenced*.
- Each edge $e \in E$ has an associated influence probability p_{ij} (probability of node i influencing j).
- Every active node i has a one-shot chance of activating a neighbor j with probability p_{ij}.
- The diffusion process stops when there are no more one-shot activations possible.

Types of Feedback strategies

OIM calls for repeated interactions with the social network to somehow (use the feedback) learn the edge probabilities directly or some surrogate model representing those probability values. The feedback that can be obtained are of the following kinds:

Node-level Feedback

Every interaction or trial (a chosen seed set is activated and the diffusion process is allowed to unfold) is conducted and the final influence spread (set of nodes that got activated) is observed.

Edge-level Feedback

In each trial, every activation attempt from a node u to its neighbor v is observed and the result (whether it was a successful activation or not) is recorded.

Common Solutions and Strategies

Common Solutions

Heuristic Based

- Random
- Maximum Degree

Explore-Exploit Policy based

- Bayesian Inference
- Combinatorial Multi-Arm Bandit

Common Solutions

Heuristic Based

- Random
- Maximum Degree

Explore-Exploit Policy based

- Bayesian Inference
- Combinatorial Multi-Arm Bandit

Common Solutions

Heuristic Based

- Random
- Maximum Degree

Explore-Exploit Policy based

- Bayesian Inference
- Combinatorial Multi-Arm Bandit

Common Solutions

Heuristic Based

- Random
- Maximum Degree

Explore-Exploit Policy based

- Bayesian Inference
- Combinatorial Multi-Arm Bandit

Bayesian Inference

- Each edge activation can be represented as a boolean random variable and hence can be assumed to have a Bernoulli distribution.
- The edge probabilities are assumed to be drawn from the probability distribution function of a *Beta distribution* [Beta was the choice of distribution since it is a conjugate prior for the Bernoulli distributions, or more generally, binomial distributions].
- For an edge from node i to j, the random variable of the influence probability P_{ij} has a density function:

$$f_{P_{ij}}(x) = \frac{x^{\alpha_{ij}-1}(1-x)^{\beta_{ij}-1}}{B(\alpha_{ij}, \beta_{ij})}$$

This makes the mean $E[P_{ij}] = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}}$ and the square of standard deviation of the distribution $\sigma^2[P_{ij}] = \frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij})^2(\alpha_{ij}+\beta_{ij}+1)}$

Bayesian Inference

- With this distribution as the prior, one can use the feedback obtained (which edges got successfully activated as a consequence of activating a particular set of seed nodes) as the knowledge gained during the trials where the chosen strategy is Explore and update the posterior distribution by using Bayes' theorem.
- Various strategies can be used to find a balance between the Explore and Exploit strategies. An ϵ-greedy one is where the agent chooses to explore with a probability of ϵ and exploit with a probability of $(1 - \epsilon)$.
- After learning the edge probabilities sufficiently, an already existing offline Influence Maximization solution can be used to identify the node set S.

Combinatorial Multi-Arm Bandit

The problem setting is one where there are multiple arms on a slot machine with each arm having an unknown reward distribution. The aim is to learn those distributions by repeatedly pulling those arms while simultaneously maximizing the expected reward. In the combinatorial setting, the only difference is that multiple arms can be pulled together which might end up stochastically pulling another set of arms. To maximize the reward (or minimize regret) in such a situation while exploring enough to learn the reward distributions can be intuitively mapped to the OIM problem as follows:

CMAB	Symbol	Mapping to IM
Base arm	i	Edge (u, v)
Reward for arm i in round s	$X_{i,s}$	Status (live / dead) for edge (u, v)
Mean of distribution for arm i	μ_i	Influence probability $p_{(\mathrm{u,v})}$
Superarm	A	Union of outgoing edges E_S from nodes in seed set S
No. of times i is triggered in s rounds	$T_{i,s}$	No. of times u becomes active in s diffusions
Reward in round s	r_s	Spread $\bar{\sigma}$ in the s^{th} IM attempt

Combinatorial Multi-Arm Bandit

At each round, select a seed set S (same as pulling the corresponding super-arm E_S). One can either select S randomly (Explore) or select it based on the available probability estimates by running an offline IM solver (Exploit). The influence diffuses through the network resulting in a set of nodes getting activated. The reward is $\overline{\sigma}(S)$: the number of active nodes at the end of the diffusion process. The mean probability estimate vector $\overrightarrow{\bar{\mu}}$ is updated based on the feedback mechanism. What follows is the above framework illustrated as an algorithm:

Algorithm 1: CMAB FRAMEWORK FOR IM(Graph $G = (V, E)$, budget k, Feedback mechanism M, Algorithm $\mathcal{A}$)

Initialize $\vec{\mu}$;
$\forall i$ initialize $T_i = 0$;
for $s = 1 \rightarrow T$ **do**
    IS-EXPLOIT is a boolean set by algorithm $\mathcal{A}$;
    if *IS-EXPLOIT* **then**
        E_S = EXPLOIT(G,$\vec{\mu}$,O,k)
    else
        E_S = EXPLORE(G,k)
    Play the superarm E_S and observe the diffusion cascade c ;
    $\vec{\mu}$ = UPDATE(c,M) ;

State of the Art

Wu *et al.*: KDD'19

- Decompose the edge probability $p_e \in [0, 1]$ on edge e into two d-dimensional latent factors on the source and destination node making up that edge. i.e.,

$$p_e = \theta_{g_e}^T \beta_{r_e}$$

  where g_e and r_e denote the source (giving) node and the destination (receiving) node of edge e respectively. The underlying philosophy is that for an edge (ij), θ_i represents the influence of i and β_j represents the susceptibility of j.
- The edge probability estimations needn't be done explicitly. One can learn the two latent factors per node to implicitly estimate p_e.
- This has an additional effect of reducing the complexity of the approach from being $O(|E|)$ to that of $O(d|V|)$.

Future Scope and Directions

Scope for future research

- **Scalability**: With social networks increasing in size rapidly, there is a need for the OIM algorithms to scale accordingly. One way to do that would be to come up with parallel versions of the algorithms. But this might be tricky for those frameworks that use the feedback from one round to select the seeds in the next.
- **Dynamic Graphs**: Already available solutions to the problem treat it as a static input. But in order to simulate a more real-world scenario there is a need to come up with solutions that take dynamic graphs as input (where nodes and edges are added and/or deleted from one time step to another).

Scope for future research

- **Epidemiology**: Intuitively, spread of infectious disease and the corresponding control using vaccination can be modeled as a network. This is where IM/OIM can have an impact by coming up with vaccination strategies (identify nodes which when infected result in maximal spread of the disease and vaccinate them first). This however needs further development because the IM/OIM framework can't be used as-is, mainly due to the following reasons:
  - The typical feedback mechanism is not applicable here since it will take a long time to estimate the effect of vaccination from one round. As a result the graph might have changed by the time there is some actionable insight making it obsolete.
  - The influencing mechanism is much different in disease spread than in social media marketing. As a result, the interaction network of a population under consideration needs to tailor made for this using domain knowledge of medical and epidemiology experts.

Thank You