CS224W: Social and Information Network Analysis **Fall 2014**

Handout: Influence Maximization

The study of social processes by which ideas and innovations diffuse through social networks has been ongoing for more than half a century and as a result a fair understanding of such processes has been achieved. Modern models of social influence have been augmented with various features allowing for arbitrary network structure, non-uniform interactions, probabilistic events and other aspects. This handout will expose you to the basic stochastic model of social influence, *i.e.*, the *Independent Cascade Model (ICM)*, and show how it can be used to find an influential set of nodes to target in order to maximize the final adoption, *i.e.*, the *Influence Maximization* problem.

1 Independent Cascade Model (ICM)

The ICM was introduced by Goldenberg et.al in 2001 to model the dynamics of viral marketing and is inspired from the field of interacting particle systems. In this model, we start with an initial set S of active individuals. Each active individual u has a single chance to activate each non-active neighbour v of his/her. However, the process of activation is deemed stochastic and succeeds with probability $p_{u,v}$ independently for each attempt. Therefore, from an initial population of active individuals the activation process spreads in a cascading manner as newly activated individuals may activate new nodes that either previous attempts failed to activate or were not before accessible.

To make things more precise and to enable mathematical treatment of the model, we are going to adopt an alternative view of the model utilizing the notion of reachability.

Definition 1 (Reachability) *Given a graph $G = (V, E)$ and a node u, define X_u^E the set of reachable nodes of V from u through the edges in E (including u).*

There is an elegant interpretation of the ICM, in terms of the reachability of nodes via paths from the initial active set S. We can picture the process of a node u activating one of his neighbours v with probability $p_{u,v}$, as flipping a biased coin and if it succeeds declare the edge *live*, otherwise declare it *blocked*. Moreover, we can without loss of generality use the *principle of deferred decision* and consider that all the coins are tossed before the process begins. Therefore, from the initial graph $G(V, E)$, we get a graph $G(V, E_{live})$ where we keep only live edges. Now, in this setting all nodes that are reachable via a live path from the initial set S would become active when the cascade process quiesced. This view is very helpful and will be used to prove a crucial property about our model.

Definition 2 (ICM) *Given a graph $G = (V, E)$ and edge probabilities $\{p_e\}_{e \in E}$, consider $\{U_e\}_{e \in E}$ independent uniform $[0, 1]$ random variables. Define the random set of active edges as $I = \{e \in E : p_e \leq U_e\}$. The Independent Cascade Model for the graph G and probabilities p defines for every initial set of active nodes S, the final set A of active nodes as $A_I(S) = \cup_{u \in S} X_u^I$.*

We can think of X_u^I as the influence set of node u under random realization of edge activations I (where I is a random variable). From here on we will assume implicitly that the graph G and probabilities $\{p_e\}_{e \in E}$ are given.

2 Influence Maximization

Our end goal is to use the knowledge of the interactions to find a set of influential nodes. In order to quantify the goodness of the initial set, the stochastic nature of the ICM necessitates the use of expectations.

Definition 3 (Total Influence) *The total influence function for the ICM is* $\sigma(S) = \mathbb{E}[|A_I(S)|]$

The problem, therefore, is given a social network, *i.e.*, a set of nodes (individuals) and the edges(interactions) between them, to select the optimal "seed" of individuals to influence so that after the activation process terminates the expected number of active nodes is maximal for a seed of size k.

Definition 4 (Influence Maximization) *Given a graph* G *with probabilities* $\{p_e\}_{e\in E}$ *and an integer* k*, the Influence Maximization problem asks for the set* S *of cardinality* k *such that* $\sigma(S)$ *is maximized.*

Theorem 1 *The Influence Maximization Problem is NP-Complete.*

Proof:(Sketch) We prove the statement through a reduction of *Set Cover*. In the Set Cover we are given a "universe" of n elements U, a collection of sets $X_1, \ldots, X_m \subset U$ and an integer k. The decision problem is whether we can select k sets out of the collection such that their union equals U (that is, "covers" U). Given such an instance of Set Cover, we show that we can construct an instance of Influence Maximization such that its solution will imply a solution to the original problem. That means we need to provide a directed graph $G = (V, E)$ and probabilities $\{p_e\}_{e\in E}$. The vertex set V consists of U along with a separate vertex v_i for each set X_i. The edge set includes only the directed edges pointing from v_i to the elements of V corresponding to the elements in X_i. We set all the probabilities of the edges equal to 1. Since, vertices corresponding to elements of U do not influence other vertices, and any vertex v_i would immediately activate the vertices corresponding to X_i, solving the Influence Maximization problem with cardinality k would also tell us whether the universe U can be covered by k sets out of $X_1, \ldots, X_m$. On the other hand the decision version of Influence Maximization obviously belongs to NP as it is possible (but non-trivial) to compute the total influence function for the optimal solution. ■

3 Submodularity and ICM

A crucial property satisfied by the ICM, that will sidestep the hardness result and enable the algorithmic treatment of Influence Maximization, is that of submodularity.

Definition 5 (Submodularity) *A set function* $f : 2^V \to \mathbb{R}$ *is called* submodular *if for all subsets* $S \subseteq T \subseteq V$ *and* $u \in V$ *the following inequality holds:*

$$f(S \cap \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T) \tag{1}$$

Intuitively, submodularity is the set-function analog of concavity. Specifically, a function is called submodular if it satisfies the "diminishing returns" property: the marginal gain by adding an element to a set S it is at least as the marginal gain by adding an element to the superset T. In other words, the higher the ground value is, the smaller is the marginal gain of adding one element. The following property of submodular function is useful in proving that the total influence function is submodular.

Lemma 1 (Conic combinations) *Let $c_1, \ldots, c_n \geq 0$ be non-negative numbers and $f_1, \ldots, f_n : 2^V \to \mathbb{R}$ be submodular functions, then $\tilde{f} = \sum_{i=1}^{n} c_i f_i$ is a submodular function.*

Proof: Let $S \subseteq T$ be subsets of V, then for every $u \in V$ we have:

$$\begin{aligned} \tilde{f}(S \cup \{u\}) - \tilde{f}(S) &= \sum_{i=1}^{n} c_i \left[f_i(S \cup \{u\}) - f_i(S) \right] \\ &\geq \sum_{i=1}^{n} c_i \left[f_i(T \cup \{u\}) - f_i(T) \right] \\ &= \tilde{f}(T \cup \{u\}) - \tilde{f}(T) \end{aligned}$$

where in the middle inequality we used submodularity of the functions f_i and positivity of the coefficients c_i. ■

Theorem 2 *The total influence function $\sigma(S)$ is monotone and submodular.*

Proof: We start by writing out the expression for the total influence. We have:

$$\sigma(S) = \mathbb{E}[|A_I(S)|] = \sum_{i \subseteq E} \mathbb{P}(I = i) \cdot |A_i(S)| \tag{2}$$

where $\mathbb{P}(I = i)$ is the probability according to the ICM that the set of active edges I is $i \subseteq E$. Since probabilities are non-negative, if we could show that $f_i(S) = |A_i(S)|$ is a submodular function, invoking Lemma 1 would complete the proof. Let $S \subseteq T \subseteq V$ and $u \in V$, then:

$$\begin{aligned} f_i(S \cup \{u\}) - f_i(S) &= \left| A_i(S) \cup X_u^i \right| - |A_i(S)| \\ &= |X_u^i| - \left| A_i(S) \cap X_u^i \right| && (3) \\ &\geq |X_u^i| - \left| A_i(T) \cap X_u^i \right| && (4) \\ &= \left| A_i(T) \cup X_u^i \right| - |A_i(T)| && (5) \\ &= f_i(T \cup \{u\}) - f_i(T) \end{aligned}$$

where in (4) we used monotonicity of $A_i(S)$ and in (3) and (5) the fundamental property $|A \cup B| = |A| + |B| - |A \cap B|$. Thus we proved the defining inequality of submodularity for $f_i(S)$. ■

4 Hill Climbing Algorithm

Submodularity of the total influence function is a property that can be exploited algorithmically to obtain a good approximation to the Influence Maximization Problem. In particular, there is a hope

Algorithm 1 Hill Climbing Algorithm

Input: a graph $G = (V, E)$, probabilities $\{p_e\}_{e \in E}$ and an integer k.
Output: Initialize $S_0 = \emptyset$
1: **for** $i = 1$ to k **do**
2: $s_i = \arg\max_{u \in V \setminus S_{i-1}} [\sigma(S_{i-1} \cup \{u\}) - \sigma(S_{i-1})]$
3: $S_i = S_{i-1} \cup s_i$
4: **end for**
Output: the set S_k.

that locally optimal choices would result in good final spread. The following natural algorithm is particularly tailored for problems where submodularity and monotonicity of the objective function coincide.

Lemma 2 (Telescoping) *For a submodular f, any set A and $B = \{b_1, \ldots, b_k\}$ it holds that*

$$f(A \cup B) - f(A) \leq \sum_{i=1}^{k} [f(A \cup \{b_i\}) - f(A)] \tag{6}$$

Proof: Let $B_i = \{b_1, \ldots, b_i\}$ with $B_0 = \emptyset$ and $B_k = B$. We start by expressing the left hand side as a telescopic sum using the above sequence of sets:

$$\begin{aligned} f(A \cup B_k) - f(A \cup B_0) &= f(A \cup B_k) - f(A \cup B_{k-1}) + \ldots + f(A \cup B_1) - f(A \cup B_0) \\ &= \sum_{i=1}^{k} [f(A \cup B_i) - f(A \cup B_{i-1})] \\ &= \sum_{i=1}^{k} [(f(A \cup B_{i-1}) \cup \{b_i\}) - f(A \cup B_{i-1})] \qquad (7) \\ &\leq \sum_{i=1}^{k} [f(A \cup \{b_i\}) - f(A)] \qquad (8) \end{aligned}$$

where we used the fact that $B_i = B_{i-1} \cup \{b_i\}$ in (7) and submodularity of f in (8). ■

Definition 6 (Marginal Increments) *Given the set S_{i-1}, the marginal increment at step i is defined as $\delta_i = f(S_i) - f(S_{i-1}) = \max f(S_{i-1} \cup \{u\}) - f(S_{i-1})$.*

Lemma 3 (Accretion) *Let S_i be the set after i-steps of the HC algorithm and T be any other set of size k. Then:*

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

Proof: Since, the two sets S_i and T can in principle be arbitrarily different we are going to use our telescoping lemma to

$$\begin{aligned} f(T) - f(S_i) &\leq f(S_i \cup T) - f(S_i) && (9)\\ &\leq \sum_{j=1}^{k} [f(S_i \cup \{t_j\}) - f(S_i)] && (10)\\ &\leq \sum_{j=1}^{k} \delta_{i+1} && (11)\\ &= k \cdot \delta_{i+1} && (12) \end{aligned}$$

where in (9) we used monotonicity and in (10) we the greedy property of the HC algorithm. Now recalling that $\delta_{i+1} = f(S_{i+1}) - f(S_i)$ and substituting the last inequality for δ_{i+1} gives the required statement. ■

Theorem 3 *The hill-climbing algorithm finds a set $\tilde{S}$ such that $\sigma(\tilde{S}) \geq (1 - \frac{1}{e})\sigma(S^*)$.*

Proof: To prove our theorem we are going to prove a stronger result, namely

$$f(S_i) \geq \left[1 - \left(1 - \frac{1}{k}\right)^i\right] f(T) \quad (13)$$

To that end we are going to employ induction on i. For $i = 0$, (13) trivially holds as $f(\emptyset) \geq 0$. Next, we carry out the inductive step:

$$\begin{aligned} f(S_{i+1}) &\geq \left(1 - \frac{1}{k}\right) f(S_i) + \frac{1}{k} f(T) && (14)\\ &\geq \left(1 - \frac{1}{k}\right) \left[1 - \left(1 - \frac{1}{k}\right)^i\right] f(T) + \frac{1}{k} f(T) && (15)\\ &= \left[1 - \left(1 - \frac{1}{k}\right)^{i+1}\right] f(T) && (16) \end{aligned}$$

where in (14) we used Lemma 3 and in (15) the inductive hypothesis. Using (13) for $i = k$ and $T = S^*$ (the optimal set of cardinality k) we get: $f(S_k) \geq \left[1 - \left(1 - \frac{1}{k}\right)^k\right] f(S^*)$. Since, the right hand side is decreasing in k, we have that always $f(S_k) \geq \lim_{k\to\infty} \left[1 - \left(1 - \frac{1}{k}\right)^k\right] \cdot f(S^*) = \left(1 - \frac{1}{e}\right) f(S^*)$ as $\lim_{x\to\infty}(1 - 1/x)^x = 1/e$. ■