

## 1 A non-monotone infection model

In the previous lecture, we began an investigation of models for the spread of infections on graphs. We assumed the following notion of *monotonicity*: once a node becomes infected, it stays infected. For certain types of diseases or behaviors, this model may be very accurate. For others, nodes will reevaluate their choice in every time step, based on the choices of their neighbors. A natural modification of the previous model posits that if in the previous time step, at least a  $p$  fraction of a node's neighbors are infected, then the node will be infected in the next step, and otherwise, it will not.

In thinking about this new definition, a first question is whether any infinite graphs can be infected starting from a finite set of infected nodes. It is pretty easy to see that the answer is "Yes": if we start with two adjacent infected nodes on the infinite line, and  $p < \frac{1}{2}$ , then the entire line will eventually become infected. In fact, for infinite graphs, Morris [3] showed the following lemma:

**Lemma 1** *If the (infinite) graph  $G$  has a (finite) contagious set with threshold  $p$  in the monotone model, then it also has a finite contagious set with the same threshold in the non-monotone case. The starting set in the non-monotone case will be the union of the starting set of the monotone case and its neighborhood.*

Notice that our example of an infinite graph with finite infectious set crucially used that  $p < \frac{1}{2}$ . It seems much more difficult to come up with examples having  $p \geq \frac{1}{2}$ . In fact, it is not even obvious how to construct arbitrarily large finite graphs that can be infected by a small set of nodes.

Notice that for the monotone case, the question with finite graphs is easy: take a star graph with threshold  $1/2$ , and start with just the center node infected. In the non-monotone case, this graph will end up oscillating with period 2, between the center node infected, and all leaf nodes infected. In fact, this is far from accidental, for one can show:

**Fact 2** *In the non-monotone case for a finite graph, the infection either converges, or oscillates with a period of 2.*

Returning to the question of the existence of small infectious sets for arbitrarily large graphs, Berger [1] answered it in the affirmative for  $p = \frac{1}{2}$ .

**Theorem 3 (Berger [1])** *There are arbitrarily large graphs that can be infected (in the non-monotone case) with  $p = 1/2$  and a starting set  $S$  with  $|S| \leq 18$ .*

At this point, it is open whether such finite infectious sets exist for  $p > \frac{1}{2}$ . However, we have the following partial result, showing that arbitrarily large  $p$  will not work.

**Lemma 4** *It is impossible to infect arbitrarily large graphs if  $p > 3/4$ .*

**Proof.** We denote by  $A_t$  the set of nodes active in round  $t$ , and write  $I_t := A_t \cap A_{t-1}$  for the set of nodes active both in rounds  $t$  and  $t-1$ . Then, let  $S_t := \bigcup_{t' \leq t} I_{t'}$  be the set of nodes which were active in any two consecutive rounds up to and including round  $t$ . Finally, let  $\delta(S_t)$  be the set of edges leaving the  $S_t$ , and  $\sigma_t := |S_t| + |\delta(S_t)|$ .

We will show that if we start with  $c$  nodes active, then  $\sigma_t = O(c^2)$  for all  $t$ . For this purpose, we show that  $\sigma_1 = O(c^2)$ , and that  $\sigma$  is a non-increasing function of  $t$ .

For the case  $t = 1$ , notice that  $S_1 = A_0 \cap A_1 \subset A_0$ . Thus,  $|S_1| \leq c$ . Since each  $v \in S_1$  is active at time  $t = 1$ , its degree can be at most  $\frac{4}{3}c$  (for at least  $\frac{3}{4}$  of its neighbors must be in  $A_0$ ). Thus, the sum of all degrees in  $S_1$  is at most  $c \cdot \frac{4}{3}c$ , which is  $O(c^2)$ .

To show that  $\sigma_{t+1} \leq \sigma_t$ , consider any node  $v \in S_{t+1} \setminus S_t$ . Due to the addition of  $v$ , the  $|S_{t+1}|$  term of  $\sigma_{t+1}$  increases by one. On the other hand, because  $v \in A_{t+1} \cap A_t$ , more than  $\frac{3}{4}$  of  $v$ 's neighbors were active at time  $t$ , and also at time  $t - 1$ . Thus, more than half of  $v$ 's neighbors are active *both* at times  $t$  and  $t - 1$ , i.e., they are in  $A_t \cap A_{t-1} \subseteq S_t$ . For all those neighbors, we lost one crossing edge from  $\delta(S_t)$  (as compared to  $\delta(S_{t-1})$ ), and gained at most one for all other neighbors of  $v$ , of which there are strictly fewer. Thus,  $|\delta(S_t)|$  decreases by at least 1 for each node  $v \in S_{t+1} \setminus S_t$ . Therefore,  $\sigma$  cannot increase overall. ■

## 2 Causing a large spread of an infection

So far, the results we studied were existential. We showed that there are large graphs that can be infected (in various models) by a small set of initial nodes. From a practical point of view, a much more relevant question is how to reach as many nodes as possible in a *given graph*. Naturally, this question has applications in using network effects for the purpose of marketing. More formally, we can look at the following question: Given a number  $k$ , which set  $S$  with  $|S| \leq k$  will eventually infect as many nodes of  $G$  as possible (in the monotone case). Unluckily, this problem not only is NP-hard, but also hard to approximate for the Morris contagion model [3].

**Lemma 5** *Unless  $P = NP$ , the above problem cannot be approximated to within  $O(n^{1-\epsilon})$  for any  $\epsilon > 0$ .*

**Proof.** We prove this with a reduction from SET COVER: Given sets  $S_1, \dots, S_m$ , each a subset of  $\{1, \dots, n\}$ , and a number  $k \in \mathbb{N}$ , are there  $k$  sets  $S_i$  whose union is the entire set?

To reduce a SET COVER instance with  $m$  subsets of a set with  $n$  elements, we create a directed graph as follows: Let  $\{s_1, s_2, \dots, s_m\}$  be nodes corresponding to the  $m$  subsets and  $\{u_1, u_2, \dots, u_n\}$  be nodes corresponding to the  $n$  elements; our construction will ensure that  $u_i$  becomes active when at least one of the nodes corresponding to sets containing  $u_i$  is active. We achieve this by connecting each  $s_k$  to the  $u_i$ 's in it, and setting a threshold of  $1/m$  for each  $u_i$ . Next, for a large constant  $c$ , we add  $N = n^c$  more nodes  $\{x_1, x_2, \dots, x_N\}$ . Each  $x_j$  is connected to all of the nodes  $u_i$ , and it becomes active only when *all* of the  $u_i$  are (i.e., the  $x_j$  have a threshold of 1). (The construction is depicted graphically in Figure 1.)

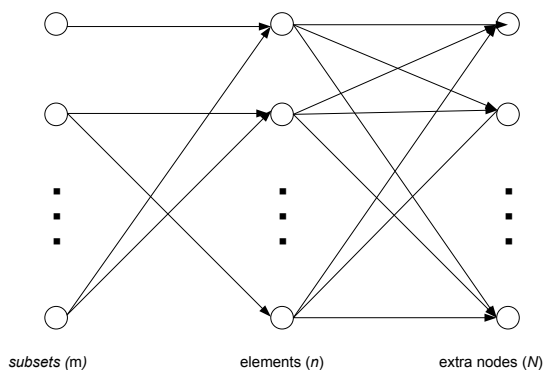


Figure 1: Construction for proving hardness of approximation

If there are at most  $k$  sets that cover all elements, then activating the nodes corresponding to these  $k$  sets will activate all of the nodes  $u_i$ , and thus also all of the  $x_j$ . In total, at least  $N + n + k$  nodes will be active. Conversely, if there is no set cover of size  $k$ , then no targeted set will activate all of the  $u_i$ , and hence none of the  $x_j$  will become active (unless targeted). In particular, fewer than  $n + k$  nodes are active in the

end. If an algorithm could approximate the problem within  $n^{1-\epsilon}$  for any  $\epsilon$ , it could distinguish between the cases where  $N + n + k$  nodes are active in the end, and where fewer than  $n + k$  are. But this would solve the underlying instance of *Set Cover*, and therefore is impossible assuming  $P \neq NP$ . ■

Notice that this proof can be modified to deal with uniform thresholds, by adding more nodes.

In the Morris model, the problem thus turns out to be completely intractable. Yet, we may be interested in finding models that are more amenable to approximation. One small modification that turns out to make a significant difference is to assume that the thresholds are uniformly (and independently) random instead of deterministic. Thus, we obtain a model where each edge has a weights  $w_e \geq 0$  such that  $\sum_{e \text{ into } v} w_e \leq 1$ . Each node  $v$  independently chooses a threshold  $\theta_v \in [0, 1]$  uniformly at random. The goal is to choose a set  $S$  with  $|S| \leq k$  reaching as large a set as possible in expectation (over the random choices of  $\theta_v$ ). We define  $f(S)$  to be objective function, i.e., the expected size of the finally infected set if we start with set  $S$  infected. We will spend the next few lectures proving the following theorem:

**Theorem 6** *There is a  $1 - \frac{1}{e}$  approximation algorithm for the problem of selecting the set  $S$  maximizing  $f(S)$ .*

The algorithm turns out to be a very simple greedy algorithm:

---

**Algorithm 1** The simple greedy algorithm

---

- 1: Start with  $S = \emptyset$ .
  - 2: **for**  $k$  iterations **do**
  - 3:   Add to  $S$  the node  $v$  maximizing  $f(S + v) - f(S)$ .
  - 4: **end for**
- 

The proof of the performance guarantee consists of three parts, captured by the following lemmas.

**Lemma 7** *A node  $v$  (approximately) maximizing  $f(S + v) - f(S)$  can be found in polynomial time.*

**Lemma 8**  *$f$  is a monotone and submodular function of  $S$ .*

Recall that a function on sets is *monotone* if  $f(S') \geq f(S)$  whenever  $S' \supseteq S$ , and *submodular* (having diminishing returns) if  $f(S' \cup \{x\}) - f(S') \leq f(S \cup \{x\}) - f(S)$  whenever  $S' \supseteq S$ . Equivalently, submodularity is characterized by the condition that  $f(S) + f(T) \geq f(S \cap T) + f(S \cup T)$  for all sets  $S, T$ . Also notice that the monotonicity of  $f$  is obvious.

**Lemma 9 (Nemhauser/Wolsey/Fischer [2, 4])** *If  $f$  is a monotone and submodular function, then the greedy algorithm is a  $(1 - \frac{1}{e})$ -approximation for the problem of maximizing  $f(S)$  subject to the constraint that  $|S| = k$ .*

We begin by proving the last lemma, which is naturally useful for other problems as well (such as SETCOVER).

**Proof of Lemma 9.** Let  $v_1, v_2, \dots, v_k$  be the nodes selected by the greedy algorithm (in the order in which they were selected), and denote  $S_i = \{v_1, v_2, \dots, v_i\}$ . Then, the marginal benefit derived from the addition of element  $v_i$  is  $\delta_i = f(S_i) - f(S_{i-1})$ . Let  $T$  be the optimal solution, and  $W_i = T \cup S_i$ .

First, the monotonicity of  $f$  implies that  $f(T) \leq f(W_i)$  for all  $i$ . Because the algorithm chooses to add the best available node in the  $(i + 1)^{\text{st}}$  iteration, and the benefit of any elements added later cannot be greater by submodularity, the total objective value for the set  $W_i$  is at most  $f(W_i) \leq f(S_i) + k\delta_{i+1}$ , and thus also  $f(T) \leq f(S_i) + k\delta_{i+1}$ .

Solving this for  $\delta_{i+1}$ , and using that  $f(S_{i+1}) = f(S_i) + \delta_{i+1}$  now shows that  $f(S_{i+1}) \geq f(S_i) + \frac{1}{k} \cdot (f(T) - f(S_i))$ . We will prove by induction that  $f(S_i) \geq (1 - (1 - \frac{1}{k})^i) \cdot f(T)$ . The base case  $i = 0$  is trivial.

For the inductive step from  $i$  to  $i + 1$ , we use the above inequality to write

$$\begin{aligned} f(S_{i+1}) &\geq f(S_i) + \frac{1}{k} \cdot (f(T) - f(S_i)) \\ &= \left(1 - \frac{1}{k}\right)f(S_i) + \frac{1}{k} \cdot f(T) \\ &\stackrel{\text{IH}}{\geq} \left(1 - \frac{1}{k}\right)\left(1 - \left(1 - \frac{1}{k}\right)^i\right) \cdot f(T) + \frac{1}{k} \cdot f(T) \\ &= \left(1 - \frac{1}{k} - \left(1 - \frac{1}{k}\right)^{i+1} + \frac{1}{k}\right) \cdot f(T) \\ &= \left(1 - \left(1 - \frac{1}{k}\right)^{i+1}\right) \cdot f(T), \end{aligned}$$

completing the inductive proof. Using the fact that  $\left(1 - \frac{1}{k}\right)^i \geq 1/e$  for all  $i \leq k$ , we obtain that the algorithm is a  $(1 - 1/e)$ -approximation.  $\blacksquare$

## References

- [1] E. Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory Series B*, 83:191–200, 2001.
- [2] G. Cornuejols, M. Fisher, and G. Nemhauser. Location of bank accounts to optimize float. *Management Science*, 23:789–810, 1977.
- [3] S. Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.
- [4] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.