

Notes on Social and Behavioral Networks

Cristian Di Pietrantonio

October 25, 2017

Contents

1	Meme flow	2
1.1	The Trace Spreading Model	2
1.2	The First-Edge Algorithm	3
2	Chernoff Bound	6
2.1	First-Edge Algorithm, revised	6
2.2	Chernoff bound	7
2.3	Chain letters	8
3	Submodular functions	11
3.1	k-max cover	11
3.2	Submodular functions	13
3.3	Facility location	16
3.4	Approximation of submodular functions	17
3.5	Learning a submodular function	20
4	Random Graph Models	22
4.1	Preferential Attachment Model	22
4.2	Small World Phenomenon	29

Chapter 1

Meme flow

On the web, information is published by one or more *sources* and often spreads quickly to other parties. The most popular kind of information that spreads quickly is a meme. From Wikipedia:

A meme is an idea, behavior, or style that spreads from person to person within a culture - often with the aim of conveying a particular phenomenon, theme or meaning represented by the meme.

In this chapter we are going to define a mathematical model that describes the spread of information. We call *node* an entity on the web that holds the information of interest; it can be a blog, a website or a social network profile. Nodes can receive or take information from other nodes, enabling the spread of the meme. Nodes can also publish the same meme independently (without “copying” each other). Whenever a node publishes information, a timestamp of the event is also available. At the end, considering a some particular meme, all we can see is a bunch of nodes having published with an associated timestamp. What we would like to know is the *social graph* that links these nodes, i.e., who retrieves information from who.

1.1 The Trace Spreading Model

Authors in [3] introduce the Trace Spreading Model, where a *trace* is a sequence of nodes ordered by their timestamp.

Define an undirected graph $G = (V, E)$, where:

- V is the set of nodes that holds the information of interest;
- There is only a source of information, chosen uniformly at random;
- $e = \{v, v'\} \in E$ iff information propagated from v to v' , $v, v' \in V$ (the actual direction is not important, as you will notice later).

- Each edge $e = \{v, v'\}$ has a *weight* $w(e)$ sampled i.i.d. in $Exp(\lambda)$, which represent the time needed for the information to propagate from v to v' . In fact, for our purpose, any distribution will do (i.i.d. is the important part).

In this model, the trace follows the shortest path tree from the source.

So, having this model and a set of traces (and every trace contains all the nodes), is it possible to reconstruct the unknown social graph G ?

1.2 The First-Edge Algorithm

Authors in [1] shows that we can reconstruct (in most cases) the graph $G = (V, E)$ with high probability with a simple algorithm.

Input: T , the set of traces
 $E \leftarrow \{\}$ //the set of edges in the graph ;
for $t \in T$ **do**
 $u \leftarrow t[0]$ //the first node in t;
 $v \leftarrow t[1]$ //the second node in t;
 $E \leftarrow E \cup \{(u, v)\};$
end

Algorithm 1: The first-edge algorithm.

Consider Algorithm 1. What it does is simply adding an edge between the first and second node in a trace, for every trace. That is because we can only be sure about the existence of that edge. The first node in a trace is the source (for that trace), since it has the lowest timestamp; the second node, which has the second lowest timestamp, could only get the information from the first node.

Theorem 1. *If there are $|T| \geq cn\Delta \log n$ traces, the first-edge algorithm can reconstruct the unknown graph $G = (V, E)$ with probability $p \geq 1 - \frac{1}{n^{2(c-1)}}$, where $n = |V|$, $\Delta = \max_v \deg(v)$ and c is a constant.*

Proof. An edge $\{u, v\}$ will be in the graph if and only if u and v appears at the beginning of at least one trace. We will prove that it happens with high probability.

Consider the following event.

$\xi_1 =$ “A given edge $\{u, v\} \in E$ is inferred using the trace t .”

The probability of ξ_1 is given by

$$Pr\{\xi_1\} = Pr\{t[0] = x\}Pr\{t[1] = y, y \in \{u, v\} \setminus \{x\} \mid t[0] = x\}, x \in \{u, v\} \quad (1.1)$$

$$= \frac{2}{n} \frac{1}{\deg(x)}, x \in \{u, v\} \quad (1.2)$$

$$\geq \frac{2}{n\Delta}. \quad (1.3)$$

In other words it is the probability of one of the two nodes u, v to appear as first node in the trace multiplied by the probability that the other appears as second node in the trace. Since the source is selected uniformly at random, every node has the same probability $\frac{1}{n}$ to be picked; for us, either u or v will do. Next, assume without loss of generality (wlog) that u is picked as first node. The second node will now be drawn from the neighbors of u (convince yourself that there can't be a node y after the first node in the trace without it being one of its neighbors). The probability that the neighbor picked is u is $\frac{1}{\deg(u)}$ that can be overestimated using Δ .

Now, the complementary event of ξ_1 is “the edge $\{u, v\}$ is not inferred using trace t ” and its probability is $(1 - \frac{2}{n\Delta})$. For the edge not to be inferred at all, there should be no trace from which that edge can be extracted.

$$Pr\{\text{An edge in the graph is not inferred}\} = \left(1 - \frac{2}{n\Delta}\right)^{|T|} \leq \left(1 - \frac{2}{n\Delta}\right)^{cn\Delta \log n} \quad (1.4)$$

We will show that this probability is tiny. To do so, we must use the following lemma.

Lemma 1.

$$1 - x \leq e^{-x} \quad (1.5)$$

Proof. Studying the function we see that

$$e^{-x} = (-e^{-x})' = (e^{-x})'' > 0$$

so its tangent in every point is below the function. Taking the tangent in $x_0 = 0$ we have

$$\begin{aligned} e^{-x} &\geq (e^{-x_0})'x + e^{-x_0} \\ &= -e^{-x_0} + 1 \\ &= -x + 1 \\ &= 1 - x. \end{aligned}$$

□

Going back to our proof we have

$$Pr\{\text{Edge } \{u, v\} \text{ in the graph is not inferred}\} = \left(1 - \frac{2}{n\Delta}\right)^{|T|} \quad (1.6)$$

$$\leq \left(1 - \frac{2}{n\Delta}\right)^{cn\Delta \log n} \quad (1.7)$$

$$\leq \left(e^{-\frac{2}{n\Delta}}\right)^{cn\Delta \log n} \quad (1.8)$$

$$= e^{\log n^{-2c}} \quad (1.9)$$

$$= \frac{1}{n^{2c}}. \quad (1.10)$$

Finally, we want to compute the probability that the algorithm fails. Let

$$\xi_{a,b} = \text{“Edge } \{a, b\} \text{ is not inferred by the algorithm”}.$$

Then

$$Pr\{\text{The algorithm fails}\} = Pr\left\{\bigcup_{\{a,b\} \in E} \xi_{a,b}\right\} \quad (1.11)$$

$$\leq \sum_{\{a,b\} \in E} Pr\{\xi_{a,b}\} \quad (\text{by union bound})$$

$$= |E| Pr\{\xi_{a,b}\} \quad (1.12)$$

$$\leq n^2 \frac{1}{n^{2c}} \quad (1.13)$$

$$= \frac{1}{n^{2(c-1)}}. \quad (1.14)$$

Notice that we can make this probability tiny as we want by increasing c , i.e. by increasing the number of traces. \square

Chapter 2

Chernoff Bound

In this Chapter we see a way to bound tail distributions, that is, the probability that a random variable assumes values that are far from its expectation.

2.1 First-Edge Algorithm, revised

In the last chapter, we studied the First-Edge Algorithm for reconstructing the social graph given a set of traces. The way we proved the algorithm to be valid can be abstracted in the following way. For each trace i , and a given edge $\{u, v\} \in E(G)$, define

$$X_i = \begin{cases} 1 & \{u, v\} \text{ are not the first two nodes in trace } i \\ 0 & \text{otherwise.} \end{cases} \quad (2.1)$$

We considered only a given edge and we computed the probability that the nodes involved in it were never at the beginning of a trace. The reason we wanted to do that was that if $\{u, v\}$ were the first two nodes in a trace then the edge was indeed in the graph.

What we were really asking for is the following:

$$Pr \left\{ \prod_i X_i = 1 \right\}. \quad (2.2)$$

If all X_i s are one then no trace had $\{u, v\}$ as first nodes. Otherwise, the product is zero and the edge is in the graph. Essentially we studied the product of random variables.

When there are a bunch of independent variables and one wants to compute, say, when one of them is different from the others, then the analysis is usually pretty simple. In most cases, though, is not so easy. Usually what is studied is the random variable X defined as

$$X = \sum_i X_i, \quad (2.3)$$

that is, the sum of the X_i s as opposite to the product. In most cases, people study this random variable, defining it from a set of variables, which may be independent, and with the aim to compute statistics like the expectation, the second moment, or the probability that the X deviate from its expectation by more than a certain amount. If X is always (or with high probability) close to its expectation we say the variable is *concentrated*. This specific random variable is ubiquitous in algorithms and random processes. We will start looking at this type of variable and its use in modeling social network processes.

2.2 Chernoff bound

We state (without proof) the Chernoff Bound, which is a powerful tool to bound a tail distribution with exponentially decreasing values.

Theorem 2 (Chernoff Bound). *Let X_1, X_2, \dots, X_n , $0 \leq X_i \leq 1, \forall i$, be a set of mutually independent random variables, and $X = \sum_{i=1}^n X_i$. Then*

$$\Pr \{X \leq (1 - \varepsilon)E[X]\} \leq e^{-\frac{\varepsilon^2 E[X]}{3}} \quad (2.4)$$

and

$$\Pr \{X \geq (1 + \varepsilon)E[X]\} \leq e^{-\frac{\varepsilon^2 E[X]}{3}}. \quad (2.5)$$

Consider the following situation. We flip n fair coins. Define

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th flip results in heads} \\ 0 & \text{otherwise} \end{cases}$$

then

$$X = \sum_{i=1}^n X_i = \# \text{ of heads.}$$

What is the probability of the event $\{X = k\}$? For X to be k there must be exactly k coins which resulted in heads. We can choose among $\binom{n}{k}$ compatible events (set of variables), each having the same probability 2^{-n} to happen. It follows that

$$\Pr \{X = k\} = \binom{n}{k} 2^{-n}.$$

What if the coin is not fair? We have the binomial distribution.

$$\Pr \{X = k\} = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (2.6)$$

We would like to know what is the probability that X assumes a value far from the expected one. We can use the Chernoff Bound.

$$\Pr \{X \leq (1 - \varepsilon)np\} \leq e^{-\frac{\varepsilon^2 np}{3}}, \quad (2.7)$$

The quantity is going down exponentially with n . Why is this useful? ϵ can go down to $\frac{1}{\sqrt{(n)}}$ and we would still get something that is going down to 0 very quickly. Suppose we choose $p = \frac{1}{2}$ and

$$\epsilon = \sqrt{\frac{\log n}{n}} \quad (2.8)$$

Then we have

$$e^{-\frac{\log n \cdot \frac{1}{2} n}{3}} = e^{-\frac{1}{6} \log n} = \frac{1}{\sqrt[6]{n}}.$$

That is, the probability of error can be made as tiny as we want by increasing n .

The Chernoff bound is an extremely important tool that is used, as an example, in all applications that deal with large datasets. This is because it states that it is not necessary to look at all the values to get a sense about what the mean is.

Consider the following scenario. There are a bunch of people that has to vote of one candidate or the other. A first attempt to predict the winner could be to ask to every person for his/her preference. Then we one would have all the necessary information, but it cannot be done in practice. A more realistic way of doing it is, because of the existence of the Chernoff bound, to pick random people. If there is some gap between the two candidates, maybe 51 vs. 49 or more, then after maybe few cases (maybe 2000 people in this example) one would be able to tell the candidate that is more likely to win. The error may be proportional to \sqrt{n} but, as long as the gap is linear in n , the prediction can be considered valid.

2.3 Chain letters

Chain letters are (were) online petitions that spread virally through emails. They work in the following way. The petition's creator writes an email explaining a certain problem he or she wants to address. Then he/she signs the petition, sends it to a set of recipients and ask them to sign and forward the email. Some recipients may ignore the email, others may sign and forward it instead. One can model the spread using a tree T , where the root is the petition's creator and every node has a child for each person he forwarded the email to. Of course, email are private. For an external observer to know the petition, at least one person must post the email on a public accessible place on the Web (e.g. public mailing lists). When a node v (person) *exposes* itself (i.e. publish the email), all its ancestors in the tree are revealed too (the path from the root to v).

We would like to know some information about the spread process, such as the reach of the petition. The full tree T is in general unknown, and what can be observed of it depends on how many and which nodes get exposed. We can assume that there exists a small probability $\delta > 0$ such that every node $v \in T$ gets exposed independently with that probability. Define T_δ to be the visible part of T reconstructed though the exposed nodes. Let V be the set of nodes in T_δ . Define $E \subseteq V$ to be the set of nodes that exposed themselves by disclosing their email to the public. Finally, let $L \subseteq E$ be the set of leaves in T_δ .

As we said, we would like to know $|V| = n$, but we only have E . Can we estimate n ? If we assume that each node in E exposed itself independently from the others, we can make a reasonable guess about n by looking at T_δ [2]. In fact

$$n\delta = |E| \quad (2.9)$$

so

$$n = \frac{|E|}{\delta}. \quad (2.10)$$

At this point, we need to find δ' such that $\delta' \approx \delta$. One can wrongly think that

$$\delta' = \frac{|E|}{|T_\delta|}. \quad (2.11)$$

The problem here are the leaves of T_δ . If they hadn't exposed themselves, we wouldn't know about them (or other nodes). From the point of view of T_δ , they are *not* exposed independently with probability δ' ; rather, they are exposed with probability 1. To compute δ' then, we exclude T_δ 's leaves. What follows is an algorithm that outputs δ' .

```

Input:  $E, L, V$ .
if  $|V| = 0$  then
  | return 0
end
if  $|V| = 1$  then
  | return 1
end
return  $\frac{|E|-|L|}{|V|-|L|}$ 

```

Algorithm 2: Estimation of δ

Finally, if $|V| \geq 2$,

$$n \approx \frac{|E|}{\delta'}. \quad (2.12)$$

Lemma 2. *If $|V| \geq 2$ then*

$$\Pr \{ |\delta' - \delta| \geq \epsilon\delta \} \leq 2e^{\frac{1}{3}\epsilon^2\delta|V-L|}. \quad (2.13)$$

Proof. For each node i , let X_i be a random variable such that

$$X_i = \begin{cases} 1 & \text{if node } i \text{ exposes itself} \\ 0 & \text{otherwise.} \end{cases} \quad (2.14)$$

Consider the three sets A , B and C built in the following way. We can scan the original tree in a bottom up fashion. At a certain time, node i is considered. If $i \notin A$, check if the node was exposed. If $X_i = 1$, add i to B and every its ancestor to A ;

otherwise, add i to C . Observe that the tripartition process does not disclose the value of X_i for all $i \in A$, and that $A = V - L$ and thus the set $E - L$ coincides with the set of exposed nodes in A . Then, we can apply the Chernoff bound, having $X = |E - L|$.

$$\Pr\{X \geq (1 + \epsilon)\delta|A|\} = \Pr\{|E - L| \geq |A|\delta + \epsilon\delta|A|\} \quad (2.15)$$

$$= \Pr\left\{\frac{|E - L|}{|A|} \geq \delta + \epsilon\delta\right\} \quad (2.16)$$

$$= \Pr\{|\delta' - \delta| \geq \epsilon\delta\} \quad (\text{We want the difference small enough})$$

$$= \Pr\{|X - |A|\delta| \geq \epsilon\delta|A|\} \quad (2.17)$$

$$\leq 2e^{-\frac{1}{3}\epsilon^2 E[X]} \quad (2.18)$$

$$= 2e^{-\frac{1}{3}\epsilon^2 \delta|V-L|}. \quad (2.19)$$

□

Chapter 3

Submodular functions

In this chapter we first look at the k -max cover problem. Then, we will see that this problem is part of a sequence of problems having the same structure and that can be solved essentially with the same algorithm. This class of problems is the one of submodular function optimization problems.

3.1 k -max cover

Given $k \geq 1$ and a set of sets S , find a $S_k \subseteq S$, $|S_k| = k$, such that $|\bigcup_{s \in S_k} s|$ is maximum. There is a similar problem, the set-cover problem, where we want to cover all the elements with the minimum number of sets. Here we have a budget of k sets, and we want to cover as many elements as possible. The following algorithm gives a greedy solution (in fact, in approximation algorithms there are not so many techniques you can use).

```
Input:  $S, k$   
 $X_0 \leftarrow \emptyset$ ;  
for  $i = 1, \dots, k$  do  
     $s \leftarrow \operatorname{argmin}_{s \in S} |X_{i-1} \cup s|$ ;  
     $X_i \leftarrow X_{i-1} \cup s$   
end  
return  $X_k$ 
```

Algorithm 3: A greedy algorithm for k -max cover.

Theorem 3. *Algorithm 3 returns a $1 - \frac{1}{e}$ approximation of the optimal solution.*

Proof. Define OPT to be the value of the optimal solution. Let $t_i = |X_i|$. The value of the greedy solution is going to be t_k . Define

$$n_i = |s_i - X_{i-1}|, \tag{3.1}$$

the number of new elements introduced at iteration i . Finally, define

$$g_i = OPT - t_i, \tag{3.2}$$

that indicates how far the greedy solution is far from the optimal one at iteration i . We want to prove

$$g_k \leq \frac{1}{e} \iff t_k \geq \left(1 - \frac{1}{e}\right) OPT \quad (3.3)$$

Lemma 3.

$$n_{i+1} \geq \frac{g_i}{k} \quad (3.4)$$

In words, it looks at how far it is from the optimal solution at iteration i and when it picks the $(i+1)$ -th set it gets at least $\frac{1}{k}$ of what it was missing in the previous iteration.

Proof. Let $\{s_1^*, s_2^*, \dots, s_k^*\}$ be the sets representing the optimal solution and O^* their union. Let $T \subseteq O^*$, maybe $T = O^* - X_i$, for any i .

$$T \subseteq \bigcup_{i=1}^k s_i^* \implies \exists i \ |s_i^* \cap T| \geq \frac{|T|}{k} \quad (3.5)$$

The fact that O^* is realized by the choice of k sets that represents the optimal solution implies the existence of at least one set in O^* that covers at least a $\frac{1}{k}$ fraction of T . This is because if all sets in O^* covered less than that fraction, then overall they could not cover T and therefore they couldn't cover O^* .

There exists one set in the optimal solution that covers at least $\frac{1}{k}$ fraction of T . Even for the specific T suggested above. But if this is true, then it means that there exists one set in S that covers at least a $\frac{1}{k}$ fraction of $O^* - X_i = T$, and Algorithm 3 is going to pick at iteration $i+1$ the one set that maximizes the number of new elements it picks. So the new set Algorithm 3 will choose, will have at least as many elements as this one set in the optimal solution, that is $\frac{g_i}{k}$. \square

Lemma 4.

$$g_i \leq \left(1 - \frac{1}{k}\right)^i OPT, \forall i. \quad (3.6)$$

Proof. We prove this by induction. For $i=0$, $g_0 = OPT$ and so $g_0 \leq (1)OPT$ it is true. Let's assume Lemma 4 is true until i and we want to prove it for $i+1$.

$$g_{i+1} = g_i - n_{i+1} \quad (3.7)$$

$$\leq g_i - \frac{g_i}{k} \quad (\text{By Lemma 3})$$

$$= g_i \left(1 - \frac{1}{k}\right) \quad (3.8)$$

$$\leq OPT \left(1 - \frac{1}{k}\right) \quad (g_i \leq OPT)$$

$$< OPT \left(1 - \frac{1}{k}\right) \left(1 - \frac{1}{k}\right)^i \quad (3.9)$$

$$= OPT \left(1 - \frac{1}{k}\right)^{i+1}. \quad (3.10)$$

□

Instantiating Lemma 4 with $i = k$ we have

$$g_k \leq \left(1 - \frac{1}{k}\right)^k OPT \leq \frac{1}{e} OPT \quad (3.11)$$

□

In the next section we define what a submodular function is and then why k-max cover is a submodular optimization problem. We will discover that we can get a $(1 - \frac{1}{e})$ approximation for any problem in this class.

3.2 Submodular functions

Given a *ground set* V , a function $f : 2^V \rightarrow \mathbb{R}$ is a set function. We would like to maximize the value of this kind of functions, but if we have no information about them, the only thing we can do is to compute f for every set.

Suppose now we know about some of f 's structure. The function f is *modular* (or *linear*) if, given some set $A \in 2^V$,

$$f(A) = \sum_{i \in A} w(i), \quad (3.12)$$

where $w(i)$ is the value of element i . The function 3.12 can be maximized easily by simply looking at every one-element set ($n = |V|$ queries or computations) and return the set which contains all the elements with positive weight.

Suppose now that we want to maximize the value of the function under the constraint that the set picked must have cardinality k . We choose the k biggest values. These problems are easy with modular functions, and hard with general set functions.

Definition 1 (Submodular function). *Submodular functions are more general than linear functions and are less general than set functions. A function f is submodular if*

$$\forall S, T \subseteq V \quad f(S) + f(T) \geq f(S \cup T) + f(S \cap T). \quad (3.13)$$

They are a generalization of modular functions. In fact, one can prove that if f is modular then Equation 3.13 holds only with equality.

Let $S = A$, $B \subseteq A$, $x \notin A$, $T = B \cup \{x\}$. Let's apply the submodularity definition

$$f(A) + f(B \cup \{x\}) = f(A \cup \{x\}) + f(B) \quad (3.14)$$

that implies

$$f(B \cup \{x\}) - f(B) \geq f(A \cup \{x\}) - f(A). \quad (3.15)$$

Definition 2 (Discrete derivative). *Given a set function $f : 2^V \rightarrow \mathbb{R}$, $A \subseteq V$, $x \in V$, we define the discrete derivative of f at A with respect to x as*

$$\Delta(x|A) = f(A \cup \{x\}) - f(A). \quad (3.16)$$

In economics, Equation 3.16 is called *marginal return*. Putting together equations 3.15 and 3.16 we get the economics Law of diminishing returns.

Definition 3 (Marginal returns property). *Given $B \subseteq A \subseteq V$ and an item $x \in V \setminus B$, adding x to B increase the value of that set more than adding that element to a bigger set A which includes B . Formally*

$$\Delta(x|B) \geq \Delta(x|A). \quad (3.17)$$

We proved that Definition 1 implies Definition 3. We now prove the opposite direction, to show that the two definitions are equivalent.

Claim 1. *Definition 3 implies Definition 1*

Proof. Let $S, T \subseteq V$ be given. There are two cases, let's look at the simple one first. Assume $S \subseteq T$, then $S \cup T = T$ and $S \cap T = S$. The same conditions of Definition 1 are present, so we can just go backwards on the steps we performed to get to Definition 3.

So assume $S \not\subseteq T$ and $T \not\subseteq S$. This means that $S - T = \{s_1, \dots, s_k\}$ is not empty.

Lemma 5.

$$\forall 1 \leq i \leq k \quad f((S \cap T) \cup \{s_1, \dots, s_i\}) - f(S \cap T) \geq f(T \cup \{s_1, \dots, s_i\}) - f(T) \quad (3.18)$$

Proof. If $i = 1$ then we get the marginal returns property. The following inequality shows the base step of the induction.

$$f((S \cap T) \cup \{s_1\}) - f(S \cap T) \geq f(T \cup \{s_1\}) - f(T) \quad (3.19)$$

This lemma is all about applying marginal returns property to longer and longer stripes of elements.

What we want to do in the next step is adding s_2 . So we have

$$f((S \cap T) \cup \{s_1, s_2\}) - f((S \cap T) \cup \{s_1\}) \geq f(T \cup \{s_1, \dots, s_2\}) - f(T \cup \{s_1\}) \quad (3.20)$$

This is true by the marginal returns property. In fact, all this inequalities will be true up until

$$f((S \cap T) \cup \{s_1, \dots, s_{i+1}\}) - f((S \cap T) \cup \{s_1, \dots, s_i\}) \geq f(T \cup \{s_1, \dots, s_{i+1}\}) - f(T \cup \{s_1, \dots, s_i\}) \quad (3.21)$$

We are assuming that these inequalities hold by induction. Now we want to add up them, and we see that the terms cancel out. We are left with

$$f((S \cap T) \cup \{s_1, \dots, s_{i+1}\}) - f(S \cap T) \geq f(T \cup \{s_1, \dots, s_{i+1}\}) - f(T) \quad (3.22)$$

And so the claim is true. \square

Now observe what happens at the last step.

$$\begin{aligned} f((S \cap T) \cup \{s_1, \dots, s_k\}) - f(S \cap T) &\geq f(T \cup \{s_1, \dots, s_k\}) - f(T) \\ f((S \cap T) \cup (S - T)) - f(S \cap T) &\geq f(T \cup (S - T)) - f(T) \\ f(S) - f(S \cap T) &\geq f(T \cup S) - f(T) \\ f(S) + f(T) &\geq f(S \cap T) + f(S \cup T) \end{aligned}$$

And we have proved the equivalence between the two definitions. \square

Claim 2. *If f and g are submodular, then $f + g$ is submodular.*

Proof. $\forall S, T \subseteq V$

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$$

and

$$g(S) + g(T) \geq g(S \cup T) + g(S \cap T)$$

Define $h(S) = f(S) + g(S)$ and then sum the inequalities. The result is still submodular (the inequality holds). \square

If f is submodular, $-f$ is submodular? Only when the property holds at equality (hence, we don't switch sign).

Claim 3. *If f is submodular, αf is submodular, $\forall \alpha \geq 0$.*

Claim 4. *If f is submodular and g is modular then $f - \alpha g$ is submodular.*

Proof. g is modular, so multiplying it for a negative number results in a modular function, which is also submodular. Finally, the sum of two submodular function is submodular. \square

3.3 Facility location

Now we are going to give an example of submodular function optimization problem. Let $\mathcal{F} = \{F_1, \dots, F_n\}$ be a set of facilities and $\mathcal{U} = \{U_1, \dots, U_m\}$ a set of users. Maybe \mathcal{F} is the set of universities in Rome and \mathcal{U} the set of (wannabe) students in Rome. Student U_i gets a gain $M_{ij} \geq 0$ for starting a program at University F_j . So what a student would do? He would go to the facility that gives the maximum gain. If that's the protocol, it is all very easy. But suppose you are the government and you can't open all the n universities, but only k of them. We have to select $S = \{i_1, \dots, i_k : U_{i_k} \in \mathcal{U}\}$ facilities so to maximize $\sum_{i=1}^m \max_{j \in S} M_{ij}$.

What is the function we are going to prove to be submodular?

$$f(S) = \sum_{i=1}^m \max_{j \in S} M_{ij} \quad (3.23)$$

The function is pretty intricate, but we are going to use the properties we just saw to prove it being submodular. Thanks to Claim 2, we can focus only on proving

$$f_i(S) = \max_{j \in S} M_{ij} \quad (3.24)$$

to be submodular.

Claim 5. $\forall S, T \subseteq \mathcal{F}$

$$f_i(S) + f_i(T) \geq f_i(S \cup T) + f_i(S \cap T). \quad (3.25)$$

Proof. Let's concentrate on $f_i(S \cup T)$. Observe that, $f_i(S \cup T)$ is going to be larger or at least as large as the other f_i s in the inequality 3.25 because f_i is just a max over the elements of the set. Suppose the maximum value of $f_i(S \cup T)$ is realized by some j^* . This element can be in S , T , or both.

$$j^* \in S \implies f_i(S) = f_i(S \cup T) \quad (3.26)$$

$$j^* \in T \implies f_i(T) = f_i(S \cup T) \quad (3.27)$$

We are trying to prove that the left hand side is no larger than the left hand side. We have seen that $f_i(S \cup T)$ is the biggest part in the right hand side, and we have shown that, no matter what, there will be one element in the LHS that has the same value. Now we have to show that the other part of the RHS is not larger than the other element in the LHS.

Suppose k^* achieves the maximum at $f_i(S \cap T)$. What do we know? k^* is going to be in both S and T . So the max at S is going to be at least as large as the max obtained with k^* . It follows that

$$f_i(S) \geq f_i(S \cap T) \quad (3.28)$$

and

$$f_i(T) \geq f_i(S \cap T) \quad (3.29)$$

so

$$\min(f_i(T), f_i(S)) \geq f_i(S \cap T). \quad (3.30)$$

□

At this point we can state that K-MAX-COVER is FACILITY LOCATION with $M_{ij} \in \{0, 1\}$. Think of users as elements E_1, \dots, E_m and facilities as sets S_1, \dots, S_n .

$$M_{ij} = \begin{cases} 1 & \text{if } E_i \in S_j \\ 0 & \text{otherwise.} \end{cases} \quad (3.31)$$

Here, $f_i(S)$ is going to be the maximum across all the sets in the solution S and is going to be 1 if at least one set in S contains element i ; zero otherwise. Finally, we are summing up over all the elements we picked.

3.4 Approximation of submodular functions

We are going to see an algorithm that returns a $(1 - \frac{1}{e})$ approximation of the optimal solution for a submodular function under a cardinality constraint. It is going to be similar to the one seen for k-max cover.

Input: f, k
 $S_0 \leftarrow \emptyset$;
for $j = 1, \dots, k$ **do**
 | Let $v_i \in ([n] - S_{i-1})$ be the element maximizing $\Delta(v_i | S_{i-1})$;
 | $S_i \leftarrow S_{i-1} \cup \{v_i\}$;
end
return S_k ;

Algorithm 4: Greedy algorithm to approximate submodular functions.

Authors in [4] proved the following theorem.

Theorem 4. *If f is submodular, non-negative and monotone then Algorithm 4 returns a $(1 - \frac{1}{e})$ approximation of $\max_{S \in \binom{[n]}{k}} f(S)$.*

Proof. Let $S^* = \{v_1^*, \dots, v_k^*\} \in \binom{[n]}{k}$ be the optimal solution.

Lemma 6.

$$\forall i \leq k-1 \quad f(S^*) \leq f(S_i) + k \overbrace{(f(S_{i+1}) - f(S_i))}^{\text{new step gain}} \quad (3.32)$$

Proof. We look at the gain obtained by adding a new element at step i . In the following lines we are *telescoping* the first line to get intermediate steps.

$$\begin{aligned}
f(S^*) &\leq f(S^* \cup S_i) && \text{(monotonicity)} \\
&= f(S^* \cup S_i) - f(\{v_1^*, \dots, v_{k-1}^*\} \cup S_i) \\
&+ f(\{v_1^*, \dots, v_{k-1}^*\} \cup S_i) - f(\{v_1^*, \dots, v_{k-2}^*\} \cup S_i) \\
&+ f(\{v_1^*, \dots, v_{k-2}^*\} \cup S_i) - \dots \\
&\dots \\
&+ f(\{v_1^*, v_2^*\} \cup S_i) - f(\{v_1^*\} \cup S_i) \\
&+ f(\{v_1^*\} \cup S_i) - f(S_i) + f(S_i)
\end{aligned}$$

On each row we can see a diminishing return:

$$\begin{aligned}
&\Delta(v_k^* | S_i \cup \{v_1^*, \dots, v_{k-1}^*\}) \\
&+ \Delta(v_{k-1}^* | S_i \cup \{v_1^*, \dots, v_{k-2}^*\}) \\
&+ \dots \\
&+ \Delta(v_2^* | S_i \cup \{v_1^*\}) \\
&+ \Delta(v_1^* | S_i) \\
&+ f(S_i)
\end{aligned}$$

So we have

$$\begin{aligned}
f(S^*) &\leq f(S^* \cup S_i) && \text{(monotonicity)} \\
&= f(S_i) + \sum_{j=1}^k \Delta(v_j^* | S_i \cup \{v_1^*, \dots, v_{j-1}^*\}) \\
&\leq f(S_i) + \sum_{j=1}^k \Delta(v_j^* | S_i) && \text{(diminishing returns)} \\
&\leq f(S_i) + \sum_{j=1}^k \Delta(v_{i+1} | S_i) && \text{(greedy choice)} \\
&= f(S_i) + k\Delta(v_{i+1} | S_i) \\
&= f(S_i) + k(f(S_{i+1}) - f(S_i))
\end{aligned}$$

□

Define $\delta_i = f(S^*) - f(S_i)$ and also, for simplicity, $\delta_{i+1} = f(S^*) - f(S_{i+1})$. Remember that we want to claim we don't lose so much in going from one step to the next. By looking at $f(S_{i+1}) - f(S_i)$ we can't say much since we don't know the function f . However, let's look at the difference between δ_i and δ_{i+1} ; essentially we used the diminishing

return property to forget about the function. Now we can just think about the costs with respect to the optimum that we pay at each step.

$$\delta_i = f(S^*) - f(S_i) \tag{3.33}$$

$$\leq k(f(S_{i+1}) - f(S_i)) \tag{by Lemma 6}$$

$$= k(\delta_i - \delta_{i+1}) \tag{3.34}$$

Now we want to show that the cost goes down.

$$k\delta_{i+1} \leq (k-1)\delta_i \implies \delta_{i+1} \leq \left(1 - \frac{1}{k}\right) \delta_i \tag{3.35}$$

So

$$\delta_0 = f(S^*) - f(S_0) \leq f(S^*) \tag{f non-negative}$$

$$\delta_1 \leq \left(1 - \frac{1}{k}\right) \delta_0 = \left(1 - \frac{1}{k}\right) f(S^*) \tag{3.36}$$

$$\dots \tag{3.37}$$

$$\delta_k \leq \left(1 - \frac{1}{k}\right) \delta_{k-1} \leq \dots \leq \left(1 - \frac{1}{k}\right)^k \delta_0 \tag{3.38}$$

$$= \left(1 - \frac{1}{k}\right)^k f(S^*) \leq \frac{1}{e} f(S^*). \tag{3.39}$$

□

This can be thought as an example of how one can start by solving a specific problem, creating a solution, and then think of it as a general solution. One can ask what the created algorithm, or the analysis for proving it works, can say about a more general problem.

Do we need the properties we assumed, i.e. submodularity, non-negativity and monotonicity?

So, non-negativity is not that important. For one thing, if the function can be negative, what does it mean to optimize it and to give a constant approximation? Maybe the optimal is negative, in which case, say, a $(1 - \frac{1}{e})$ approximation would be impossible. In the context of maximization, multiplying the optimal (which we assume negative) for a constant $0 < x < 1$ gives a greater value than the optimal (contradiction). Thus, multiplying by a constant can change whether you are maximizing or minimizing. For what we have just said, non-negativity is necessary. On the other hand, you could remove that assumption keeping, in this particular case, a bound with $f(S^*) - f(S_0)$. It won't yield a constant approximation, because it doesn't make sense with negative functions, but you would get something that depends on this gap. Essentially, non-negativity is used to get a nice, clean theorem.

What about monotonicity? That actually helps. If it removed, then what we can get is a $\frac{1}{2}$ approximation, under some computational assumptions, but not a $(1 - \frac{1}{e})$ approximation. It would require a different way of reasoning about the problem and a much more complicated algorithm. So removing it means getting a worst approximation. Finally, we don't want to remove submodularity, we started from there.

Other questions are about constraints. Why do we care about optimizing the value of the function at sets of cardinality k ? One could say "I care about other sets". If the function is monotone and we want to maximize it without any constraint on the set, what should I do? Return the full set. Monotonicity is actually very important if the optimization problem comes with some constraint. As example, what if we do not want exactly k elements but at most k elements. In this case not having monotonicity is an important issue (maybe the empty set is the one that has more value).

3.5 Learning a submodular function

Algorithm 4 is able to approximate the value of the function by looking at $O(kn)$ values instead of $\Theta(2^n)$. It performs a number of steps that is logarithmic in the size of the input. In cases when there is no need to look at all the input, the algorithm is said to be sublinear.

We saw that linear (modular) functions can be understood by only looking at singleton sets (plus the empty set), that is, $n + 1$ values. It is natural to think that we can't do the same for submodular functions, and what follows is the idea to prove it.

What is a construction that can actually prove that, while there is an algorithm that efficiently optimize the function, there is no algorithm that can learn the function efficiently? As usual we want to use easy building blocks, and then mix up them according to the claims we saw earlier (summing them up, essentially).

For $T \subseteq [n]$, define the function

$$f_T(S) = \begin{cases} 0 & \text{if } S \subseteq T, \\ 1 & \text{otherwise.} \end{cases} \quad (3.40)$$

The function f_T is a building block that we want to use to build the final submodular function F . Consider

$$\mathcal{S} \subseteq U = \binom{[n]}{\frac{n}{2}} = \left\{ Q : Q \subseteq [n] \wedge |Q| = \frac{n}{2} \right\}$$

F will depend on \mathcal{S} , but the fact we don't know actually what \mathcal{S} is will lead us to the result we want to prove. So, let's define

$$F_{\mathcal{S}}(A) = \sum_{T \in \mathcal{S}} f_T(A). \quad (3.41)$$

There are exponentially many functions f_T , which are all the possible building blocks. We are looking for a subset of them, and one subset is to be picked without it being

disclosed; now, the value of the submodular function F is the sum of the values of the functions at a particular set.

Claim 6. f_T is submodular.

Proof. Let's pick two sets $A, B \subseteq [n]$, and apply the definition of submodularity.

$$f_T(A) + f_T(B) \geq f_T(A \cup B) + f_T(A \cap B). \quad (3.42)$$

There are four cases to be considered.

1. $A, B \subseteq T$. Then we have $0 + 0 \geq 0 + 0$, which holds.
2. $A \subseteq T$ and $B \not\subseteq T$ (and the specular case). Computing the function values we have that $0 + 1 \geq 0 + 1$.
3. $A, B \not\subseteq T$. In this case we have $1 + 1 \geq 0 + x$. We can't tell exactly the value x of the function computed at intersection, but it can be at most 1, so the inequality holds.

□

Since the sum of submodular functions is submodular, $F_S(A)$ is submodular. We want to prove that knowing F requires a lot of queries. Doing it formally would require many definitions, but it is a pretty easy algorithm. Your algorithm's call is to learn F by querying it. The answer it gets at each query is some integer representing the number of functions f_T that evaluates to 1 at set A .

How big and how small that number can be? The function F is nonnegative, so the minimum is 0 and the maximum is the cardinality of S , which can be as large as 2^n . $F_S(A)$ is an integer from 0 to 2^n , and thus requires n bits for each query. We will show that the number of bits needed to represent F is much larger, exponential. Why is it the case? How big is the set of sets U ?

$$|U| = \binom{n}{\frac{n}{2}} = \Theta\left(\frac{2^n}{\sqrt{n}}\right). \quad (3.43)$$

Since we are picking any subset \mathcal{S} of this big set of sets, to be able to know which subset \mathcal{S} is, we actually have to know for each set in U whether i picked it or not. One bit is needed for each set, summing up to $|U|$ bits. At each query we get only n bits, therefore at least $\frac{|U|}{n}$ queries are needed (i.e., an exponential number).

Why do we need to know exactly S ? Well, if we don't know S , the value of F can't be known.

This is why learning submodular functions is hard. But we can optimize them with efficient algorithms, doing leverage on certain properties as we saw.

Chapter 4

Random Graph Models

In this chapter we are going to take a look at random graph models developed to represent properties of social networks. The first one is the Preferential Attachment model, a very famous one, introduced to explain the *power law* that characterize the degree distribution in big social graphs. The other one justifies the fact that between any two people there is a minimal path with really few "hops", fact known as *small world phenomenon*. One thing that we anticipate is that is really difficult to find a single model that contains all the social networks properties; instead researchers were able to produce algorithms that in general won't work, but on special cases they perform really very well.

4.1 Preferential Attachment Model

It has been observed that the degree distribution in big social graphs, like the one of Facebook or Twitter, is really close to a *power law* distribution.

Definition 4. A probability distribution P follows a power law if

$$P(x) = x^{-\alpha}, \alpha > 1. \quad (4.1)$$

In the social network context it means that it is exponentially more likely to pick "normal people" with few friends or followers rather than popular profiles, called "celebrities" or "authorities".

It turns out that the Preferential Attachment Model is able to capture this social networks' aspect, whereas others random models like $G(n, p)$ (the one proposed by Edgar Gilbert) do not show this regularity. In $G(n, p)$, for example, we expect each node to have the degree very close to the mean, because of the Chernoff Bound.

Nodes with high degree in a social network are few, but they exist. So, as an example, an advertisement agency could pay those celebrities to publicize a product, enabling a spread of information due to the high number of connections those nodes have.

If we plot a power law distribution on a *loglog plot*, that is, a plot where on both axes are represented the log of coordinates of each point, we get a straight line that says

how many nodes for each degree one can get, which is kind of unusual to observe in real world phenomena, thought to be more “chaotic”. A first scientific question one can ask is how can it be that something so striking.

We would rather expect that there exists some average degree from which values would deviate with a an exponential law, resulting in very small deviations from the mean; something similar to what the Chernoff bound states. And yet is not what happens here. In understanding why this the case, the Preferential Attachment Model is really important. It was rediscovered by many people. In physics, this model is usually attributed to Barabási and Albert, but also discovered in mathematical biology, and in math. It is called also called “rich get richer” model, and we will see why.

First of all, it differs from $G(n, p)$ because it is evolutionary, not fixed. It is possible to stop at n nodes if needed, but it is not required. There are many variants. In our discussion, the model starts with one node with a self loop. Then, at each time step t , a new node is added, which then is linked to an existing node in the graph. The new node will choose its neighbor in some random way. In the simplest model it chooses one neighbor, but it can be generalized to k neighbors (this is not considered here). The neighbor is selected with a *rich get richer* policy. This means that a node is chosen with a probability proportional to its degree. Richer nodes, the ones with high degree, have a greater probability to be chosen as neighbor of a new node.

Formally, given the graph G_{t-1} , the probability of a node $v \in V(G_{t-1})$ to be chosen at time t to be the neighbor of the new node is

$$P(v) = \frac{\deg(v)}{\sum_{v' \in V(G_{t-1})} \deg(v')} \quad (4.2)$$

It follows that the probability distribution may change every time a new node is added. It is evolutionary, and we can’t make an independent choice for each edge. When we are in such a situation, studying the graph is hard, in fact there are many things we don’t know about this model. But it is possible to study the degree distribution and to show that it is actually a power law distribution.

At first glance, assuming a graph with n nodes, it seems that to add the $(n + 1)$ -th node one has to look at the full graph. This means that, if $n = 1000$, then we have to look at 1000 nodes to decide where to put the edge. The first observation is, if we only care about the degree distribution, we can reduce the knowledge about the graph that we care about; in particular, if we know the vector of degree counts at time $t - 1$ then it is possible to compute the new vector of degrees at time t . This is the main trick we will be using.

Let’s define some notation. Let

$$X_i^{(t)} = \text{“Number of nodes with degree } i \text{ in } G_t\text{”}. \quad (4.3)$$

The following properties are always true (we start with one node with a self loop):

$$X_2^{(1)} = 1, \quad X_i^1 = 0 \quad \forall i \neq 2. \quad (4.4)$$

$$|E(G_t)| = t, \quad |V(G_t)| = t \quad (4.5)$$

Furthermore, the sum of all degrees in the graph is twice the number of edges.

Can we compute the expected number of nodes with degree, say, one? In the end, we want to be able to know the expected number of nodes of any degree, but now let's focus on the base case. That number depends on the past, since the process is evolutionary, which we can represent as the vector $\overline{X^{(t-1)}}$ containing the degree distribution at $t - 1$ time step. We first assume a given realization of the vector $\overline{X^{(t-1)}}$.

Define

$$\xi_1 = \text{"The new node chooses a degree 1 neighbor"}. \quad (4.6)$$

Then, when adding a new node, either ξ_1 or its complementary will happen. In the first case, the number of nodes stays the same, since we add and remove one; in the second case, we increase the count by one.

$$\begin{aligned} E \left[X_1^{(t)} | \overline{X_1^{(t-1)}} \right] &= (X_1^{(t-1)} + 1)(1 - \Pr\{\xi_1\}) + X_1^{(t-1)} \Pr\{\xi_1\} \\ &= (X_1^{(t-1)} + 1) \left(1 - \frac{X_1^{(t-1)}}{2(t-1)} \right) + X_1^{(t-1)} \frac{X_1^{(t-1)}}{2(t-1)} \\ &= X_1^{(t-1)} - \frac{(X_1^{(t-1)})^2}{2(t-1)} + 1 - \frac{X_1^{(t-1)}}{2(t-1)} + \frac{(X_1^{(t-1)})^2}{2(t-1)} \\ &= \left(1 - \frac{1}{2(t-1)} \right) X_1^{(t-1)} + 1. \end{aligned}$$

So now, by the law of total probability,

$$\begin{aligned} E \left[X_1^{(t)} \right] &= \sum_{c \geq 0} \Pr\{X_1^{(t-1)} = c\} E \left[X_1^{(t)} | X_1^{(t-1)} = c \right] \\ &= \sum_{c \geq 0} \Pr\{X_1^{(t-1)} = c\} \left(\left(1 - \frac{1}{2(t-1)} \right) c + 1 \right) \\ &= \sum_{c \geq 0} \Pr\{X_1^{(t-1)} = c\} \left(1 - \frac{1}{2(t-1)} \right) c + \sum_{c \geq 0} \Pr\{X_1^{(t-1)} = c\} \\ &= \left(1 - \frac{1}{2(t-1)} \right) \sum_{c \geq 0} \Pr\{X_1^{(t-1)} = c\} c + 1 \\ &= \left(1 - \frac{1}{2(t-1)} \right) E \left[X_1^{(t-1)} \right] + 1. \end{aligned}$$

Lemma 7.

$$E \left[X_1^{(t)} \right] = \left(1 - \frac{1}{2(t-1)} \right) E \left[X_1^{(t-1)} \right] + 1. \quad (4.7)$$

What we get is a recursive formula to compute the expected number of nodes with degree one. The “plus one” is the new node we add, and the others nodes with degree one will almost be preserved except for the $\frac{1}{2(t-1)}$ fraction we might end up losing because of a one degree node removal (by attaching the new edge to it).

How can we hope to do the same thing for degrees larger than one? In this case, the difference is that the new node won't have degree i , so what could happen? The number of nodes having degree i might not be affected, or can increase or decrease by one. If, for example, we cared about nodes with degree three, and the new one attaches to one of such nodes, then the count decreases. If the new node attaches to a previously degree two node, then the count increases. Formally,

$$\begin{aligned}
E[X_i^{(t)} | \overline{X^{(t-1)}}] &= X_i^{(t-1)} + Pr\{X_i^{(t)} - X_i^{(t-1)} = 1\} - Pr\{X_i^{(t)} - X_i^{(t-1)} = -1\} \\
&= X_i^{(t-1)} + \frac{i-1}{2(t-1)} X_{i-1}^{(t-1)} - \frac{i}{2(t-1)} X_i^{(t-1)} \\
&= \left(1 - \frac{i}{2(t-1)}\right) X_i^{(t-1)} + \frac{i-1}{2(t-1)} X_{i-1}^{(t-1)} \\
&= E[X_i^{(t)} | X_i^{(t-1)}, X_{i-1}^{(t-1)}].
\end{aligned}$$

It follows that

$$\begin{aligned}
E[X_i^{(t)}] &= \sum_{c,d \geq 0} Pr\{X_i^{(t-1)} = c \wedge X_{i-1}^{(t-1)} = d\} E[X_i^{(t)} | X_i^{(t-1)} = c, X_{i-1}^{(t-1)} = d] \\
&= \sum_{c,d \geq 0} Pr\{X_i^{(t-1)} = c \wedge X_{i-1}^{(t-1)} = d\} \left(\left(1 - \frac{i}{2(t-1)}\right) c + \frac{i-1}{2(t-1)} d \right) \\
&= \sum_{c \geq 0} Pr\{X_i^{(t-1)} = c\} \left(1 - \frac{i}{2(t-1)}\right) c + \sum_{d \geq 0} Pr\{X_{i-1}^{(t-1)} = d\} \frac{i-1}{2(t-1)} d \\
&= \left(1 - \frac{i}{2(t-1)}\right) E[X_i^{(t-1)}] + \frac{i-1}{2(t-1)} E[X_{i-1}^{(t-1)}].
\end{aligned}$$

Lemma 8.

$$\forall i \geq 2, E[X_i^{(t)}] = \left(1 - \frac{i}{2(t-1)}\right) E[X_i^{(t-1)}] + \frac{i-1}{2(t-1)} E[X_{i-1}^{(t-1)}]. \quad (4.8)$$

Our original goal was to prove that

$$E[X_i^{(t)}] \propto i^{-\alpha} t \quad (4.9)$$

that is, a power law behavior. This is a very rough estimate but it is representative of what we would like to achieve; in other words, to bound the expectation with the

number of steps we had so far. Currently we cannot achieve this because we can express the expectation at time t only as a function of the expectation at the previous time step.

We need to solve recurrent linear systems (Lemma 7, Lemma 8). Lemma 7 is easy, since it has only one free variable; Lemma 8 is not as easy. It is actually pretty much complicated, and exists general, hard, techniques to solve it. In such situations, the best thing to do is to guess the solution and check that it works. Still, if one wants to solve the recursion, you can evaluate the base case (Lemma 7), and then plug the solution into the general case (Lemma 8) iteratively.

We will give a solution to bound $E \left[X_i^{(t)} \right]$ and check if it is correct.

Theorem 5. Let $E_i = \frac{4}{i(i+1)(i+2)}$ be our guess on the expectation $E \left[X_i^{(t)} \right]$.

$$E_i t - 1 \leq E \left[X_i^{(t)} \right] \leq E_i t + 1 \quad (4.10)$$

It is not the exact solution, but it is within a constant from it, and since E_i grows with t , the constant factor doesn't matter in practice.

Proof. As said previously hinted, we will do a double induction to prove the theorem. The outer one is on i , and the inner one on t . So, assume $i = 1$, the base case of the outer induction. Let's proceed with the inner one.

Since $i = 1$, the graph has only one node with a self loop, by construction. Why 1 was chosen as error constant? We started from one node with degree two. If you think about, we could have started from any other graph, but we had to change the error factor.

In the inner base case, $t = 1$. Then

$$E \left[X_1^{(1)} \right] = 0 \quad (4.11)$$

and

$$\frac{2}{3} - 1 \leq E \left[X_1^{(1)} \right] \leq \frac{2}{3} + 1. \quad (4.12)$$

Actually, Inequality 4.12 is true for any i . Now we assume that Equation 4.10, with $i = 1$, is true up until $t - 1$. Let's prove that it is true also at time t . We prove the upper bound first.

$$E[X_1^{(t)}] = \left(1 - \frac{1}{2(t-1)}\right) E[X_1^{(t-1)}] \quad (4.13)$$

$$\leq \left(1 - \frac{1}{2(t-1)}\right) \left(\frac{4}{6}(t-1) + 1\right) + 1 \quad (\text{By inductive hypotesis})$$

$$= \frac{2(t-1)}{3} - \frac{2(t-1)}{6(t-1)} - \frac{1}{2(t-1)} + 1 \quad (4.14)$$

$$= \frac{2(t-1)}{3} + \left(1 - \frac{1}{3}\right) + \left(1 - \frac{1}{2(t-1)}\right) \quad (4.15)$$

$$= \frac{2}{3}t + \left(1 - \frac{1}{2(t-1)}\right) \leq \frac{2}{3}t + 1. \quad (4.16)$$

Now, the lower bound.

$$E[X_1^{(t)}] = \left(1 - \frac{1}{2(t-1)}\right) E[X_1^{(t-1)}] \quad (4.17)$$

$$\geq \left(1 - \frac{1}{2(t-1)}\right) \left(\frac{4}{6}(t-1) - 1\right) + 1 \quad (\text{By inductive hypotesis})$$

$$= \frac{2(t-1)}{3} - \frac{2(t-1)}{6(t-1)} + \frac{1}{2(t-1)} + 1 \quad (4.18)$$

$$= \frac{2}{3}t - \frac{2}{3} - \frac{1}{3} + \frac{1}{2(t-1)} + 1 \quad (4.19)$$

$$= \frac{2}{3}t + \frac{1}{2(t-1)} \geq \frac{2}{3}t - 1. \quad (4.20)$$

The inner induction is terminated. We have proved for the base case $i = 1$. Assume Equation 4.10 is true up until $i - 1$, for any t . We what to show that it is true also for the number of nodes with degree i . Let's now prove the upper bound.

$$E[X_i^{(t)}] = \left(1 - \frac{i}{2(t-1)}\right) E[X_i^{(t-1)}] + \frac{i-1}{2(t-1)} E[X_{i-1}^{(t-1)}] \quad (4.21)$$

$$\leq \left(1 - \frac{i}{2(t-1)}\right) (E_i(t-1) + 1) + \frac{i-1}{2(t-1)} (E_{i-1}(t-1) + 1) \quad (4.22)$$

$$= \left(1 - \frac{i}{2(t-1)}\right) \left(\frac{4(t-1)}{i(i+1)(i+2)} + 1\right) \quad (4.23)$$

$$+ \frac{i-1}{2(t-1)} \left(\frac{4(t-1)}{(i-1)(i)(i+1)} + 1\right) \quad (4.24)$$

$$= \frac{4(t-1)}{i(i+1)(i+2)} + 1 - \frac{2}{(i+1)(i+2)} - \frac{i}{2(t-1)} + \frac{2}{i(i+1)} + \frac{i-1}{2(t-1)} \quad (4.25)$$

$$= \frac{4(t-1)}{i(i+1)(i+2)} + 1 + \frac{4}{i(i+1)(i+2)} - \frac{1}{2(t-1)} \quad (4.26)$$

$$= \frac{4t}{i(i+1)(i+2)} + 1 - \frac{1}{2(t-1)} \leq E_i t + 1. \quad (4.27)$$

The lower bound proof have essentially the same steps. We show it for completeness.

$$E[X_i^{(t)}] = \left(1 - \frac{i}{2(t-1)}\right) E[X_i^{(t-1)}] + \frac{i-1}{2(t-1)} E[X_{i-1}^{(t-1)}] \quad (4.28)$$

$$\geq \left(1 - \frac{i}{2(t-1)}\right) (E_i(t-1) - 1) + \frac{i-1}{2(t-1)} (E_{i-1}(t-1) - 1) \quad (4.29)$$

$$= \left(1 - \frac{i}{2(t-1)}\right) \left(\frac{4(t-1)}{i(i+1)(i+2)} - 1\right) \quad (4.30)$$

$$+ \frac{i-1}{2(t-1)} \left(\frac{4(t-1)}{(i-1)(i)(i+1)} - 1\right) \quad (4.31)$$

$$= \frac{4(t-1)}{i(i+1)(i+2)} - 1 - \frac{2}{(i+1)(i+2)} + \frac{i}{2(t-1)} + \frac{2}{i(i+1)} - \frac{i-1}{2(t-1)} \quad (4.32)$$

$$= \frac{4(t-1)}{i(i+1)(i+2)} - 1 + \frac{4}{i(i+1)(i+2)} + \frac{1}{2(t-1)} \quad (4.33)$$

$$= \frac{4t}{i(i+1)(i+2)} - 1 + \frac{1}{2(t-1)} \geq E_i t - 1. \quad (4.34)$$

□

As final thought one can observe that this model is very linear, since we get the recursive equations; also, if you care only about the degree distribution, then you only need the degree count vector. In general, knowing a model well enough can make run your computations faster, discarding information you don't need. There are some things we can carve out of this lesson. As example, how big a clique can be in social network? Say it can contain 20 people more or less (compute the largest clique is NP-complete).

In this model, the biggest clique is of size two, since it is a tree. It follows that it isn't a great representative of a social network under this aspect. Yet, it can give us one of the most striking properties, that is the power law degree distribution.

4.2 Small World Phenomenon

In the Sixties Stanley Milgram and other researchers conducted an experiment on the average path length for social networks of people (in the physical world, of course) in the United States. He wanted to claim that the six degree of separation property is valid (actually, he only thought that the average path length was small, then discovered the mean value).

Here's what he did. He picked something like a thousand people in Boston, Massachusetts, from the address book in a random fashion; then, he selected one guy in New York (a friend of his), let's call him *target*. He sent a letter to each of the 1000 people along the following lines:

I would like this message to reach *target*, which resides in NYC, but you can only forward this message to people whom you know personally, and asking them to do the same if they are not *target*.

Each guy tries to reach *target* by sending the letter to the closest person he knows. What Milgram observed is that the average length of the chains of letters was 6.

In the following decades, this experiment was confirmed by other studies. The main consequence is the understanding of fact that the social graph diameter is small. In fact Jon Kleinberg observed something more striking: not only the chain among two random people is small in average, but people can find the next step without looking at the full graph. Each of the guys didn't know the shortest path, but only a sense of where their friend were.

Kleinberg tried to find a model that could explain the small diameter and allow the discovery of the chain efficiently.

In the model Kleinberg proposed, people live on a bi-dimensional grid, $n \times n$, and each person lives in exactly one cell of the grid, denoted by some coordinates (i, j) . We can think of it also as a graph where the nodes are

$$V = \{(i, k) : 1 \leq i, j \leq n\}. \quad (4.35)$$

The simplest model works as follows. Each person has short links to its geographical neighbor people. In addition, each person u has one long range contact v (denoted by $u \rightsquigarrow v$), where u and v are not geographical neighbors; this can be motivated by the fact that sometimes people travel and become friends with people who live far away. A person (i, j) chooses as long range friend (k, l) (thus establishing a link) as follows.

$$Pr\{(i, j) \rightsquigarrow (k, l)\} \propto (|i - k| + |j - l|)^{-2} = l_1((i, j), (k, l))^{-2}, \quad (4.36)$$

where l_1 is the Manhattan distance. The exponent might feel arbitrary but we will show that, if the long range links are chosen in this way, then the expected length of the chains will be polylogarithmic. Any other value different from -2 will lead to a polynomial length. In a sense, one of the most striking practical observations Kleinberg made is that Equation 4.36 is the density at which long links should exist for this property to hold, and people have observed that the same type of density is what happens to exist in real social networks. The back-then Yahoo Messenger researchers, analyzing all the users' connections, confirmed that Equation 4.36 represents the correct guess form which links distribution is induced.

In terms of the graph, the actual probability is

$$Pr\{u \rightsquigarrow v\} = \frac{d(u,v)^{-2}}{\sum_{z \neq u} d(u,z)^{-2}} \quad (4.37)$$

where $d = l_1$. Observe that the long range links are chosen independently.

We would like to prove that chains have length $\log^2(n)$ in average. This isn't the same as saying that there exist one path of $\log^2(n)$; that wouldn't give us what we want. In fact, it can be shown that there always exists a path of length $\log(n)$, that is shorter. Rather, we want to claim that the algorithm we are going to see requires $\log^2(n)$ many steps (where each step is associated to a link) to find chains in practice. Also it is a tight bound, no other algorithm can do better.

Input: a node s , destination t , the message M ;

$c \leftarrow s$;

while $c \neq t$ **do**

$w \leftarrow$ neighbor of c that has the smallest l_1 distance to t (break ties arbitrarily);
 c forwards (M, t) to w ;
 $c \leftarrow w$;

end

Algorithm 5: GEO-GREEDY algorithm

Algorithm 5 resembles the Milgram experiment. At each iteration, the current person c forwards the letter to the closest friend to the target t he knows.

Before proving any theorem, let's discuss a proof technique and make some observations along the way. First of all, the process we are studying can be in principle hard to study, but there are some properties of the random graph that help a lot.

The first one is that each node chooses its long range contact independently at random. Equation 4.37 can be thought as a distribution over random graphs. Why this can help us? If we get to a new node, then we can assume that exactly at the time we get there the long range contact is chosen, since the choice of the long range contact is independent of other random choices we made so far. The other thing is that we never get to the same node twice. You can always make one step closer to the destination. The distance of the current node in the Algorithm 5 decreases by at least one at each step.

These properties allow us to study the algorithm and produce the random graph as the algorithm proceeds. As Algorithm 5 goes to a new node, a random independent choice is made at random. So it is easier than what happened in the Preferential Attachment model.

What we want to prove is that, no matter which node we are, there is a reasonable chance that we will be able to halve the distance from ourselves to the final destination. Suppose that, using the random choice, we can prove that the probability of halving the distance is at least some p (think of it as a large number). Then, after $\frac{1}{p}$ steps we will have divided by two the distance to the destination with constant probability. This proof idea has been seen different times by now, in the chain letters and coin tossing, as examples.

If t is the final destination, let's look at a large ball around t . If we are some node c there is a reasonable chance that our long range contact will end up inside this ball. The radius of the ball is half of the distance from t that we are currently at.

Theorem 6. *For any source s and destination t , GEO-GREEDY will make an expected number of steps to go from s to t of $O(\log^2(n))$.*

Proof. Remember that, in the end, we want to show that the probability for our long range link to end up in the ball is high. For this purpose, it is necessary to estimate the number of nodes in the ball. But first, let's give a bound to the normalizing factor in Equation 4.37. By looking closer to it, we can derive the following statement.

$$\sum_{z \neq u} d(u, z)^{-2} = \sum_{i \geq 1} i^{-2} |\{z : d(z, u) = i\}| \quad (4.38)$$

We actually have to bound the number of nodes at any geographical distance from any given node. Let $S_u^r = \{z : d(z, u) = r\}$. What is the value of $|S_u^r|$? It is the number of nodes on the surface of the ball of radius r and centered in u . We will upper bound the size of the surface in the following way.

$$S_u^r = \{z : d(z, u) = r\} \quad (4.39)$$

$$\subseteq \{(i + k, j + r - k) : 0 \leq k \leq r\} \quad (4.40)$$

$$\cup \{(i + k, j - r - k) : 0 \leq k \leq r\} \quad (4.41)$$

$$\cup \{(i - k, j - r - k) : 0 \leq k \leq r\} \quad (4.42)$$

$$\cup \{(i - k, j + r - k) : 0 \leq k \leq r\} \quad (4.43)$$

The reason why we partitioned the edges in this way is that it is easy to compute the cardinalities. In fact we easily state that

$$|S_u^r| \leq 4(r + 1). \quad (4.44)$$

We can improve the upper bound a little bit, because we are overcounting by an additive factor of four. By removing the elements in the intersections (the elements “at the corners”) we can state the following Lemma.

Lemma 9.

$$|S_u^r| \leq 4(r+1) - 4 = 4r. \quad (4.45)$$

But is this bound tight for any node in the graph (or grid)? The nodes at the edges can get a better upper bound. A node in the corner will have only one of the above sets, while a node on the edge will have two. As r becomes bigger and bigger, we will exceed the size of the grid at some point in time, since it is finite, and the surface will be empty. This leads us to the following statement.

Lemma 10.

$$\forall r \leq \frac{n}{2} \quad |S_u^r| \geq r \quad (4.46)$$

We give the idea to prove this lemma. Divide the grid in 4 parts. A node u is in one of those 4 parts; then it means we can pick a node at distance $\frac{n}{2}$ in both the adjacent quadrants' direction, defining a set of $r+1 \geq r$ nodes on the segment that connect the two picked nodes. We move $\frac{n}{2}$ steps in both direction, and pick the nodes that define the edge set.

To upper bound the normalizing factor shown in Equation 4.38, we will make use of the following observations.

Observation 1.

$$\sum_{i=1}^t \frac{1}{i} \leq 1 + \ln t + 1 \quad (4.47)$$

Observation 2.

$$1 + \log_b x = \log_b bx \quad (4.48)$$

Proof.

$$1 + \log_b(x) = \log_b(b) + \log_b(x) \quad (4.49)$$

$$= \log_b(bx) \quad (4.50)$$

□

From Equation 4.38 it can be derived that

$$\sum_{z \neq u} d(u, z)^{-2} = \sum_{i \geq 1} i^{-2} |\{z : d(z, u) = i\}| \quad (4.51)$$

$$\leq \sum_{i \geq 1}^{2n-2} i^{-2} 4i \quad (\text{Lemma 9})$$

$$= 4 \sum_{i \geq 1}^{2n-2} \frac{1}{i} \quad (4.52)$$

$$\leq 4(1 + \ln(2n-1)) \quad (\text{Observation 1})$$

$$\leq 4(1 + \ln(2n)) \quad (4.53)$$

$$= 4 \ln(2en) \quad (\text{Observation 2})$$

$$\leq 4 \ln(6n). \quad (4.54)$$

We have a more algebraic bound on the normalizing factor, and we can state the following Lemma.

Lemma 11.

$$\Pr \{u \rightsquigarrow v\} \geq \frac{d(u, v)^{-2}}{4 \ln(6n)}, \quad (4.55)$$

Now we can go back to the algorithm's behavior, analyzing how much time we need to halve the distance from where we are to final destination. We say that Algorithm 5 is in phase j if its current node c satisfies the constraint

$$2^j \leq d(t, c) \leq 2^{j+1}. \quad (4.56)$$

Observe that phases go in the unusual direction. The algorithm starts from phase j , with j that can be as large as $\ln(n+1)$ in the worst case, and then j decreases. Essentially, we want to show that the time to go from one phase to the other is small.

Define the random variable X_j to be

$$X_j = \text{number of steps that we spend in phase } j. \quad (4.57)$$

The total number of steps X the algorithm will need, then, is

$$X = \sum_{j=0}^{1+\log(n)} X_j. \quad (4.58)$$

In this way we are bucketing the number of steps the Algorithm 5 takes according to phases. The reason why is that it's easy to upper bound the expected number of steps in a phase, because the distances between a node of a specific phase to the destination are going to be all within a constant factor to each other. If we manage to show that X_j is logarithmic then we are done. What does it mean that a phase, in expectation, will require at most $\log(n)$ many steps? It can happen because we have to take at least one long range link, and we will prove it now.

As we said, we want to hit a ball around the target when choosing the next long range link, and it happens with roughly the probability of being a node in the ball times the number of nodes in the ball. We said "roughly" because each node has not the same probability to be chosen, but they are not so different. We need to know how many nodes are in the ball and, for the probability of getting it the ball to be high, we need to lower bound that number. This is different from what we were doing before. To prove the lower bound on the probability of reaching a specific node, in Lemma 11, we had to upper bound the number of nodes in the sphere, whereas here we want the number of nodes in the ball to be large.

Here we care about the content of the sphere too, because it is sufficient to hit any node in the sphere, actually the closer to the center the better.

Definition 5. We define the ball of radius r and centered in u as

$$B_u^r = \{u : d(u, v) \leq r\}. \quad (4.59)$$

To lower bound the volume of the ball we can observe that

$$|B_u^r| \geq \left| B_u^{\lceil \frac{r}{4} \rceil} \right|, \quad (4.60)$$

since the radius of the ball in the RHS is smaller, making the ball cover less nodes. The reason for $\frac{r}{4}$ is that the lower bound in Lemma 10 works for radius values up to $\frac{n}{2}$ and the actual radius of the ball we care about might be large as the diameter of graph, roughly $2n$. What is the size of $B_u^{\lceil \frac{r}{4} \rceil}$? It is the sum of the sizes of the smaller spheres that are inside the ball. This is why we need to use Lemma 10. We make the ball smaller so that $2n$ will become $n/2$. Some contribution will be lost, but it has to be done.

$$\left| B_u^{\lceil \frac{r}{4} \rceil} \right| = \sum_{i=0}^{\lceil \frac{r}{4} \rceil} |S_u^i|. \quad (4.61)$$

Let's take out the cardinality of the sphere with radius 0, which is one.

$$\left| B_u^{\lceil \frac{r}{4} \rceil} \right| = \sum_{i=1}^{\lceil \frac{r}{4} \rceil} |S_u^i| + 1 \quad (4.62)$$

$$\geq \sum_{i=1}^{\lceil \frac{r}{4} \rceil} i + 1 \quad (\text{Lemma 10})$$

$$= 1 + \frac{\lceil \frac{r}{4} \rceil (\lceil \frac{r}{4} \rceil + 1)}{2} \quad (4.63)$$

$$\geq \frac{r^2}{32}. \quad (4.64)$$

Lemma 12.

$$|B_u^r| \geq \frac{r^2}{32}. \quad (4.65)$$

We now have all the pieces we need. We know how big the ball is going to be and we have a good lower bound on the probability of our current node to have a long range link to any node in the ball.

Take phase j and consider any node $x \in B_t^{2^j}$. Let c be the current node in Algorithm 5.

$$d(c, x) \leq d(c, t) + d(t, x) \quad (4.66)$$

$$\leq 2^{j+1} + 2^j \quad (4.67)$$

$$\leq 2^{j+2}. \quad (4.68)$$

What we got is that, for any node in the final ball we are trying to hit, the distance between c and that one node is not that large. Let's apply Lemma 11.

$$Pr\{c \rightsquigarrow x\} \geq \frac{d(c, x)^{-2}}{4 \ln(6n)} \quad (4.69)$$

$$\geq \frac{2^{-2j-4-2}}{\ln(6n)} \quad (4.70)$$

$$= \frac{2^{-2j-6}}{\ln(6n)}. \quad (4.71)$$

Inequality 4.71 shows the probability of c to have the long range contact inside the ball in phase j . Why is this useful? We have the square of the radius many nodes in the ball. The radius is 2^j , which yields to 2^{2j} by substitution; multiplying that quantity with the lower bound shown in Inequality 4.71 results in a constant over $\log(n)$, which is large enough.

$$Pr\left\{\exists x \in B_t^{2^j} : c \rightsquigarrow x\right\} = \sum_{x \in B_t^{2^j}} Pr\{c \rightsquigarrow x\} \quad (4.72)$$

$$\geq \sum_{x \in B_t^{2^j}} \frac{2^{-2j-6}}{\ln(6n)} \quad (4.73)$$

$$\geq \frac{(2^j)^2}{32} \frac{2^{-2j-6}}{\ln(6n)} \quad (4.74)$$

$$= \frac{2^{2j-5-2j-6}}{\ln(6n)} \quad (4.75)$$

$$= \frac{2^{-11}}{\ln(6n)} = \frac{1}{2048 \ln(6n)}. \quad (4.76)$$

The constant is large, but we didn't pay too much to make it smaller; it doesn't matter for our purposes anyway. Now, how many steps the algorithm will spend in phase j ? We have a reasonable probability of hitting the ball and going to the next phase.

Observation 3.

$$\sum_{i=0}^{\infty} (1-x)^i = \frac{1}{x}, \quad 0 < x < 1. \quad (4.77)$$

Let's compute the expected number of steps.

$$E[X_j] = \sum_{i=1}^{\infty} i \Pr\{X_j = i\} \quad (4.78)$$

$$= \sum_{i=1}^{\infty} \Pr\{X_j \geq i\} \quad (4.79)$$

$$\leq \sum_{i=1}^{\infty} \left(1 - \frac{1}{2048 \ln(6n)}\right)^{i-1} \quad (4.80)$$

$$= \sum_{i=0}^{\infty} \left(1 - \frac{1}{2048 \ln(6n)}\right)^i \quad (4.81)$$

$$= 2048 \ln(6n). \quad (4.82)$$

Finally, notice that we are summing the expected value $\log(n)$ times, yielding to an expected running time of $\log^2(n)$.

□

There are a few things that could be said. One is that this proof only shows that in expectation the algorithm takes $\log^2(n)$ steps; what can be shown is that this happens with high probability. The second thing is, as observed by Kleinberg, if the probability doesn't use the constant -2 , then the time to completion becomes polynomial. In the exposed work he proposed a model according to which people chooses long range contacts. The model was validated years later looking at data, a very surprising outcome. Other observations are on the model. We started from a grid, but it is not that important. The proof has been recreated for other geometries. Finally, one can regards this as a beautiful example of using theory to guess about reality, without being able to meet that guess.

Bibliography

- [1] Abrahao et al. “Trace Complexity of Network Inference”. In: *Knowledge Discovery and Data Mining* (2013).
- [2] Chierichetti, Kleinberg, and Liben-Nowell. “Reconstructing Patterns of Information Diffusion from Incomplete Observations”. In: *Advances in Neural Information Processing Systems* (2011).
- [3] Gomez-Rodriguez, Leskovec, and Krause. “Inferring Networks of Diffusion and Influence”. In: *Knowledge Discovery and Data Mining* (2010).
- [4] Nemhauser, Wolsey, and Fisher. “An analysis of approximations for maximizing submodular set functions”. In: *Mathematical Programming* (1978).