

# AC290r.1: Influence Maximization Problem

Stephen Carr and Charles Liu

October 13, 2015

## 1 Introduction

### 1.1 Goal and Motivation

From a marketing perspective, social networking sites provide a large number of users with personal and persuasive information regarding products. With limited resources, one would want to know how to reach advertisement to the most number of people in this personalized manner. This problem can be reduced to a graph, where each user is a node and each relationship between users is represented by an edge with some weight to indicate the strength of their relationship. Which user, or set of users, has the ability to spread your product to the greatest number of potential customers?

### 1.2 Data set

The user network of the online restaurant reviewing site Yelp is used in this study. Each account represents the node, and the edges are to be generated by looking at which accounts posted reviews of the same restaurant within a certain time period of one another. This information (the user, their total number of reviews, and how many reviews they have written within 1 month of another user) was pulled from the Yelp site, turned into a library format, and then imported into a python development environment using network x. For most of the prototyping of a solution, a subset of the North Carolina network was taken, which consisted of 240 nodes and ??? edges.

### 1.3 Independent Cascade, $I(s)$

We start with a set of nodes  $s = \{n_1, n_2, \dots, n_k\}$  and wish to find the number of nodes which are activated through our probabilistic network of influence. So for each node  $n$  in our original set  $s$ , we look at all nodes connected to  $n$  by an edge and use the beta distribution to determine if the connected node  $n_c$  activates (we ignore  $n_c$  if it is already activated). The beta distribution  $\mathbf{B}$  is given by:

$$\boxed{\int_{-\infty}^{\infty} \frac{dx}{\mathbf{B}(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} = 1} \quad (1)$$

Where  $\beta$  is taken to be the number of reviews node  $n_c$  has written within one month of node  $n$  writing a review of the same restaurant, and  $\alpha$  is taken to be the number of reviews  $n_c$  has written. A uniform distribution is taken between 0 and 1 and if its value is less than  $\mathbf{B}(\alpha, \beta)$  then  $n_c$  activates and we repeat the same process for neighbors of  $n_c$ . This recursive process continues until either all nodes are activated or no activated nodes are left in the loop. The number of nodes activated is called  $I(s)$ .

## 2 Stochasticity

### 2.1 Averaged Trial Function, $f_N(s)$

Since each step of the independent cascade is a random sampling of the beta distribution, there is no guarantee that two different measurements of  $I(s)$  will be the same, or even "close" to one another, as seen in 1.

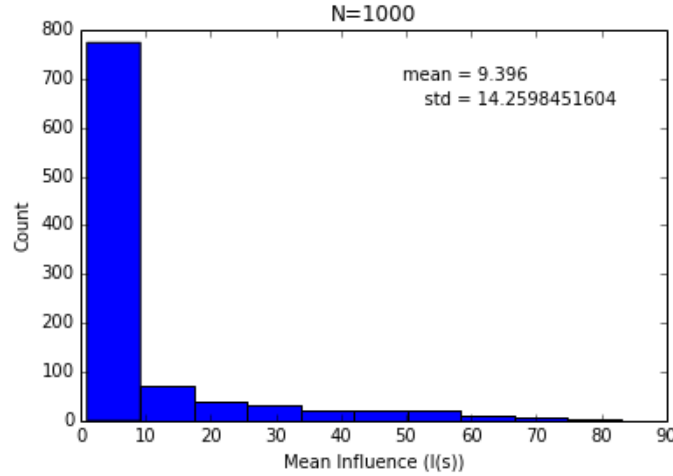


Figure 1: Histogram of values  $I(s)$  over 1000 iterations for a fixed node in the 240 North Carolina subgraph.

Because of this, we must define a new measurement of a node's influence, namely its mean over  $N$  samples:

$$f_N(s) = \frac{1}{N} \sum_{i=1}^N I(s) \quad (2)$$

### 2.2 Properties of $f_N(s)$

Obviously we cannot take  $N$  arbitrarily large, so we must understand how  $f_N$ 's properties vary with  $N$ . Taking a 240 node subset of North Carolina, we obtain

the following relationship between  $N$  and the standard deviation of  $f_N$  in 2.

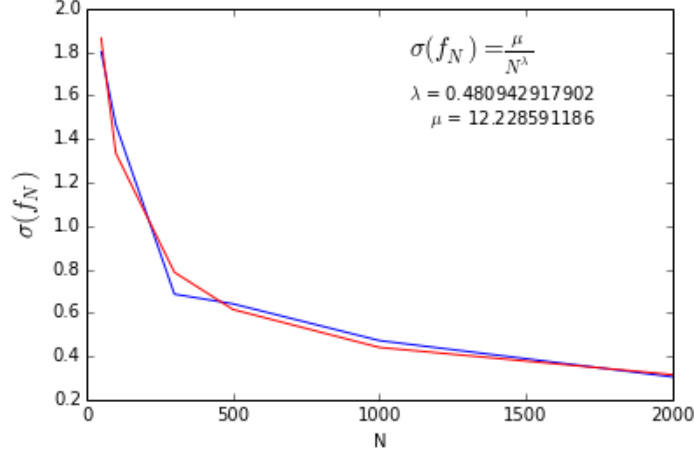


Figure 2: Fitted regression for the standard deviation of  $f_N(s)$ . This was with 100 trials for each  $N$  value. The blue line is the simulation, while the red line is the fitted relationship.

### 2.3 Greedy Algorithm

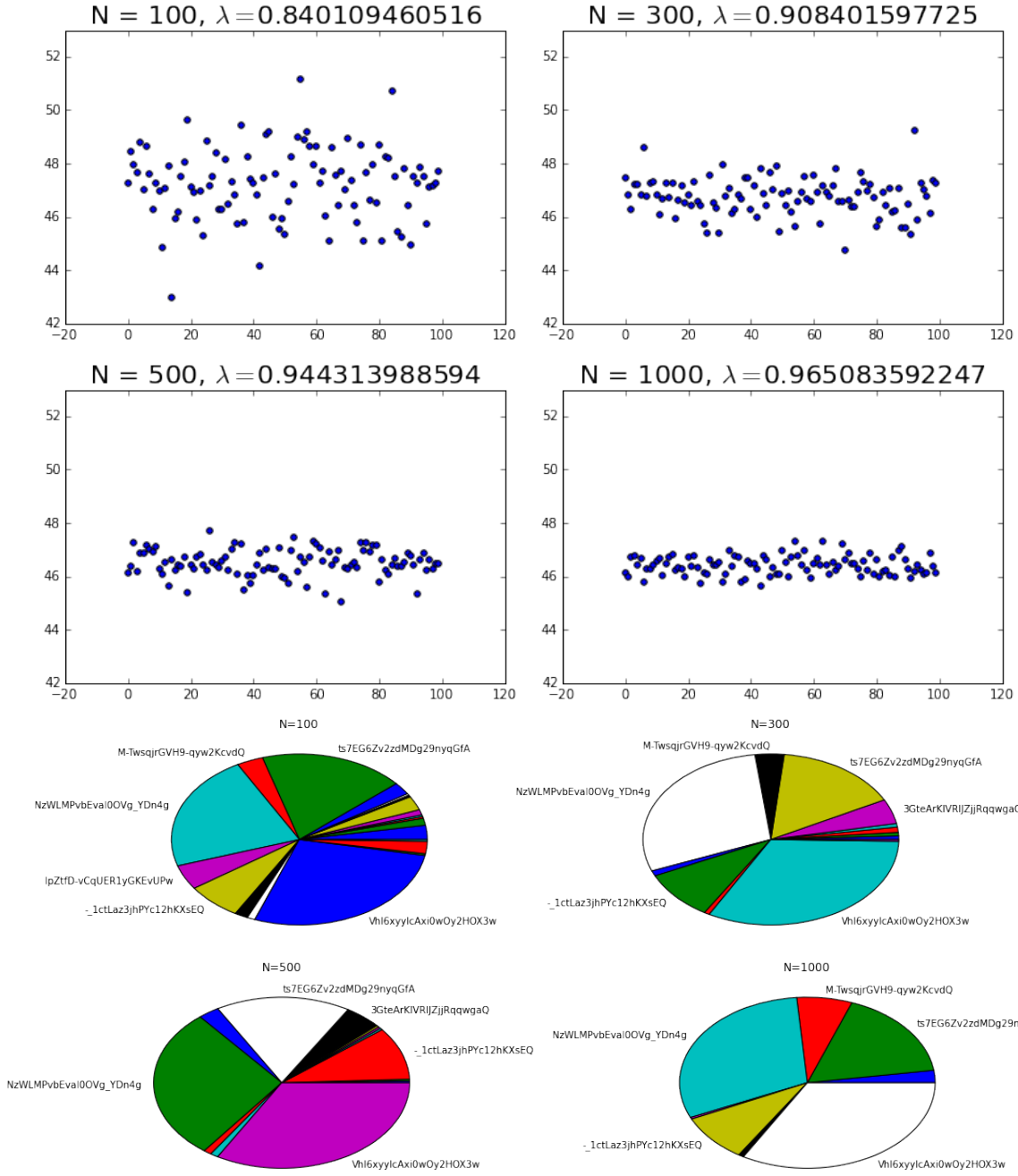
Although this problem is inherently non-linear in its parameters (the starting set of nodes), we can try to solve it in a linear way. This is called the "Greedy Algorithm": Given nodes  $s_i = \{n_1, n_2, \dots, n_i\}$ , check every remaining node to find node  $n_{i+1}$  that maximizes  $f(s_{i+1})$  where  $s_{i+1} = \{n_1, n_2, \dots, n_i, n_{i+1}\}$ .

The greedy algorithm of order  $k$  looks for  $s_k$  satisfying the above condition. The greedy algorithm is an "embarrassingly parallel" problem. Using Spark, in one iteration of the greedy algorithm, we can distribute the nodes evenly and run the influence function. For  $k = 3$  on our 240 node North Carolina subset we found variation in the maximal  $I(s)$ .

Table 1: 100 runs with  $k=3$ ,  $t=10$

N	min	max	$\lambda$
100	42.98	51.16	0.840
300	44.73	49.24	0.908
500	45.04	47.696	0.944
1000	45.66	47.31	0.965

The scatterplots of the max values over the 100 trials show the convergence of results as  $N$  increases.



We can similarly see the convergence through the number of nodes selected over the course of our trials.

## 3 Implementation

### 3.1 MCMC and Simulated Annealing

For a large problem with  $N$  nodes, the greedy algorithm would need to check roughly  $N^k$  node arrangements (first  $N$ , then  $N - 1$ , etc.). For  $k > 2$  this will not be feasible. Instead, we implement a Monte Carlo method of picking sets of nodes at random and evaluating their influence. Although this method will not guarantee we find the optimal solution, we can define an iterative method (Markov Chain) that on average will find a better solution after each step.

The idea of a Markov Chain Monte Carlo (MCMC) algorithm is to use the function of interest to assign a weight to every position in parameter space, and use this weight to add control to the "random steps" of a Monte Carlo method. Our function is the mean influence,  $f_N(s)$ , and our parameter space is the unordered combination of  $k$  nodes  $s = \{n_i\}$ . Instead of always taking a "random step" around parameter space, we assign to each step a probability of success:

$$P(s_1, s_2) = e^{\frac{-\Delta E}{T}}, \Delta E(f(s_2), f(s_1)) \quad (3)$$

Where  $s_1$  is our current set of nodes and  $s_2$  is a new proposed set. Controlling the Monte Carlo method in this manner is known as Simulated Annealing, as the probability formula mimics the Boltzmann factor of statistical mechanics. The Boltzmann factor describes the relative probability of a particle being in state  $s$  when it is in a thermal equilibrium of temperature  $T$ . Annealing such a system means heating it up and then slowly cooling it down, allowing the particle(s) to settle into a new arrangement, hopefully of lower energy. For our problem we take the "Energy" to be  $-f_N(s)$ , since we want to maximize  $f_N(s)$ .

The simplest form for  $\Delta E$  is  $f(s_1) - f(s_2)$ . In this case, if  $f(s_2) > f(s_1)$  we take  $s_2$  with probability 1, otherwise there is a chance of taking  $s_2$  even though its estimated influence is less than our current state. This gives us an additional parameter to control our MCMC method: the "Temperature"  $T$  controls how likely we are to take a step away from a maximum. So when we are not "near" a local maximum, a high  $T$  allows us to sweep the parameter space quickly, but as we near a solution we take a low  $T$  to make the MCMC very sensitive to small changes in  $f_N(s)$ .

### 3.2 Advantages and Shortcomings

As the MCMC is an iterative process that gives a solution from the first step that slowly improves with each iteration, it is better than the greedy algorithm which needs to finish in its entirety before an answer can be output. Furthermore, the greedy algorithm does not guarantee to find the answer, and it should find the same sub-optimal solution everytime (provided  $N$  is large enough). The MCMC will "eventually" find the global maximum influence, as it will eventually cover all possible permutations of nodes. So for large graphs, the MCMC method is

clearly superior, so the only work left to do is to find how to speed up each iteration without sacrificing accuracy. A number of methods to speed up the MCMC process is covered next.

### 3.3 Choice of Parameters for Speed

#### 3.3.1 $\Delta E$

For small graphs,  $\Delta E = f(s_1) - f(s_2)$  will work. But for large graphs, this will allow  $P(s_1, s_2)$  to be unsuitable large or small. Instead we take  $\Delta E = \log(f(s_1)) - \log(f(s_2))$ , allowing  $P(s_1, s_2)$  to be tractable both conceptually and computationally.

#### 3.3.2 $N$

Table of  $\mu$  and  $\lambda$  as a function of  $f_N(s)$  for NC small, NC large, (And US?)

#### 3.3.3 $T$

Start with  $T = 1$ . After each successful step, set  $T_{n+1} = \alpha * T_n$  where  $\alpha < 1$ .

#### 3.3.4 Allowed Nodes ( $s_i$ )

Only allow the top 5% or 10% of nodes (by edge count or by edges of neighbors). For generating  $s_{n+1}$  given  $s_n$ , randomly swap one of the nodes of  $s_n$  to any other node ("closeness" is controlled by keeping the other nodes the same, NOT by staying "near" the changed node on the graph).

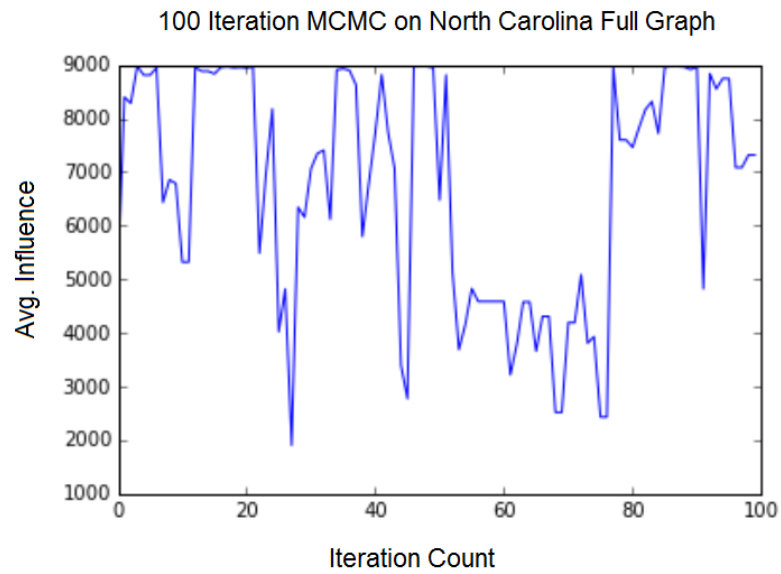
#### 3.3.5 Re-annealing Control

If no new local maximum is found in  $k$  attempts, set  $T = 1$  to "restart" the search process (i.e. allow for "large steps" again but start from the current set of nodes).

## 4 Results

### 4.1 North Carolina (mini)

### 4.2 North Carolina (full)



Max influence = 8975.7

Nodes (by hash): u'9ki\_I7oW6YxnUZBCmxs5Q', u'G1lP-1YGO1m61G4ly5fYTQ',  
u'iCePpGSAGAOSYZ9TkCQ09w'

All MCMC runs seem to max out around 9000 influence...?