

**CS599: Structure and Dynamics of Networked Information (Spring 2005)**  
**04/18/2005: Finding a good node to add in the greedy algorithm**  
**Scribes: Affan Syed and Hui Zhang**

In the last two classes, we showed that the greedy algorithm gave us a  $(1 - 1/e)$ -approximation for the problem of choosing a node set of largest expected total influence [2]. We did not yet show how to actually find such a node. In fact, it is not known how (or whether) one can find, in polynomial time, the node giving largest marginal gain at any point. However, we can prove that we can find an almost best node, which will be enough for our purposes.

**Lemma 1** *In polynomial time, we can find a  $(1 - \epsilon)$ -approximately best node to add.*

**Proof.** First, we describe an algorithm for estimating the marginal gain of adding a node. If we can get sufficiently accurate estimates for each node, then a choice based on these estimates will give us an approximately best node.

The algorithm is simply to simulate the random process multiple times, and take the average of the number of additional nodes reached over all these simulations. This average will certainly in the limit converge to the expectation, but the question is how quickly, i.e., how often is “multiple” times? If some “gain values” had very low probability of occurring, but were extremely high, then we may need a large number of simulations to get a good estimate. Fortunately, in our case, the values we sample have a natural upper bound of  $n$ , the number of nodes. We will see that the number of iterations will not need to be very high as a result.

Formally, we will use the Chernoff-Hoeffding Bound, as given by the following theorem.

**Theorem 2 (Chernoff-Hoeffding Bound [1])** *If  $X_1, \dots, X_k$  are independent random variables with  $0 \leq X_i \leq b_i$  for all  $i$ , and  $X = \sum_i x_i$ , and  $\mu = E[X]$ , then for all  $\Delta \geq 0$ ,*

$$\text{Prob}[|X - \mu| \geq \Delta] \leq 2e^{-\frac{\Delta^2}{\sum_i b_i^2}}.$$

In our case, we let  $X_i$  be the outcome of the  $i^{\text{th}}$  simulation. Thus,  $0 \leq X_i \leq n$  for all  $i$ , so  $b_i = n$ . Also, because at least one node is active in each outcome (the start node itself), we have that  $X_i \geq 1$ , and thus  $\mu \geq k$ . Choosing  $\Delta = \epsilon\mu$ , we obtain that

$$\text{Prob}[|X - \mu| \geq \epsilon\mu] \leq \text{Prob}[|X - \mu| \geq \epsilon k] \leq 2e^{-\frac{(\epsilon k)^2}{kn^2}} = 2e^{-\frac{\epsilon^2 k}{n^2}}.$$

Thus, if we are aiming for a  $(1 \pm \epsilon)$ -approximation with probability at least  $1 - \delta$ , it is sufficient to require that  $2e^{-\frac{\epsilon^2 k}{n^2}} < \delta$ . Solving for  $k$  gives that  $k > \frac{n^2}{\epsilon^2} \log \frac{2}{\delta}$  simulations are sufficient. By a Union Bound over all (at most  $n$ ) iterations of the greedy algorithm, and all  $n$  nodes in each iteration, all simulations have error at most  $\epsilon$  with probability at least  $1 - n^2\delta$ .

If we want, for instance, success probability at least  $1 - \frac{1}{n^4}$ , we can choose  $\delta = \frac{1}{n^4}$ , and then, we need to run  $O(\frac{n^2}{\epsilon^2} \log n)$  iterations to get a  $(1 \pm \epsilon)$  accurate estimation.

It still remains to verify that we actually obtain a close to best node when picking a node based on  $\epsilon$ -estimates. In the worst case, the picked node’s gain is over-estimated by a factor of  $(1 + \epsilon)$ , while the true best node’s gain is under-estimated by a factor of  $(1 - \epsilon)$ . But because the picked node appeared to be better, its gain must have been within a factor of  $\frac{(1-\epsilon)}{(1+\epsilon)} \geq 1 - 2\epsilon$  of the best node’s. So we have a  $(1 - 2\epsilon)$ -approximate best node in each iteration, in polynomial time. ■

By essentially mimicking our earlier proof of the Nemhauser-Wolsey-Fischer theorem, we can show the following stronger version:

**Theorem 3** *If we choose a  $(1 - \epsilon)$ -approximate best node in each iteration of the greedy algorithm, then we get a  $1 - \frac{1}{e} - \epsilon'$  approximation algorithm, where  $\epsilon' \rightarrow 0$  as  $\epsilon \rightarrow 0$ , polynomially fast.*

## References

- [1] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. of the American Statistical Association*, 58:13–30, 1963.
- [2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence in a social network. In *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining*, pages 137–146, 2003.