CS 7641 Project

Maximizing the Spread of Influence through a Social Network

Aemen Lodhi
Sabyasachi Deyati

1 Introduction

In this project we explore the results presented by Kempe et al. [4] in "Maximizing the Spread of Influence through a Social Network". We evaluate the experimental aspects of the paper on the dataset used by Kempe et al. We also evaluate their propositions on an additional dataset and find that their results are qualitatively consistent. The rest of the report is organized as follows. In section 2 we present a summary of the paper. Section 3 describes the dataset used by Kempe et al. In section 4 we discuss our implementation of the mechanisms discussed in the paper followed by a discussion of our results in section 5. We discuss additional experiments that we carried out in section 6. Finally, conclude in section 7.

2 Summary of the paper

In [4] Kempe et al. consider the problem of ideas/influence propagation through a social network. A social network can be considered to be a graph of relations and interactions among a group of individuals. The main question addressed in the paper is that if a subset of individuals in the network can be convinced to adopt a certain idea/product, can they in turn cause a cascade of adoption in the network? Subsequently, who should be chosen in the subset in first place so that the adoption is optimized. In other words, the paper studies $diffusion processes$ on the network. These problems have been discussed in the context of different applications including marketing, technology diffusion etc. The ideas in the paper have general application and are not limited to a particular application.

The main question is how to choose the first few individuals (or the initial subset) to *seed* the diffusion process. Assuming that there is a cost associated with influencing those first set of individuals, we would like to choose the smallest number of such individuals who can influence the maximum individuals in the network. Thus, it is a problem in discrete optimization. Various heuristics have been proposed in the literature for this purpose based on the well-studied notions of $degree centrality$ and $distance centrality$.

The optimal solution is $NP-hard$. The main contribution of the paper is to show (theoretically) that the optimal solution for influence maximization can be efficiently approximated to within a factor of $1-1/e-\epsilon$ where e is the base of the natural logarithm and ϵ is any positive real number. Kempe et al. propose a natural greedy hill-climbing algorithm which achieves this performance guarantee. This performance guarantee achieves 63% of the theoretical performance limit. Their proposed algorithm also outperforms existing heuristics and algorithms. The authors also compare their proposed algorithms with existing heuristics on a network dataset that exhibits many features of large-scale social network.

2.1 Diffusion models

An *influence model* determines how a node in the network influences its neighbors. The authors compare their algorithms under three different models of influence which are described in this section.

2.1.1 Linear Threshold

In this model a node v is influenced by each neighbor w according to a weight b_{vw} such that:

$$\sum_{w \in neighbor\ of\ v} b_{vw} \leq 1$$

The dynamics of the process is given below. Each node v is pre assigned a value θ_v where $\theta_v \in$ [0,1] . Thus, θ_v is the threshold of the node in influence propagation. This represents the weighted fraction of $v's$ neighbor that must become active in order for v to become active. Given a random choice of θ_v for all nodes and a initial population A_0 (all other nodes are inactive) the diffusion process unfolds deterministically in every time step: in step t all the nodes that were active in step $t-1$, will remain active and we will activate node v if total weight of its active neighbors is at least θ_v. The process ends when no more new activation is possible.

2.1.2 Independent Cascade

This model also starts with initial population A_0 and unfolds itself by the following randomized rule. If a node v first got active in time step $t-1$ then in time step t it will have a single chance to influence its inactive neighbors w. It succeeds with probability $p_{v,w}$ which is a parameter of the system and independent of the node its connection, its neighbors and previous history. (if w has multiple newly activated neighbors their attempts are sequenced in random order) If v is successful in activating w then w will be active in time step $t+1$. But whether or not v is successful it will become inactive from time step $t+1$. The process ends when no more activation is possible.

2.1.3 Weighted Cascade

For weighted cascade the activating probability is not governed by system parameter any more. It is now depends on local node connection and its neighbors. So $p_{v,w}$ will not be same throughout the network. Each edge from node u to v have a probability of $1/d_v$ of activating its neighbor v. (d_v is the number of neighbors of node v).

2.2 Initial set choice algorithms

The following four algorithms and heuristics are used in the paper for the choice of initial subset. The maximum size of the initial subset is set at k.

Random: k nodes are randomly chosen as initial active nodes.

High Degree: k nodes with the highest degree are chosen.

High Centrality: k nodes with highest distance centralities are chosen as initial population.

Greedy: This is the proposed algorithm in the paper. The algorithm as described in the paper [4] is quoted here:

"For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let $S*$ be a set that maximizes the value of f over all $k-element$ sets. Then $f(S) \geq (1-1/e).f(S*)$; in other words, S provides a $(1-1/e)$- approximation."

3 Dataset

The dataset used in the paper is a scientific collaboration network. The dataset is available at [2]. The following description has been given at the link:

"Arxiv HEP-PH (High Energy Physics - Phenomenology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to High Energy Physics - Phenomenology category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes."

The network features are given in table 1.

Characteristics	Values
Nodes	12008
Edges	237010
Nodes in largest WCC	11204 (0.933)
Edges in largest WCC	235268 (0.993)
Nodes in largest SCC	11204 (0.933)
Edges in largest SCC	235268 (0.993)
Average clustering coefficient	0.6115
Number of triangles	3358499
Fraction of closed triangles	0.6595
Diameter	13
90-percentile effective diameter	5.8

Table 1: Dataset characteristics

4 Our implementation

The diffusion models and algorithms were implemented in MATLAB. Figure 1 gives the implementation of the linear threshold model, figure 2 gives the implementation of the independent cascade model and figure 3 gives the implementation of weighted cascade model. We also used the MATLAB Boost Graph library [3] for graph algorithms.

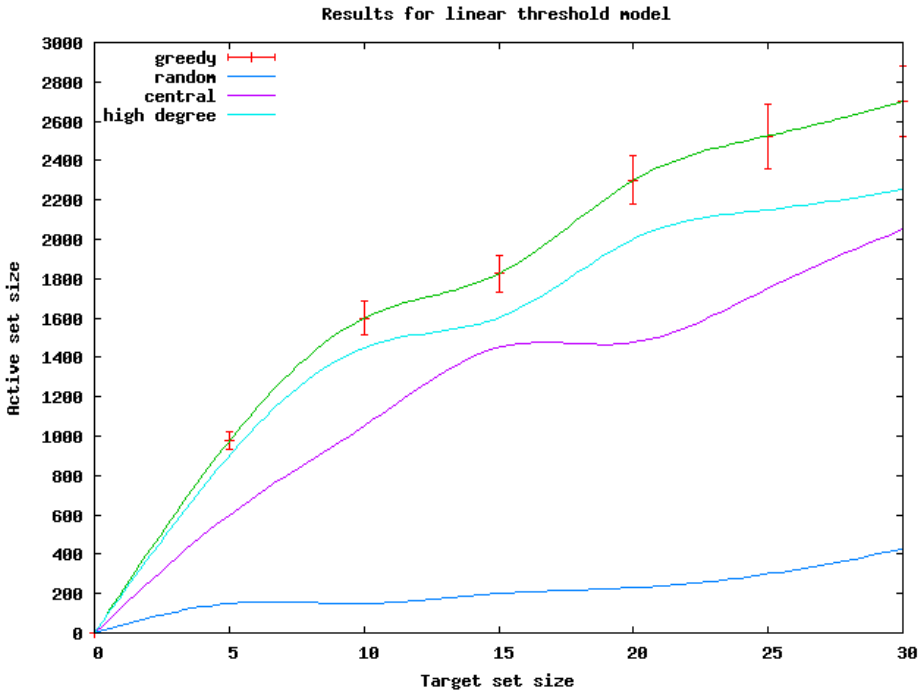

Figure 1: Linear Threshold diffusion model

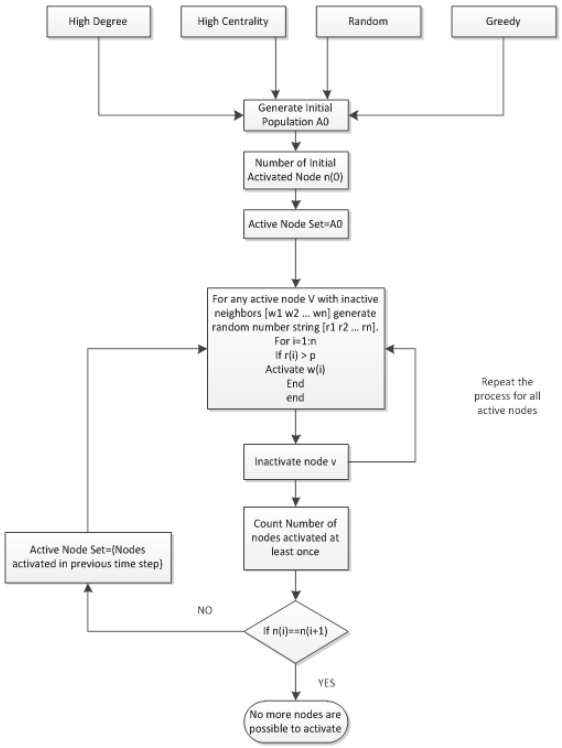

Figure 2: Independent Cascade

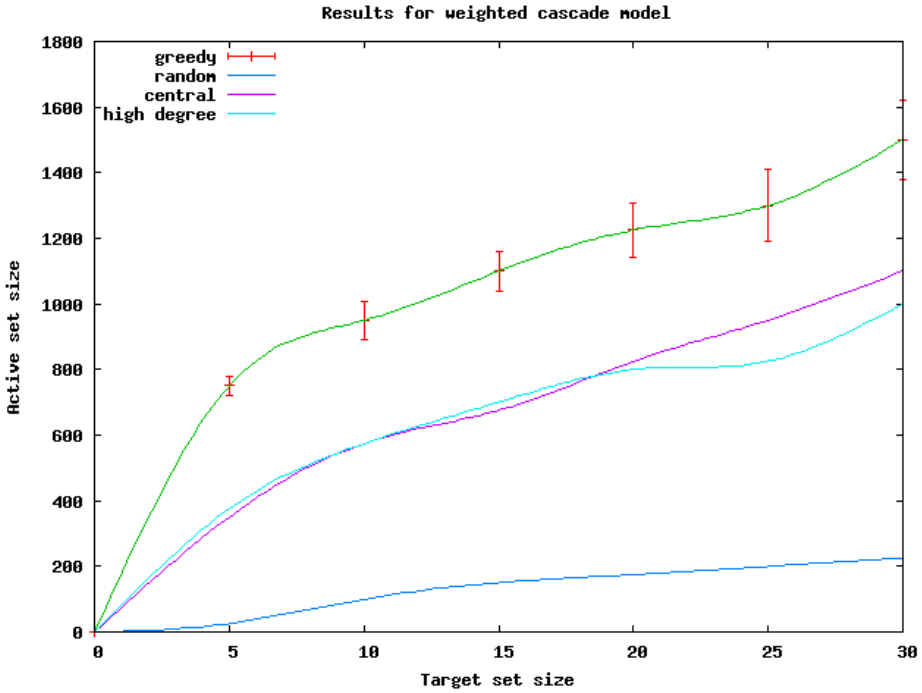

Figure 3: Weighted Cascade

5 Results

We carried out the experiments with the three diffusion models and four algorithms. Our results verify the findings of Kempe et al [4] that the greedy algorithm out performs other heuristics and algorithms. Our results show greater number of nodes in the final active set than the paper since the dataset size has increased over time. Our results show that the proposed greedy algorithm always outperforms other heuristics. We find that randomly choosing the nodes is the worst choice. We also find that choosing the nodes based on high degree slightly outperforms the choice based on centrality.

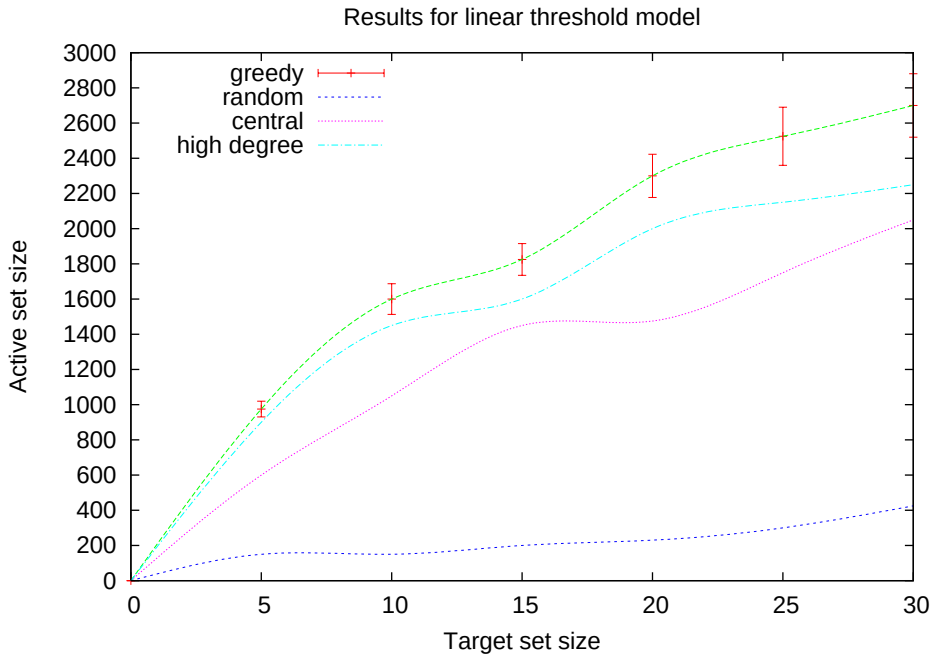

Figure 4: Results for the linear threshold model

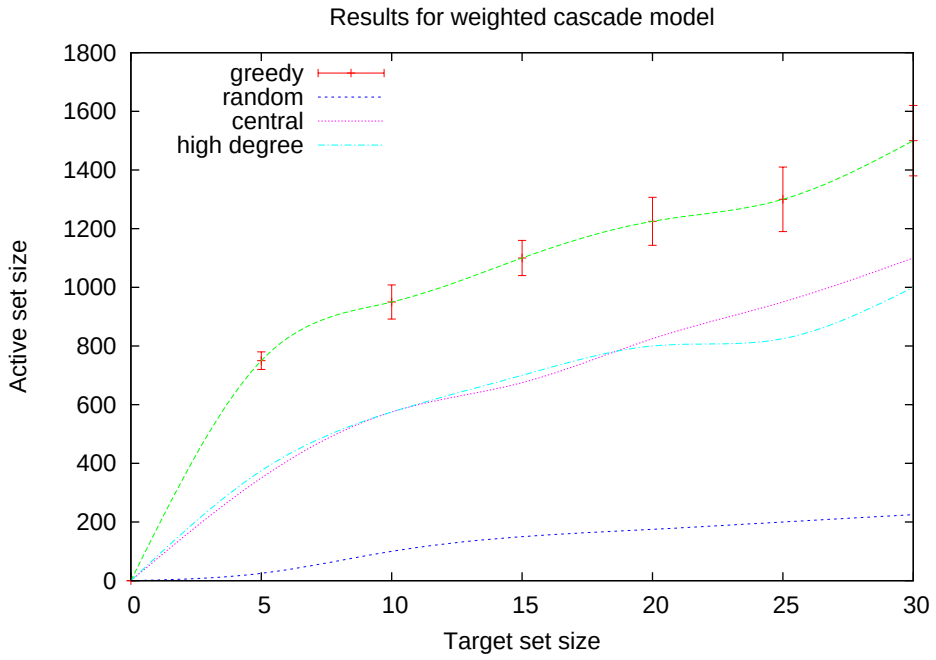

Figure 5: Results for the weighted cascade model

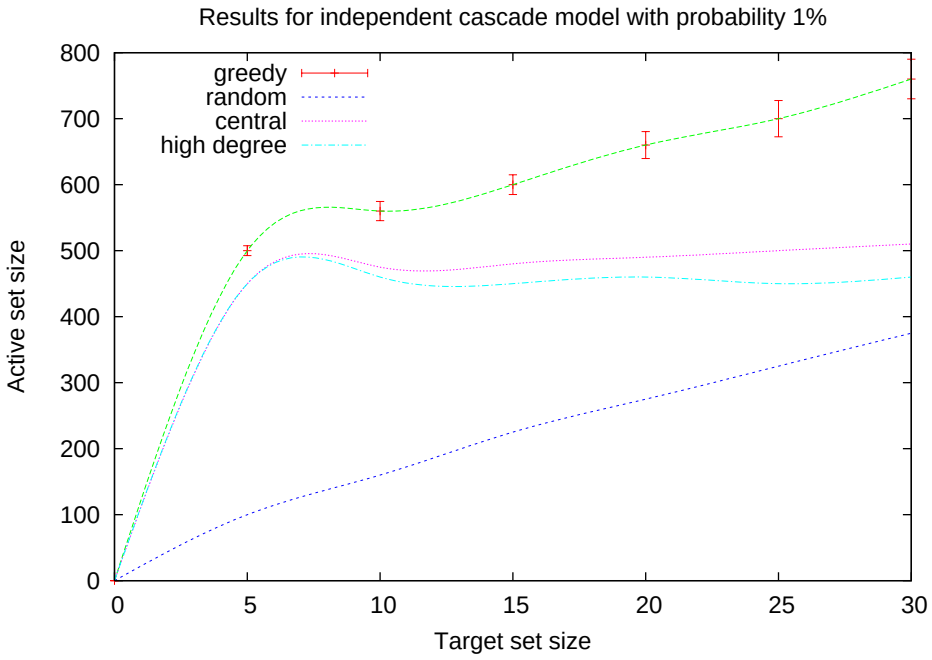

Figure 6: Independent cascade model with probability 1%

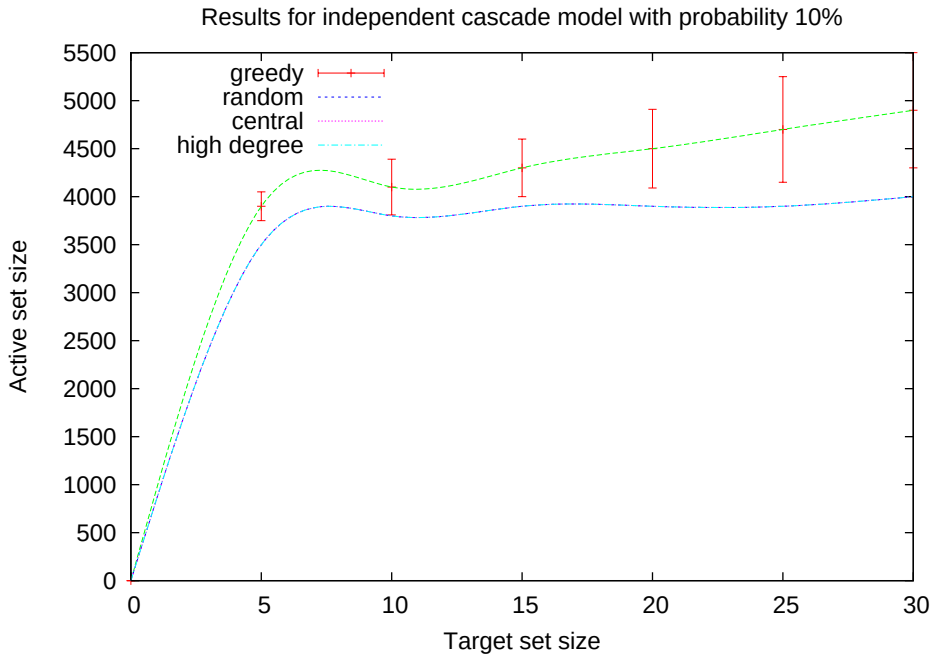

Figure 7: Independent cascade model with probability 10%

6 Additional Experiments

We carried out additional experiments with a different collaboration network. The dataset and description for this collaboration network is available at [1].

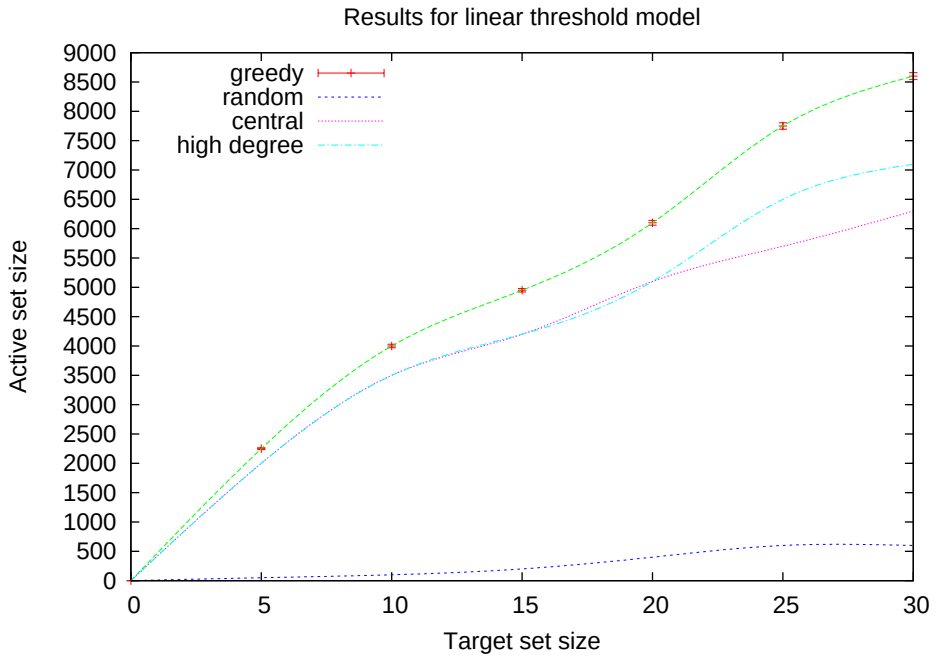

Figure 8: Results for the linear threshold model

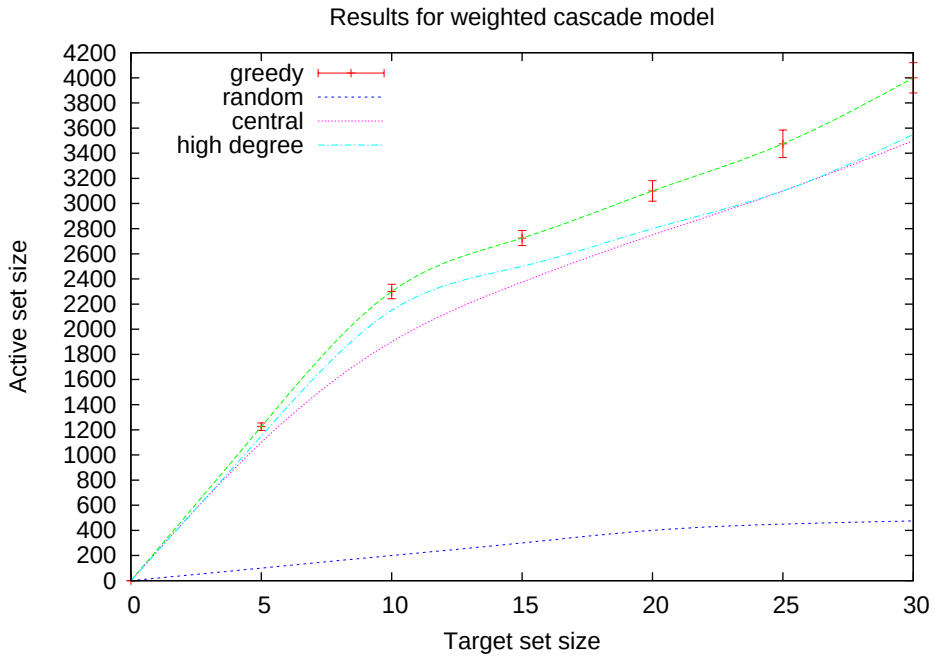

Figure 9: Results for the weighted cascade model

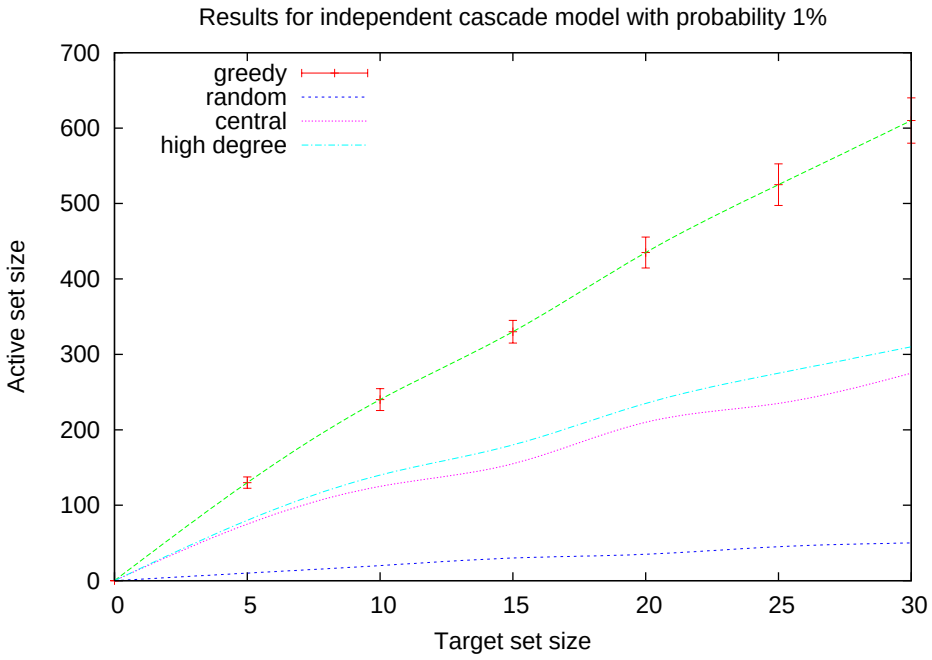

Figure 10: Independent cascade model with probability 1%

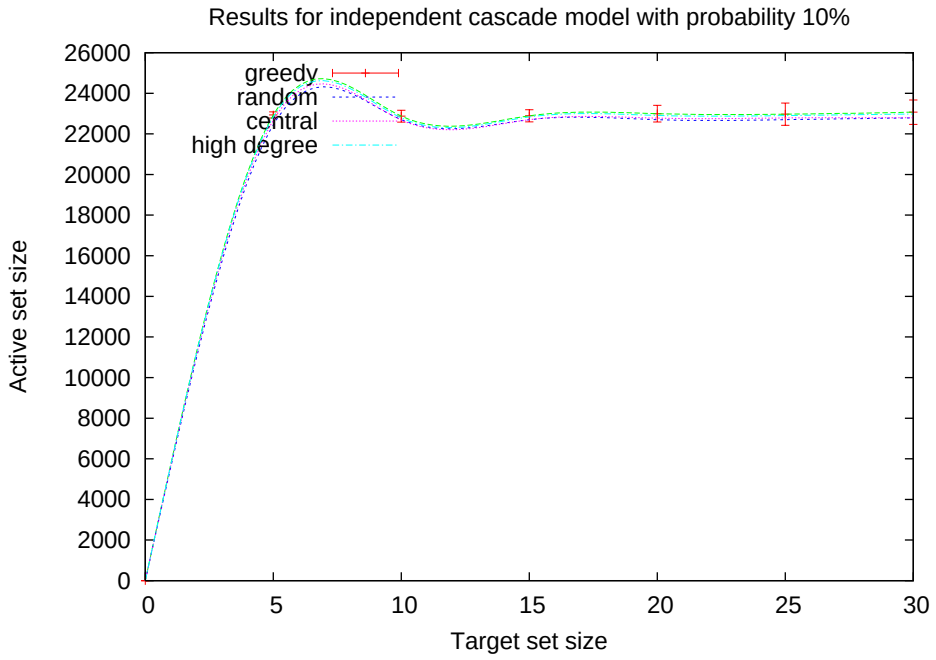

Figure 11: Independent cascade model with probability 10%

7 Conclusion

Our results support the conclusions of the paper that their proposed greedy algorithm out performs existing heuristics. However, the algorithmic complexity of their algorithm is still high enough to demand further changes in their mechanism. We propose that some sampling mechanism be added to their algorithm so that it becomes more useful for large-scale datasets.

8 References

References

[1] High Energy Physics - Citation network. `http://snap.stanford.edu/data/cit-HepPh.html`.

[2] High Energy Physics - Phenomenology collaboration network. `http://archive.ics.uci.edu/ml/datasets/SECOM`.

[3] MATLAB Boost Graph Library. `http://dgleich.github.com/matlab-bgl`.

[4] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.