**Reaction Paper # 5: Influence Maximization**

Mohammad Elkholy  
Computer Science and Engineering  
 The American University in Cairo  
 New Cairo, Cairo, Egypt  
moelkholy@aucegypt.edu

**SUMMARY**

**Efficient Influence Maximization in Social Networks**

The authors in this paper aimed to improve the running time of the original greedy algorithms used for identifying the best initial k set of nodes that maximize influence, as well as provide a new heuristic for choosing nodes that is much faster than the greedy algorithm but provides similar performance. The authors also try to mix the original greedy algorithms with the CELF optimization to achieve a 700x speedu in running time. The CELF optimization makes use of the submodular nature of influence maximization to reduce the number of evaluations needed, this improvement however can still take up to a frew hours on a graph with number of nodes ~ , so it is not very scalable for larger networks.

The original algorithm has a runtime of O(knRm), where k is the number of nodes in the starting set S, n is the number of nodes in the network, m is the number of edges in the network and R is the number of simulations conducted per node. The CELF optimization uses submodularity property in order to speed up the algorithm by noticing that the incremental influence spread decreases as S increases in size, we don’t need to reevaluate spread values for a large number of the nodes as they were already calculated in the previous round, thus eliminating a lot of computation.

The authors then suggest improvements to the different cascade models already studied beforehand. They first start by the Independent Cascade Model, suggesting to use BFS or DFS to create a linear scan of the subgraph G’ created by eliminating the edges from the seed set S of nodes that will not be propagated through. This creates the set of Vertices reachable from S to G’. Doing a linear scan generates the reachable nodes from S as well as the reachable nodes from any arbitrary node v in V\S, thus the additional number of elements that v contributes to reaching is the the number of nodes reachable by v if v is not reachable by S, and 0 if v is reachable by S. This method provides an order O(n) speedup on the original algorithm. They then suggest a new algorith that mixes both their algorithm and CELF optimization, by computing the influence spread estimates for all vertices in the first round then using the CELF optimization for later iterations, this is done for other models as well.

They also provide an improvement on the weighted cascade model, where a node u is activated with probability 1/dv where dv is the degree of node v. This leads to an imbalance in the probability of activation, as u activating v is not the same as v activating u. This led them to create a directed graph where each edge between u and v in undirected graph is replaced by one from u to v and one from v to u. Then, for each edge uv, they remove it from the graph with probability 1-1/dv. We can then find the reachable nodes from set S and each vertex v not in S from the generated graphs. This however is a problem as the linear scan now will take much longer as it is directed. Thus, they resort to Cohen’s randomized algorithm which uses an unbiased estimator of the exponential distribution to estimate the spread after having generated a directed acyclic graph by collapsing strongly connected components into one node with a weight. They incorporate this in their algorithm, needing T iterations to estimate which leads to a speedup over original algorithm if T < n, which is always the case as T = 5 leads to very good results.

The authors then provide a new heuristic by which we could pick the starting set S of nodes. They call it a degree discount. This measure decreases the bias created by the extra edges from the Seed set to the new node to be activated. This is known as a single discount. However, for Independent Cascade with very small p, influence spread is not very sensitive to removal of edges with probability 1-p, which has been reported for up to p = 0.1. Thus further discount helps reduce bias more without affecting the result. They decided to ignore multi-hop neighbors for this calculation for small p. They made assumptions that dv = O(1/p) and tv = o(1/p) which they believed is reasonable for their experimental networks and other social networks. This algorithm has a running time of O(klogn+m) which is much faster than the greedy algorithm and their own modified ones.

Their experiments supported their claim that their modified algorithm as well as their discount degree heuristic provides very similar results to the original algorithm where the degreediscount heuristic was vastly superior running time wise. Their new improved algorithm exchanged blows with the CELF algorithm while their mixed ones were always faster. They concluded by suggesting that degree discount performs better in influence spread, and that the mixed algorithms should be used when running time is not a concern but influence spreed guarantees are essential.

**On the Submodularity of Influence in Social Networks**

The author in this paper provides a prove that local submodularity is preserved globally for diffusion processes. This has already been verified for the independent cascade model and linear threshold model but was yet to be proven more generally. They are set to prove that if a function satisfies submodularity on a general set S, then it is also submodular on subsets of S.

To prove this, they introduce two ideas:

* antisense coupling
* piecemeal growth
* need-to-know representation

Piecemeal growth describes the division of the diffusion process into several stages or steps, each with their specific distribution in such a way that preserves said distribution.

Need to know representation deals with handling the thresholds of node activation between stages such that it is preserved across the different stages. We do not need to know about the thresholds at each step in the process is whether the increase was enough to hit the threshold. It turns out that this maintains the distributions of each step of the process.

The last step was to use coupling to prove global submodularity. They introduce 4 processes, each with 3 stages of some combinations of distributions . They then construct needed results using submodularity definition and set properties, proving that submodularity is maintained given a monotone submodular function on a subset S.

**COMMENTS**

**Efficient Influence Maximization in Social Networks**

This paper was very interesting. Previous papers tackling this problem suggested to move away from heuristics due to their poor performance, but they failed to provide greedy algorithms that can run efficiently that produced good results. This paper however managed to provide running time improvements on those greedy algorithms as well as provide a heuristic for picking the seet set in a very efficient manner. My only complaint is their assumptions about dv and tv in social networks. These assumptions are necessary for their proof and as such their applicability on other social networks comes under question.

**On the Submodularity of Influence in Social Networks**

Highly theoretical paper that my lack of knowledge of probability and measure theory renders me unable to make any comments. I liked the idea of coupling so I will be reading more about that.

**REFERENCES**

[1] Chen, Wei & Wang, Yajun & Yang, Siyu. (2009). Efficient influence maximization in social networks. 199-208. 199-208. 10.1145/1557019.1557047.

[2] Elchanan Mossel and Sebastien Roch. 2007. On the submodularity of influence in social networks. In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (STOC '07). Association for Computing Machinery, New York, NY, USA, 128–134. https://doi.org/10.1145/1250790.1250811