# Mining Influential Nodes in Social Networks to Maximize the Spread of Influence with Compact Communities

Project Proposal for CS 7930 - Spring 2016

Honey Varghese, Jonathan Arndt, Curtis Larsen

## Related Papers

Our project idea began with a pair of papers related to social network structure and maximizing influence spread in social networks.

The first paper is "*Finding compact communities in large graphs*" by Cruesefond, et al. [1]. A graph community is a collection of nodes that are more highly connected than the graph as a whole. Several varying definitions of communities are used in the literature and algorithms for finding them, have been studied. Most algorithms have time complexities that make them unsuitable for use with large networks. Communities in social networks can be used for targeted marketing of products and ideas. The LexDFS based hierarchical clustering algorithm of this paper identifies compact communities in graphs with efficient O(n log(n)) runtime complexity, making it possible to use on large social networks. LexDFS is a modified Depth First Search that tracks the generation of the search that each vertex was visited, and prefers recently discovered vertices for expansion. This causes clustered vertices to be visited in close generations. Finally, their algorithm finds compact clusters of vertices by weighting edges using the difference between generation numbers of the vertices participating in the edge. Then use these weights to find clusters of vertices that have small edge weights between them. This paper does not explore the idea of applying the identified communities for the selection of seed nodes in the maximum influence spread problem.

The second paper is "*Spanning graph for maximizing the influence spread in Social Networks*" by GAYE et al. [2]. The influence spread problem is to maximize the spread of a message through a network to activate the largest number of nodes after a fixed time. This problem is known to be NP-Hard. The message spreader selects some set, of size k, of initial seed nodes. The seed nodes spread the message to their neighbors with some probability using their influence in the network. Then, all activated nodes spread the message via their connections in the next time cycle. This repeats until the spread of the message stops, and the activation of nodes stops. The number of activated nodes at this time is the size of the influence spread.

There is a cost associated with selecting seed nodes in the network, so a limited number of seeds should be selected. Selecting the fixed number of seeds to maximize the spread has been studied with many different algorithms identified, including High Degree Heuristic Model, Degree Discount Heuristic Model, Centrality Measure, Hill Climbing and others. This paper designs a pair of algorithms, one for undirected graphs and one for directed graphs, to remove cycles from a network before applying the Centrality Measure algorithm to select the seeds. They studied the effectiveness of their method by running simulations on two network data sets and comparing with the Centrality Measure seed selection and the modified Centrality Measure seed selection. They found that influence propagation was moderately improved with their method.

After additional reading we found several papers of interest, including the detailed study of the Maximum Influence Spread problem and the Hill Climbing algorithm by Kemp et al. [4], including a methodical approach using experiments to empirically compare seed selection algorithms. A work by Wang et al. [3] develops and applies an efficient community based greedy algorithm for finding influential nodes in a network.

# Project Proposal

We propose an algorithm that identifies the top k influential nodes in a social network using the compact community detection algorithm from Creusefond et al., as the community identification portion of the greedy algorithm of Wang et al.The algorithm will identify k compact clusters and will select one potential candidate from each cluster to start the propagation of information. The highest degree heuristic will be used to select the seed node within a cluster. Our project will compare this algorithm to the greedy hill climbing, CELF++, high degree, centrality, and random seed selection algorithms. The comparison will include empirical results for the size of influence spread and the run time complexity of each algorithm. Our goal is to prove that this algorithm is of sufficient speed to be used feasibly on large social networks to achieve large influence spread. This would then be valuable to marketers desiring to effectively spread their message through a small set of influential nodes in the network.

## Research Problem

Utilize efficient algorithms to maximize the influence spread in large social networks, using k seed nodes.

## Research Question

Would seed nodes selected using compact cluster detection algorithm help in maximizing the influence spread while maintaining scalability? Would it improve scalability while maintaining influence spread?

## Question Interest

Marketers of goods and services desire to spread their message to large groups of people in social networks. An effective method to identify a small group of people to spread their message will maximize their gain while minimizing their cost. For example, identifying a small number of key influential people and giving promotional merchandise to them will spawn loyalty and privilege promoting a cascading effect to their connections. Mass messages will be perceived as spam whereas messaging from friends are acceptable and more likely to be received positively. In fact, it is known to influence purchasing decisions to the extent that 2/3 of the United States economy is driven by those kind of personal recommendations Iribarren et al. [9].

## Existing Research

Various seed selection models have been proposed and studied. This is a short list of the most effective and currently popular:

a) Hill-Climbing Algorithm : It is not efficient, but guarantees the quality of seed nodes selection [4]
b) CELF Algorithm (and CELF++) [6]: It exploits the property of submodularity of social networks to improve hill-climbing
c) Heuristic Algorithm[4]: Degree-Centered heuristic and Distance-Centered heuristic: Commonly used in the literature as estimates of a node's influence. They are capable of producing acceptable results in much less time than Hill-Climbing.

## Project Plan

We plan to develop our seed selection algorithm and then compare it to other standard seed selection algorithms. We will use empirical experiments to measure the expected influence spread and algorithm run time efficiency. For analysis the network datasets from Stanford Large Network Dataset Collection [7], and the Stanford Network Analysis Platform (SNAP) [8] will be utilized. Each seed selection algorithm will be used along with the Independent Cascade influence model, with uniform probabilities on the edges. The outcome of the experiments will be a set of graphs showing the active set size achieved by each of the seed selection algorithms, and respective run times.

## Solution Evaluation

The solution will be evaluated by size of the influence set it produces and run time compared to the baseline algorithms for the same values of k.  Larger influence set size or faster runtime will be considered a viable result.

## Timeline and Milestones

3/5 -  SNAP trials (data and libraries at small scale)
3/19 - Create Seed selection algorithms
3/19 - Influence spread simulation (IC and/or LT)
3/26- Data collection/selection from SNAP
4/2  -  Preliminary results
4/9  - Algorithm improvements, design and completion
4/16 -Finish  Simulation and results comparison
4/23 - Report

# Project

Link for Snap - python: http://snap.stanford.edu/snappy/index.html

# Bibliography

1- Jean Creusefond, Thomas Largillier, and Sylvain Peyronnet. 2015. Finding compact communities in large graphs. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15), Jian Pei, Fabrizio Silvestri, and Jie Tang (Eds.). ACM, New York, NY, USA, 1457-1464. DOI=http://dx.doi.org/10.1145/2808797.2808868

2- Ibrahima GAYE, Gervais Mendy, Samuel Ouya, and Diaraf Seck. 2015. Spanning graph for maximizing the influence spread in Social Networks. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15), Jian Pei, Fabrizio Silvestri, and Jie Tang (Eds.). ACM, New York, NY, USA, 1389-1394. DOI=http://dx.doi.org/10.1145/2808797.2809309

3- Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. 2010. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 1039-1048. DOI=http://dx.doi.org/10.1145/1835804.1835935

4- Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Theory OF Computing 11.4 (2015): 105-147. (137-146 for baseline greedy algorithm)

5- Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01). ACM, New York, NY, USA, 57-66. DOI=http://dx.doi.org/10.1145/502512.502525

6- Goyal, Amit, Wei Lu, and Laks VS Lakshmanan. "Celf++: optimizing the greedy algorithm for influence maximization in social networks." Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 47-48.

7- Jure Leskovec and Andrej Krevl. 2014. Stanford Large Network Dataset Collection. http://snap.stanford.edu/data

8- Jure Leskovec and Rok Sosic. 2014. Snap.py: SNAP for Python, a general purpose network analysis and graph mining tool in Python. http://snap.stanford.edu/snappy

9- José Luis Iribarren and Esteban Moro. 2011. Branching dynamics of viral information spreading, Phys. Rev. E 84, 046116. http://markov.uc3m.es/~emoro/ps/PRE2011-2.pdf

10- Granovetter M: Threshold Models of Collective Behavior. The American Journal of Sociology 1978, 83(6):23.

11- Schelling TC: Miromotives and Macrobehavior. 1978.