**Finding “Seed” Users with Maximum Influence in Their Social Circles**

*Mark Weaver, Maxwell Yi*

[mrw93@txstate.edu](mailto:mrw93@txstate.edu)

[mky7@txstae.edu](mailto:mky7@txstae.edu)

***Abstract --* Social media, since its creation, has allowed us to remain connected with the people we interact with the most. There exists algorithms that even aid in filtering what we want or do not want to see on our social feeds; however, these *social circles* become increasingly complex, and for marketers and businesses, it becomes time consuming and costly. Therefore, we develop an algorithm specifically designed to locate these social circles by node clustering in order to find the most effective *seed* users, propagating information targeted to such users and allowing us to maximize user-awareness to a product.**

1. Introduction

The internet, as we know it today, contains more information than we can fathom. Social networks such as Twitter and Facebook allow us to remain connected with people we care about and even people we idolize. Everyday, an average person is exposed to endless streams of information by close friends, relatives, celebrities, and more, so much that we can consider this an ‘information overload’ [McAuley and Leskovec et al. 2014]. Typically, we tend to organize our social interactions manually, but with exponentially advancing technology, social media is able to sort social circles and information streams based on what we click, what we like, what we comment on, what we watch, etc. Personalized search engines takes advantage of such sorting methods, reordering search results based off of our previous searches and clicks. So we ask, how might we use such algorithms to identify and target these circles for products and business demographics.

We study Julian McAuley and Jure Leskovec’s research, *Discovering Social Circles in Ego Networks*, to further analyze how social circles are formed and the algorithms that define them. Furthermore, we analyze the information within the defined social circles to maximize user awareness and discover the most effective *seed* users for a product. We will describe this as *node clustering*, a network of connections between a user and their friends/interactions.

1. Problem
2. *Description*

Consider the problem of a new marketing strategist looking to exploit an existing social network, to identify which users would be the most effective seed users, and to maximize user awareness of a product by propagating that information to targeted social circles and groups. To do this we need to define a machine learning task that automatically  identifies users’ social circles.

We pose this problem as a node clustering and optimization problem on a user’s network, a network of connections between their friends. By studying past research on social circles, we will be able to define an algorithm that allows marketing strategists to push a product to a given demographic based on a circle’s information. Such node clusters will carry information regarding users’ choices in their social stream preference. What they view, like, comment on, and share will allow for the algorithm to further define the target audience; however, social circles contain a vast amount of information that share similar qualities between themselves.

*B. Theory*

As previously stated, an individual’s social network is vast. One can control their own network’s organization into social circles based on relationships and interests. We describe the social circles as *ego networks* in which the owner is the *ego* and surrounding nodes are its *alters* [McAuley and Leskovec et al. 2014].

We visualize an ego network with nodes, and we measure node connections with two metrics: *degrees*--the direction connections between nodes--and *shortest path*--the amount of hops needed to traverse from one node to the next*.* Almost every node is connected, and in each social circle, there exists a node that we describe as the most influential, that is, the node with the greatest *closeness centrality*. We describe the closeness centrality as the summation of the length of the shortest paths between the nodes and all other nodes in the data set.

There exists another centrality known as *degree centrality* (the node with the highest degree); however, this does not define a node’s influence. A node’s influence relies on its degree status and its shortest path values.

By finding the most influential node, we broaden our search to find the top 3 central nodes. These nodes will have the most influence on a social circle and allow us to maximize the outreach of a product or marketing campaign.

It is important to note that the ego is not the most influential. While the ego may have the highest degree centrality, it will not have influence due to the varying distances between its alters. Therefore, we may exclude the ego from our calculations.

*C. Application & Methods Used*

Below are two datasets that display metadata for two popular social media websites. Each dataset contains a set of users and all of the circles, edges, ego features, features, and feature names associated with each user. Figure. 1 contains node information from the website “Twitter.” Additionally, Figure. 2 contains node information from the website “Facebook.”

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 107614 |
| Edges | 13673453 |

*Figure 1. Twitter Dataset Statistics*

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 4039 |
| Edges | 88234 |

*Figure 2. Facebook Dataset Statistics*

Using the information listed in figures 1 & 2, we can define a model that can be applied arbitrarily. There are several ways in which we can define them. One of which is node clustering.

In order to use node clustering to find seed users with maximum influence, we observe and measure a node’s degree. Taking a look at our facebook data set we can use this configuration to display an ego network that shows degree centrality. When analyzing our facebook dataset we chose to view the ego network of node 3437 at random. The visualization of node 3437’s ego network is shown below in Figure 3.

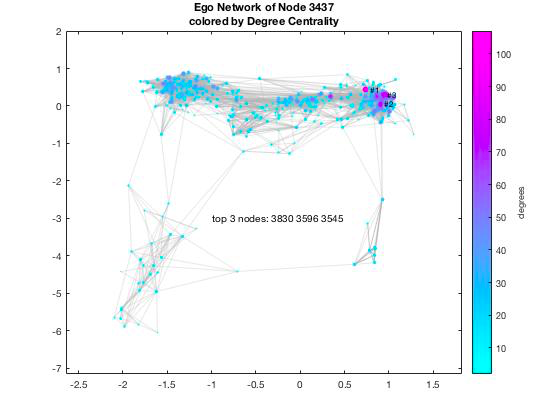
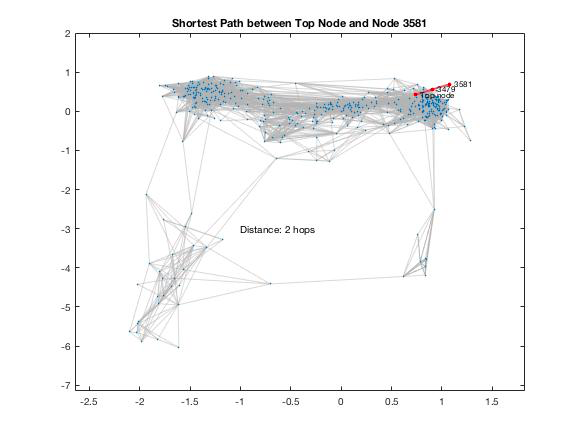
*Figure 3. Ego Network of Node 3437 Colored by Degree Centrality*.

Figure 3 shows use the top 3 nodes of centrality (excluding the ego node) according to degree centrality, but as we stated before, degree centrality does not provide the most influential nodes because of the varying distances between its alters. Thus, we continue by calculating the distances between nodes by using the method of shortest path. In Figure. 4 below we see the same ego network but with the shortest path between the top node and node 3581.



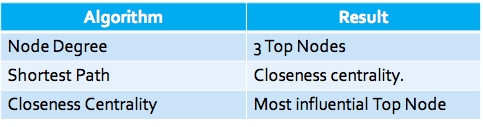
*Figure 4. Shortest Path Between Top Node and Node 3581*

If we know the distance of the shortest path between the top node and any other node, then we can use the distances measured by shortest paths to compute closeness centrality. Below is the equation to compute the closeness centrality

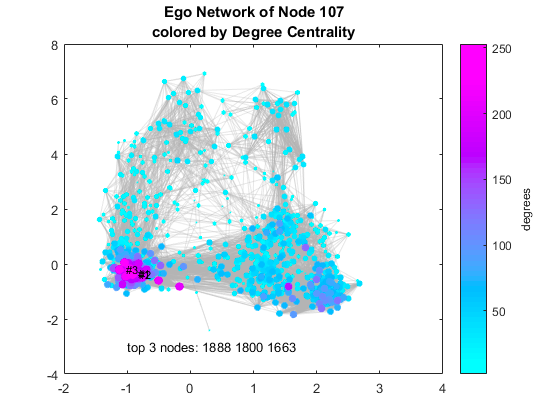
where *C(x)* is the closeness centrality, *N* is the number of nodes in the graph, and *d(y,x)* is the distance between the vertices *y* and *x*.

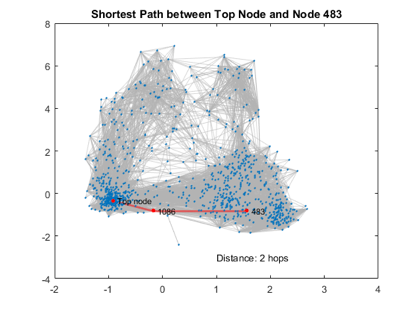
III. Results

After calculation, we find the nodes with high closeness centrality scores since these will have the most influence in the ego network. In our data, we find the top three nodes with the greatest degrees excluding the ego. Figure 5 describes the relation between node degree, shortest path, and closeness centrality.

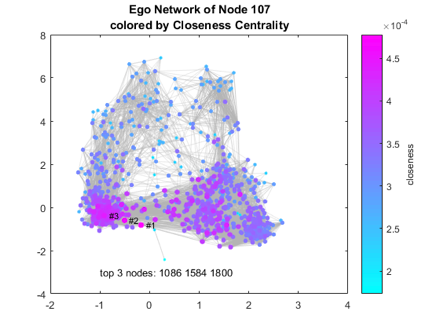


*Figure 5. Algorithmic relationships between node definitions*

*Figure 6. Ego Networks Colored by Degree Centrality*



*Figure 7. Shortest Path example*

*Figure 8. Ego Networks colored by Closeness Centrality*

The node degree and shortest path have a direct correlation in finding the top nodes of the dataset.

IV. Conclusion

So far, we have analyzed the datasets to the best of our ability in order to create a comprehensible summary of our algorithm. While the models are still in the process of derivation, we have found algorithms that will allow us to find social circles among users using popular social media. In having this, we are able to discover circles with basic information regarding user preference on social media, allowing us to readily develop a target audience for a given product.

We have observed that the more popular a user is (by the node degree), the more influence they hold. Although, to more accurately predict a user’s influence, we also need to calculate the centrality of the node because the more central a node is, the closer it is to all other nodes. Therefore the nodes (users) with the highest degree and highest centrality are the most influential nodes (users) in dataset; their information will propagate fastest through their ego network, thus making them the perfect seed user.

We observe that the top three nodes

V. Future Work

Further developing the algorithm may lead us to a greater predictability outcome for future ego networks as our data is not absolutely accurate; simultaneously, ego networks are constantly growing at an exponential pace which makes it more difficult to accurately target the most influential users. Because of such rapid growth, our algorithm will have to update frequently. It is undetermined at which rate updates will occur, but it would minimize error in targeting the most influential nodes.

Furthermore, it would be ideal for the algorithm to autonomously learn different node patterns and characteristics. More specifically, the algorithm should begin to learn information from a user’s public profile and make predictions for private profiles based off of the given information. This will allow the algorithm to find more specific social circles based on interests. Unfortunately, this may lead to some inaccuracy within the ego network and decrease our datasets, but product advertisement may be pushed more efficiently and expand its audience. In doing so, marketing strategies may become autonomous; campaigns may be pushed to ‘special-interest’ social circles in addition to those more general.

VI. Contributions

**References**

*Barbieri, Nicola, Bonchi, Francesco, and Manco, Giuseppe. Topic-aware social influence propagation models. Knowledge and information systems, 37(3), pages 555–584, 2013.*

*David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social ´ network. In KDD, pages 137–146. ACM, 2003.*

*Goyal, Amit, Bonchi, Francesco, and Lakshmanan, Laks VS. A data-based approach to social influence maximization. Proceedings of the VLDB Endowment, 5(1), pages 73–84, 2011a.*

*Hongyuan Zha, Isabel Valera, Le Song, Manuel Gomez-Rodriguez, Mehrdad Farajtabar, and Nan Du. Shaping Social Activity by Incentivizing Users. In Advances in Neural Information Processing Systems 27, pages 2474-2482, Curran Associates, Inc., 2014.*

*Shure, L. (2016, February 3). Visualizing Facebook Networks with MATLAB. Retrieved April 4, 2018, from https://blogs.mathworks.com/loren/2016/02/03/visualizing-facebook-networks-with-matlab/*

*Vaswani, S., Kveton, B., Wen, Z., Ghavamzadeh, M., Lakshmanan, L.V.S. & Schmidt, M.. Model-Independent Online Learning for Influence Maximization. Proceedings of the 34th International Conference on Machine Learning, in PMLR 70, pages 3530-3539, 2017.*

*Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In KDD, pages 199–208. ACM, 2009.*