**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. The vastness of social media increases the difficulty to pinpoint the most influential users of an individual network. Additionally, a user’s access to streams of information from their network is variant upon their relationships with other users and other users’ choice to publicly release their information streams. 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.

In order to target node clusters by interest, we must begin by finding the users with the most influence in a given network. Additionally, an algorithm must be developed in which we may find the most influential users autonomously. From there, we can further distinguish the interests in a social network and target more specific node clusters with greater influence over each sub-network.

*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 in the context of the spread of information, 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 | 81,306 |
| Edges | 1,768,149 |
| Ego Networks | 1,000 |
| Diameter (longest shortest path) | 7 |

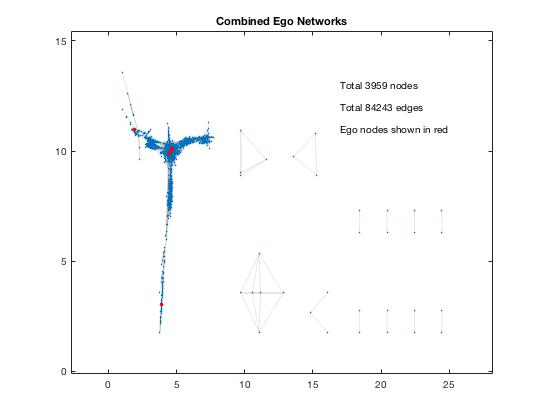
*Figure 1. Twitter Metadata Dataset Statistics*

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 4,039 |
| Edges | 88,234 |
| Ego Networks | 10 |
| Diameter (longest shortest path) | 8 |

*Figure 2. Facebook Metadata Dataset Statistics*

From our Facebook dataset we were given the profile and network data from 10 ego-networks, consisting of 193 circles and 4,039 users. Circle information was gathered by a survey of ten users, who were asked to manually identify all of the circles to which their friends belonged.

Using this information, we can define a model that can be applied arbitrarily. There are several ways in which we can define this data. One of which is node clustering. Our first model we decided to create is displayed in Figure 3 where we describe an undirected graph of the combined ego network of all ten different ego nodes within our dataset to visualize our given portion of the social network as a whole.



*Figure 3. Combined Ego Networks*

Although, examining our social network as a whole did not show any obvious patterns or correlations that might help us determine optimal seed users; in order to use node clustering to find seed users with maximum influence, we observe and measure a node’s degree. Taking another look at our facebook data set, we can use this configuration to display an ego network that shows degree centrality. This way we can see the popularity of nodes within an ego network. Decreasing the node pool, also, will allow us to better target influential nodes within a given ego network. 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 4.

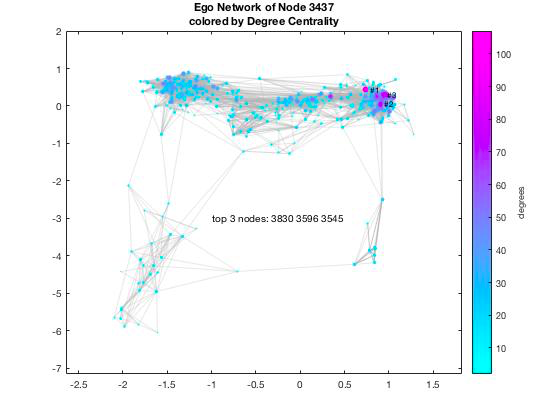
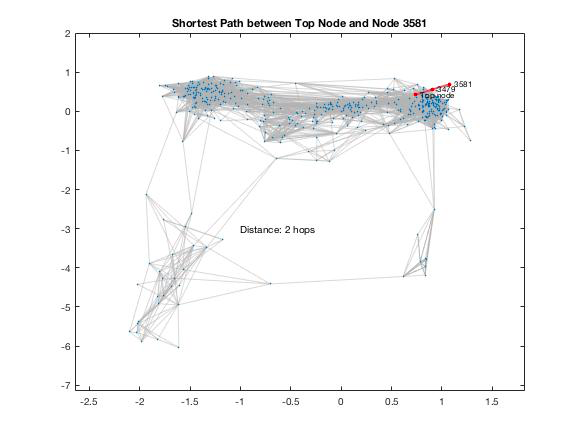
*Figure 4. Ego Network of Node 3437 Colored by Degree Centrality*.

Figure 4 describes 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 in the context of access of information such as, the flow of information over a social network due to the varying distances between the node and all of its alters. Thus, we continue by calculating the distances between nodes by using the method of shortest path. In Figure. 5 below we see the same ego network but with the shortest path between the top node and node 3581.



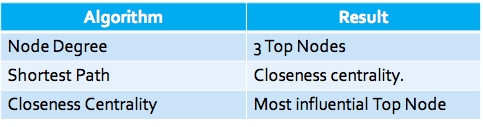
*Figure 5. Shortest Path Between Top Node and Node 3581*

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,

and Normalized Closeness Centrality,

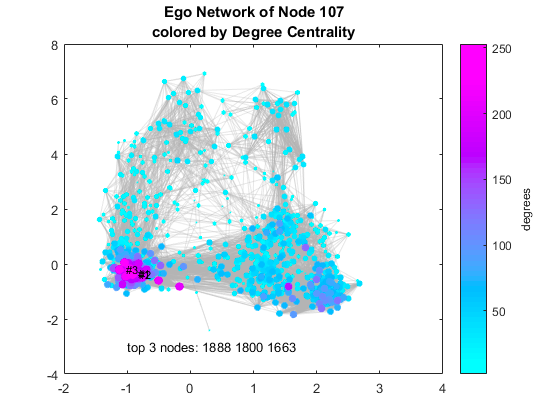
where *C(x)* is the closeness centrality, *N* is the number of nodes in the graph, and *d(y,x)* is the distance between the vertices *y* and *x*. Note that, because the equation is a reciprocal of farness, we search for closeness scores of lower values; the lower the value, the more central a node will be. We can then rank all of the centrality scores of each node to find the most central nodes.

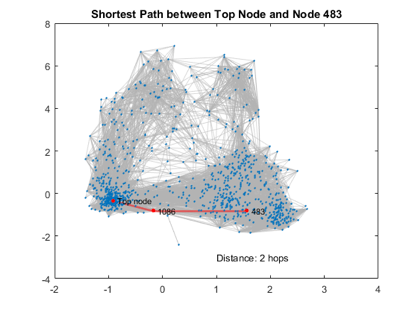
III. Results

After the implementation of all of our algorithms and calculations, we find the nodes with closeness scores of high centrality 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 6 describes a quick break down of our findings about the relation between node degree, shortest path, and closeness centrality.

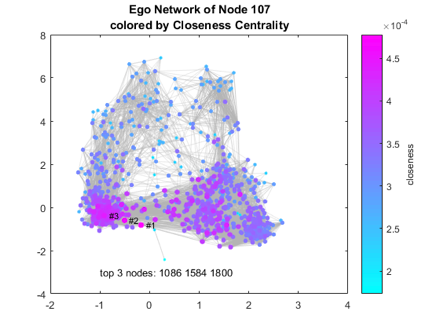
*Figure 6. Algorithmic relationships between node definitions*

We continue our comparison between different ego networks, observing each degree centrality to recognize similarity patterns. Figure 7 demonstrates the ego network of node 107; however, we score each node based on their degree centrality, depicted by color.

*Figure 7. Ego Networks Colored by Degree Centrality*

It can be said that both ego networks contain approximately 2-3 concentrated node clusters with similar shortest path distances, and our top influential nodes can be found within the clusters of the highest degree centralities. In order to further understand the importance of the shortest path, we find the top node and follow the path to a random node. Figure 8 describes the shortness path between the top node and node 483.

*Figure 8. Shortest Path example*

*Figure 9. Ego Networks colored by Closeness Centrality*

In Figure 9, we further analyze the ego network of node 107 by scoring each node based on closeness centrality depicted by color. Note that the node degree and shortest path have a slight correlation in finding the top nodes of the dataset. Closeness centrality and degree centrality share a common characteristic; that is, the most influential nodes are found in the node clusters with the highest of each. Notice, however, the spread of closeness centrality in figure 8. If the top influential nodes are characterized by their closeness centrality, then it would, arguably, make little difference between targeting nodes within the alternate node cluster.

The degree centrality will allow for us to maximize the outreach of any campaign pushed to the influential target nodes; closeness centrality defines our top three, but degree centrality allows for a greater chance of distribution throughout the ego network. This demonstrates the importance of finding degree centrality in tandem with closeness.

However, that does not qualify that degree centrality holds equal importance when finding the most influential nodes. Node degree centrality describes a user’s amount of connections in their individual social circle; node closeness centrality allows a user more direct access to information within their social circle. For example, one user on Facebook is able to have 4,000 friends maximum on their personal profile. User A’s profile contains 3,500 friends and User B has 800. While User A has an incredible amount of connections with varying people, User B may contain closer relationships with his limited friend list. For this reason, User B will have a greater influence than User A. In relation to our data analytics, User B has a stronger influence than User A despite User A’s higher degree centrality score. In such a case, a marketer will have more success raising a product’s awareness through an ego network by targeting User B.

IV. Conclusion

1. *Conclusion*

So far, we have analyzed the datasets to the best of our ability in order to create a comprehensible summary of our algorithm and purpose. 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 when it comes to context of sheer number of contacts, like audience size or number of supports. Although, to more accurately predict a user’s influence in the context of opinion forming or direct access and spread of information, 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 the dataset; their information will propagate fastest through their ego network, thus making them the perfect seed user.

The ego networks found from our datasets reveal that there are multiple nodes that seem to have an influence as valuable as the top three. Additionally, there exists other nodes with connections that could be vital to the expanding a product’s exposure in an ego network. Therefore, we may increase the degree value to find, perhaps, the top 5 or top 10 influential nodes of future networks. Additionally, other node clusters exist that could benefit our overall purpose in exploiting a network to a product.

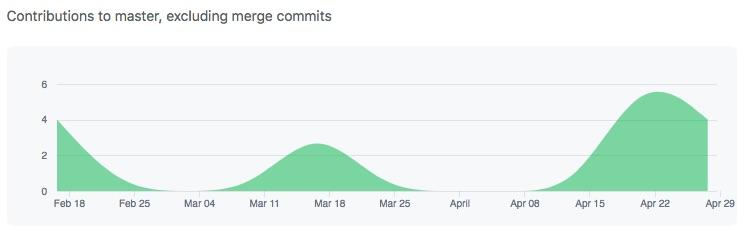
We are only given basic information from the users within our given datasets; in each user, there is additional information in terms of their interests, their relationships, and their access to other users’ information (basic and extraneous). We found that our datasets contain two to three node clusters that contain, what seems to be, users with influence among their given cluster. It is questionable whether or not the interests in the alternate clusters differ from the cluster with the most influential nodes. Further analysis indicates that their closeness centrality degree is approximately the same as that of the primary cluster (the cluster containing the most influential nodes). We find the possibility that the alternate node cluster might have significant influencers for users of different interest, allowing for marketing strategists to, more effectively, propagate their campaigns and products.

It can also be said that different user networks overlap with one another. We may continually pick nodes at random and analyze their ego networks to find that some nodes are shared between them. Taking this into account, we conclude that influence is not only defined by the closeness centrality of a node, but also defined by a node’s shared connections among overlapping networks.

*B. Contributions*

Both contributors definitely played an integral role in the process of this project coming to fruition. This is seen below in Figure 10, Figure 11, and Figure 12 from the github repository, both participants made timely and necessary contributions.

Each figure may not reflect the exact amount of the workload that was shared between each collaborator, which makes an explanation of the shared workload appropriate.



*Figure 10. Contributions to Project*



*Figure 11. Contributions made by Mark Weaver*



*Figure 12. Contributions made by Maxwell Yi*

The The original topic (real world problem) that was decided upon and the datasets that were used were discovered by Mark Weaver. As was the majority of the research that was found and completed. But, the algorithms used to visualize and analyze our data were discovered and written by Maxwell Yi along with most of the powerpoint presentation.

The project proposal, intermediate project report, and the final project report were all written in tandem with both collaborators contributing equally to accomplish each of the tasks required.

*C. 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.

In our datasets, we observe that the top three nodes are centralized in one area of the network; while these are the most influential nodes, will this provide an efficient outreach to the rest of the network? Ideally, a derivation of our algorithm might allow us to target other influential nodes in other clusters of the ego network. As we observed earlier, there exists multiple clusters within an ego network. Whether or not the top nodes are truly influential will be determined after a product is pushed.

Additionally, as previously discussed, overlapping networks may increase a node’s influence among multiple networks. We would like to study and develop an algorithm to find nodes that are shared between multiple ego networks and discover their centralities overall. A combination between shared networks and high centralities would significantly increase a node’s influence over a large ego network; moreso, a combination between shared networks and high centralities would allow businesses and marketing strategists to dramatically increase their product’s awareness and to propagate the networks to the fullest extent.

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.

**Resources**

*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.*