**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*

A social circle has its own respective ego network which we visualize with nodes; 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.

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.

*C. Technical Proof*

asdfasdf

*D. Application*

asdfasdfasd

*E. Methods used*

In order to use node clustering to find “seed” users with maximum influence, we looked at degrees as a metric to evaluate nodes, because the more friends a node has, the better connected it is.

In order to get a more accurate idea of which node in each ego network is the most influential, we decided to use another common metric known as shortest path. While degrees measure direct connections only, shortest paths consider how many hops at minimum you need to make to traverse from one node to another.

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. Which is calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph.

equation for closeness centrality

C(x) = 1/∑yd(y,x)

1. Data 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. Fig. 2 contains node information from the website “Twitter.” Additionally, Fig. 3 contains node information from the website “Facebook.”

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 107614 |
| Edges | 13673453 |
| Diameter (longest shortest path) | 6 |

*Figure 2. Twitter Dataset Statistics*

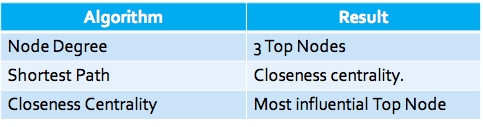
|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 4039 |
| Edges | 88234 |
| Diameter (longest shortest path) | 8 |

*Figure 3. Facebook Dataset Statistics*

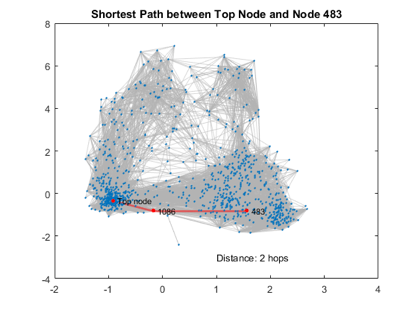
Using the information listed in figures 2 & 3, we can define a model that can be applied arbitrarily. There are several ways in which we can define them. Our derivations will come from simple Linear Regression and Cost Function models. The primary functions to describe the social circles will be following an “unsupervised algorithm to optimize the latent variables and the profile similarity parameters to best explain the observed network data” [McAuley and Leskovec et al. 2014].

IV. Results

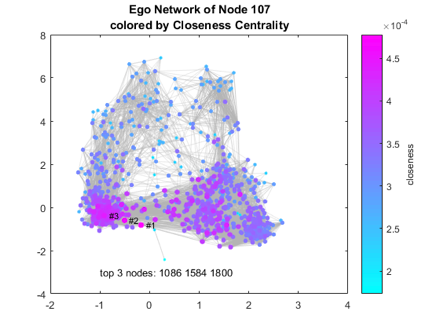
Those with high closeness scores are the ones you want to start with when you want to spread news through your ego network.



When we use the node degree to find the top 3 nodes and then compute the shortest path from a top node to any other node the result looks as such.



When we use the distance of the shortest path to compute the closeness centrality of the ego network of a node, the result is as follows.



V. Completed So Far (Conclusion)

So far, we have analyzed the datasets to the best of our ability in order to create a comprehensible summary. While the models are still in the process of derivation, we have 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 can tell that the more popular a user is (by the node degree) the more influence they hold. Although, To more accurately predict a users 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 because information will propagate fastest through their ego network, thus making them the perfect “seed” user.

VI. Future Work

What remains to be done, is to find a machine learning algorithm for node clustering and optimization that that automatically identifies users’ social circles.

Continuing this research will allow us to take the most influential users and their features to define social circles related to certain interests.

Further developing algorithms will allow markets to exploit influential users to increase a product’s marketability autonomously.

VII. Contributions