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

Mark Weaver Maxwell Yi

[Mrw93@txstate.edu](mailto:Mrw93@txstate.edu) [mky7@txstate.edu](mailto:mky7@txstate.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 Google+ 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**

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 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. The algorithm may therefore be broadened to cover a wide demographic, allowing a business or strategist to increase their reach and, potentially, profit.

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. 1 contains node information from the website “Twitter.” Additionally, Fig. 2 contains node information from the website “Google +.”

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 107614 |
| Edges | 13673453 |
| Nodes in largest WCC | 107614 (1.000) |
| Edges in largest WCC | 13673453 (1.000) |
| Nodes in largest SCC | 69501 (0.646) |
| Edges in largest SCC | 9168660 (0.671) |
| Average clustering coefficient | 0.4901 |
| Number of triangles | 1073677742 |
| Fraction of closed triangles | 0.6552 |
| Diameter (longest shortest path) | 6 |
| 90-percentile effective diameter | 3 |

*Figure 1. Twitter Dataset Statistics*

|  |  |
| --- | --- |
| Dataset statistics | |
| Nodes | 81306 |
| Edges | 1768149 |
| Nodes in largest WCC | 81306 (1.000) |
| Edges in largest WCC | 1768149 (1.000) |
| Nodes in largest SCC | 68413 (0.841) |
| Edges in largest SCC | 1685163 (0.953) |
| Average clustering coefficient | 0.5653 |
| Number of triangles | 13082506 |
| Fraction of closed triangles | 0.06415 |
| Diameter (longest shortest path) | 7 |
| 90-percentile effective diameter | 4.5 |

*Figure 2. Google+ 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. 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].

1. **Completed So Far**

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.

1. **What Remains to Be Done**

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