**Nate Bachmeier**

Week 2: Explore Data Analysis

March 17th, 2019

TIM-8101 Principals of Computer Science

Northcentral University

# Sample Research Area

One of the challenges with exploring highly connected data sets is not to miss the forest for the trees. This is partially due to the locality of relationships, which forces us to start at a specific point and walk forward. To address these challenges data visualization can be leveraged to gain a broader understanding.

All figures were generated with Gephi and the Facebook data set provided by Leskovec (Bastian M., 2009) (McAuley & Leskovec, 2012).

## Environment Configuration

A visualization layer was not provided with Stanford’s Snap.py repository. The lack of one being provided out of the box led to the need for an open source solution. The industry has generally standardized on Apache TinkerPop for graph storage and Gremlin as the query engine. An ecosystem of tools like Gephi have been created around these open tools.

To get Leskovec’s data set into TinkerPop, a short script was created to generate the relevant Gremlin commands. Next, Gephi was connected to TinkerPop for visual filtering and inspection of the graph.

## Observation: The Overall Structure

The initial expectation was that the visualization would show multiple disconnected Ego networks, as the sampling was intended to be random. However, from the visualization it was immediately apparent that relationships exist to connect all networks in the cohort. This would suggest a limited diversity of user data to the original study. If there are not enough examples for the unsupervised algorithm, then it can run into challenges when applied to other markets.

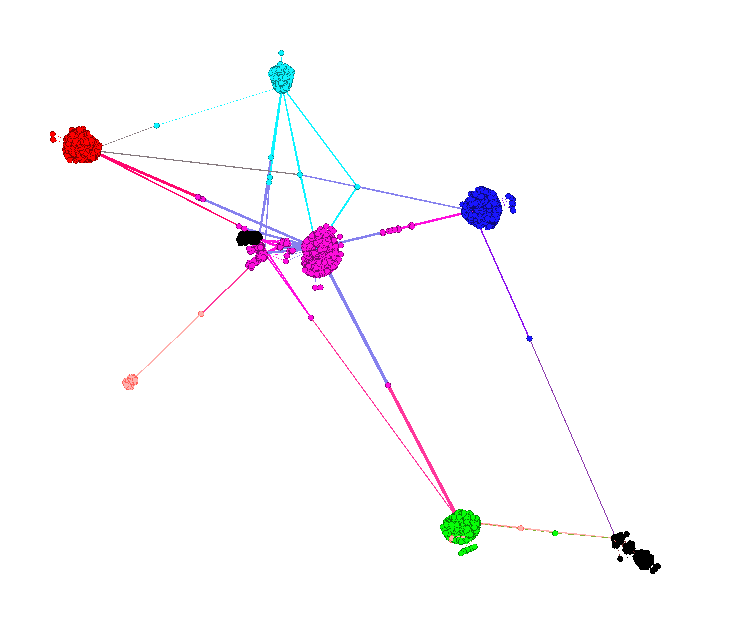


Figure 1: Facebook Network

## Observation: Connectivity of Root Nodes

Next the root nodes 0 and 107 were selected at random and their connectivity investigated. As parts of separate ego networks, it was surprising to find they related with a shortest path of length=1. The connecting edge was anonymized locale\_127 which suggests that the lived in a similar location.

Continuing to explore the edges between these two returns: education\_degree\_id\_22, education\_school\_id\_52, education\_type\_53, education\_type\_54, education\_year\_id\_66, education\_year\_id\_70, and gender\_78. From this information it appears that these users attended the same degree program at the same time. This validates our expectations based on the visual representations from Gephi.

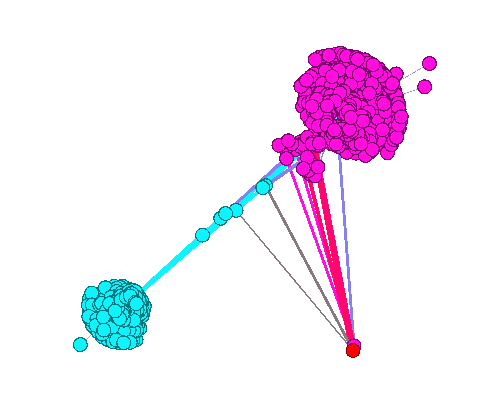


Figure 2: Network 0 and 107

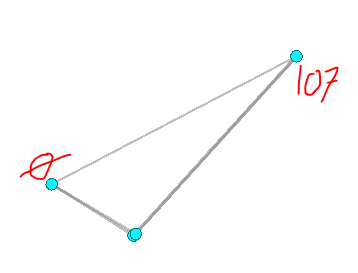


Figure 3: Shortest Path Between Nodes

# References

Bastian M., H. S. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. AAAI.

McAuley, J., & Leskovec, J. (2012). *Social circles: Faebook.* Retrieved from Stanford Network Analysis Project: http://snap.stanford.edu/data/ego-Facebook.html