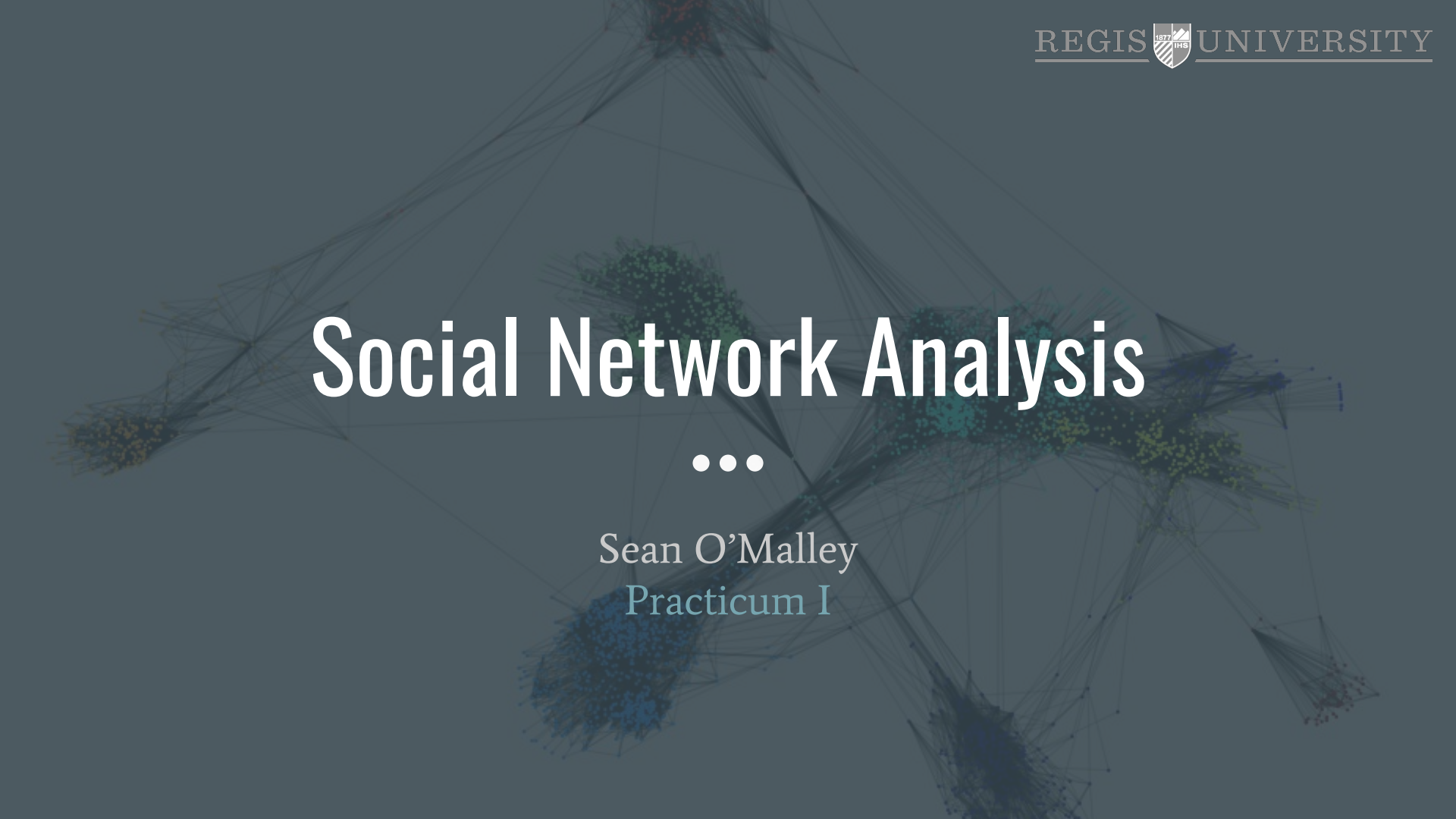


Social Network Analysis

Sean O'Malley

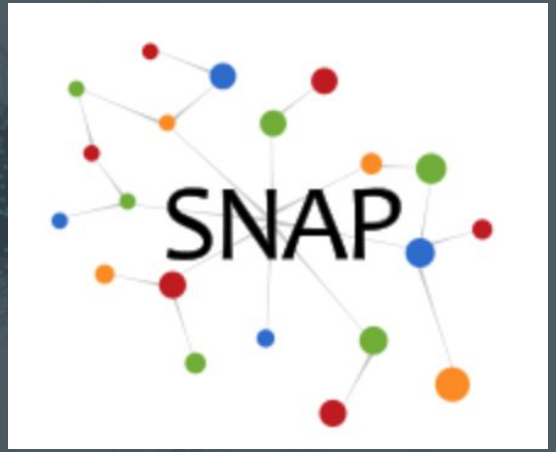
Practicum I

Abstract: Project

- The intent of this analysis is to be able to take a large network of undirected connections and make sense of the network, while also giving shape to the possible groups and influencers within.
- The analysis is framed as a test case for understanding of a community the minimal input of an edgelist.
- In my current role, I have been researching and practicing data science in the poorest areas of Perú with a specific focus on the communities of Pamplona Alta and Ayaviri with the aim of understanding urban and rural poverty in the developing world better.
- We hope to one day be able to use a similar data set map and interpret community structures in order to help them more.

Abstract: Data

- Stanford Network Analysis Project is a collection of 50 large network datasets.
- The dataset we use is a compressed text file friend connections from Facebook collected from survey participants.
  - EG: (4354,6901)
- The structure represents the reality of communities extremely well, offering a fantastic testing ground for in depth community analysis simply using a list of connections



Abstract: Analysis

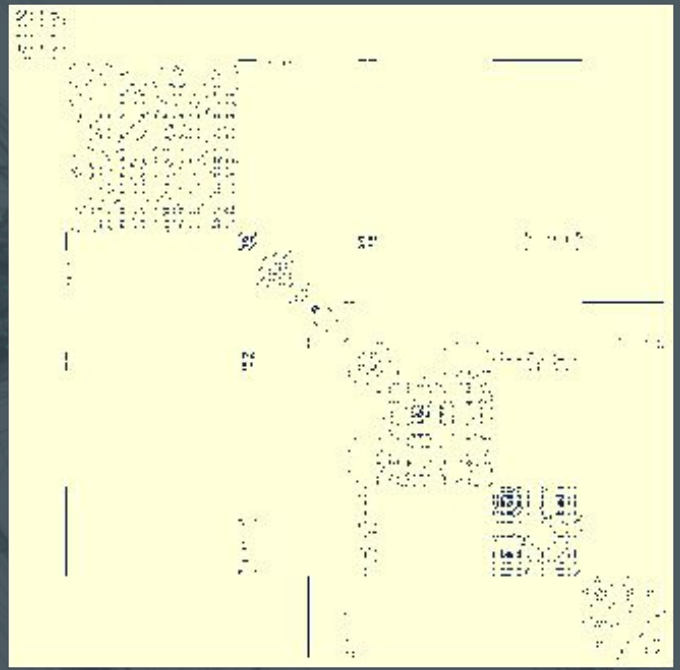
- The analysis used is network analysis, which comes from the concept of network (graph) theory.
- Network theory is the representation of symmetric relations between discrete objects.
- This analysis can be used to graph and understand various social and transportation networks.
- Primarily, network analysis helps us model relationships between entities, determine entity importance and find community structures within a network.

Process:

- Exploratory Data Analysis
  - Degree
  - Degree Centrality
  - Betweenness Centrality
- Communities
  - Girvan-Newman Algorithm
  - Compare Communities
- Influencers
- Subplots
- Triangles
- Friend Suggestion Engine
- Summary

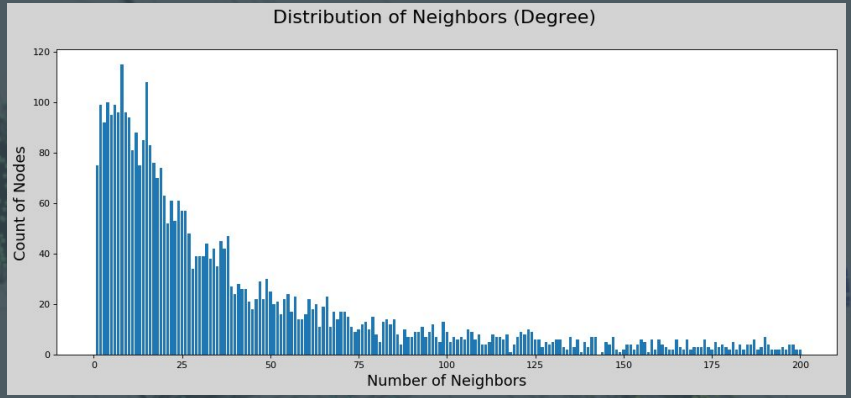
Exploratory Data Analysis: The Network

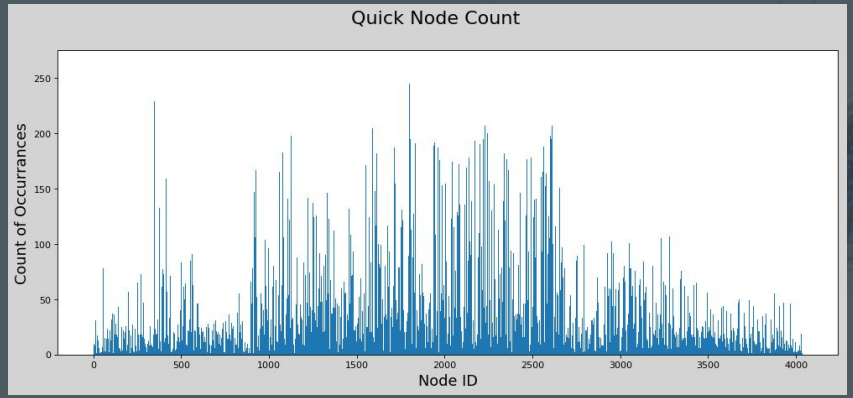
- In the matrix plot to the right, notice the high connectivity within certain grid segments and low connectivity with other segments, implying a natural community structure simply based upon ingest.
- A matrix plot returns the matrix form of the graph where each node is one column and one row, and an edge between the two nodes is indicated by the value 1 (dark value).
- The chart confirms our statistical analysis, showing that there are separate groups that are largely unconnected to one another in the larger network.



Exploratory Data Analysis: Degree

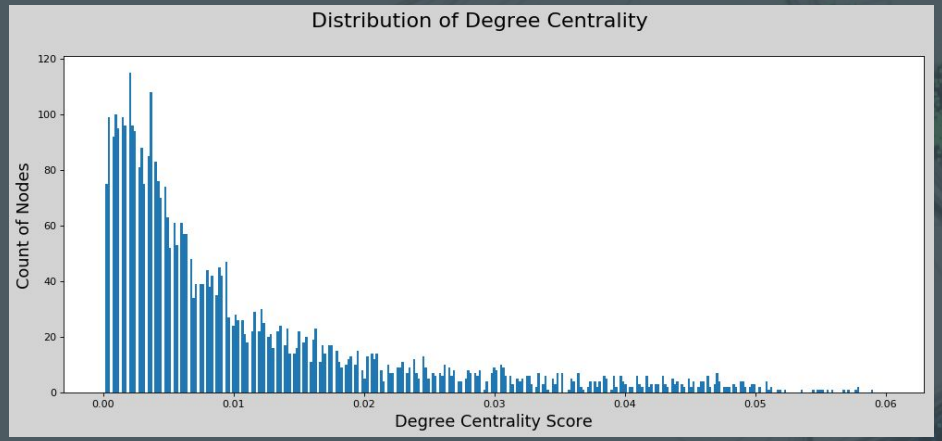
- The degree of a node in a network is the number of connections it has, while the degree distribution is the probability distribution of the degrees of nodes across the entire network.





- The second graph of the number of times a node occurred in the list of relationships, this is the degree.
- Visually we see that there a majority of nodes have below 50 connections, while a portion of nodes occur 150+ times in the list of relationships.

Exploratory Data Analysis: Degree Centrality



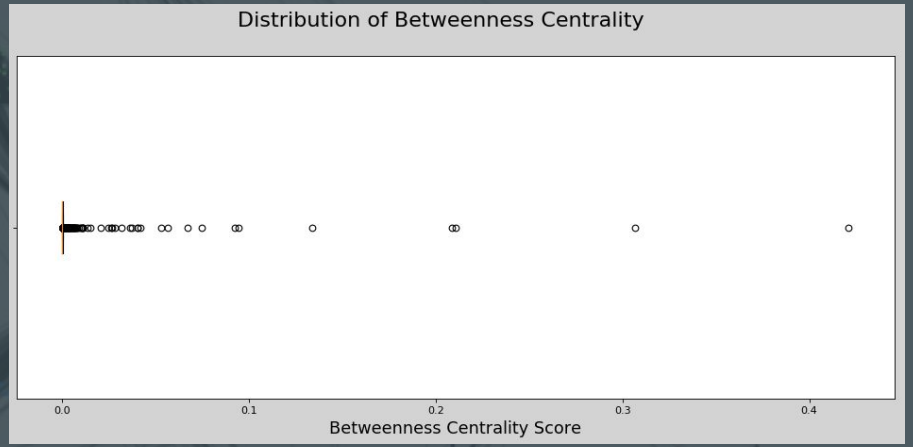
- The degree centrality takes the degree process a step further, by giving an importance score based on the number of links held by each node.
- Providing a measure of node connectivity in a simple equation

$$\text{Degree Centrality} = \frac{\text{Number of Neighbors}}{\text{Number of Total Possible Neighbors}}$$

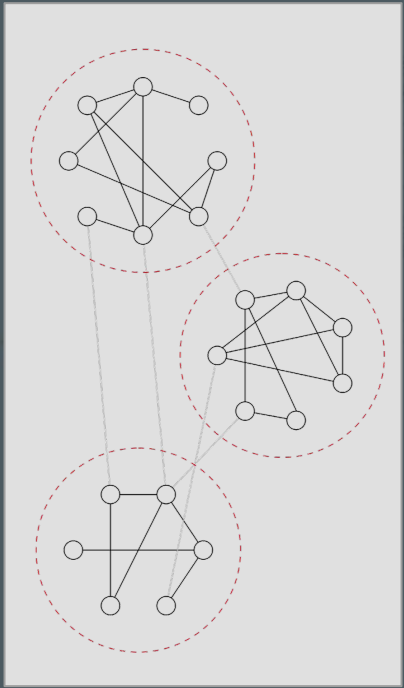
- The distribution to the left suggests that a majority of nodes are poorly connected in regards to the larger network

Exploratory Data Analysis: Betweenness Centrality

- Betweenness centrality is based on the concept of the shortest path, which states, for every pair of nodes in a connected graph there is a shortest path available between the those nodes.
- The betweenness centrality score is a quantification of the count of shortest paths that pass through a specific node.
- Thus, the betweenness centrality score works as another measure for node (user) importance.
- The distribution score to the right suggests that a majority of nodes are far from influential, but there are others that yield a large amount of influence in the entire network.



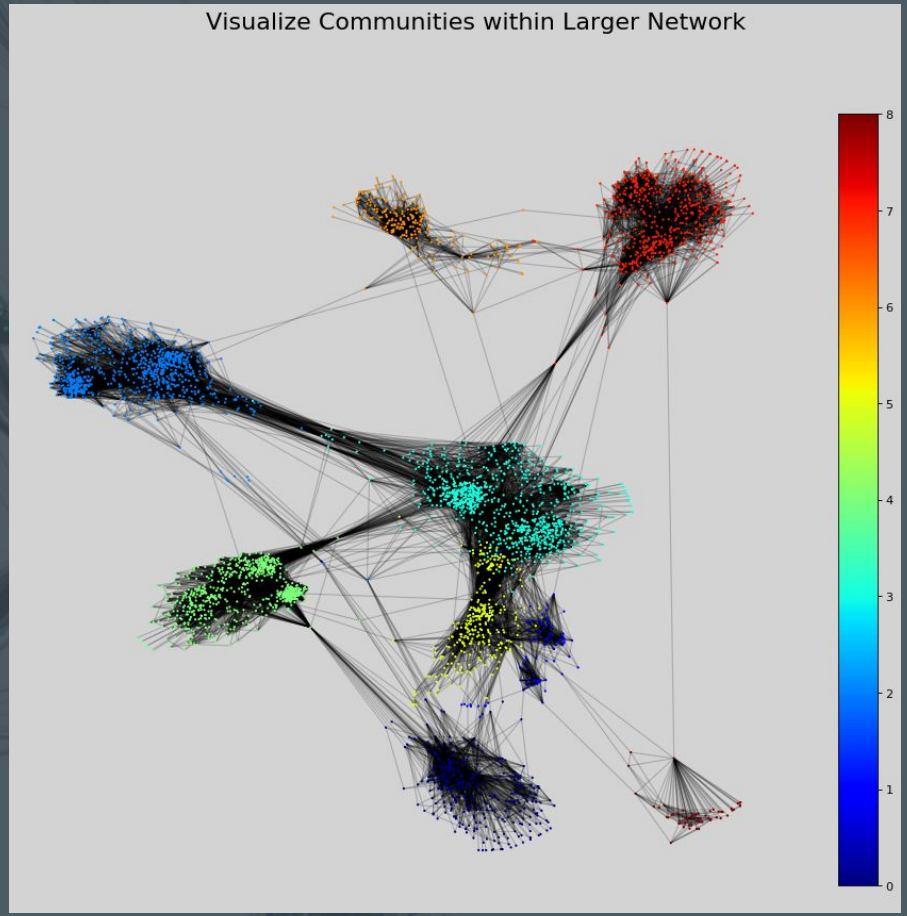
Find Communities: The Girvan-Newman Algorithm



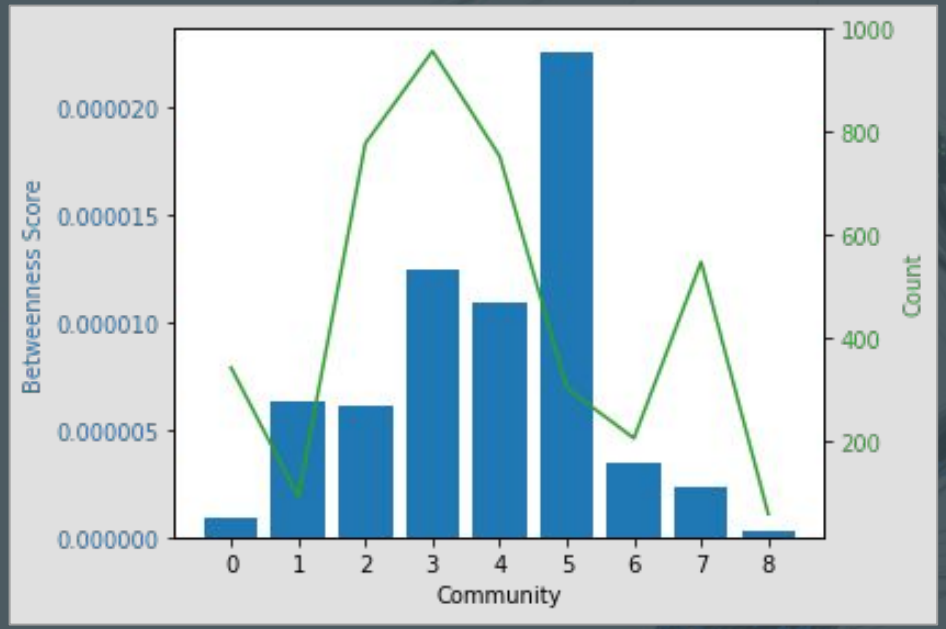
- The Girvan–Newman Algorithm detects communities by progressively removing edges from the original network, the remaining network connected components are the ‘communities.’
- The algorithm focuses on edges that are most likely between communities, which is an extension of edge betweenness, which we analyzed earlier.
- Step by step, the Girvan Newman Algorithm determines the betweenness of all existing edges in the network, then removes the highest betweenness score. After which the the effect of this removal is then measured by a recalculation of the betweenness scores. This process is then repeated until there are no longer edges remaining.

Find Communities: Visualize

- This visualization has the greatest explanatory power of any analysis thus far, helping us visualize communities in the network by color.
- There are six clearly defined communities and three in which there are some overlap between the communities.
- A majority of the communities also have specific nodes serving as a conduits to other communities, suggesting the presence of influencers.
- This confirms our betweenness centrality analysis, where we saw a majority of nodes having low betweenness scores (nodes that exist solely within the smaller community structures) and also nodes with extra-ordinarily high scores (the influencers).



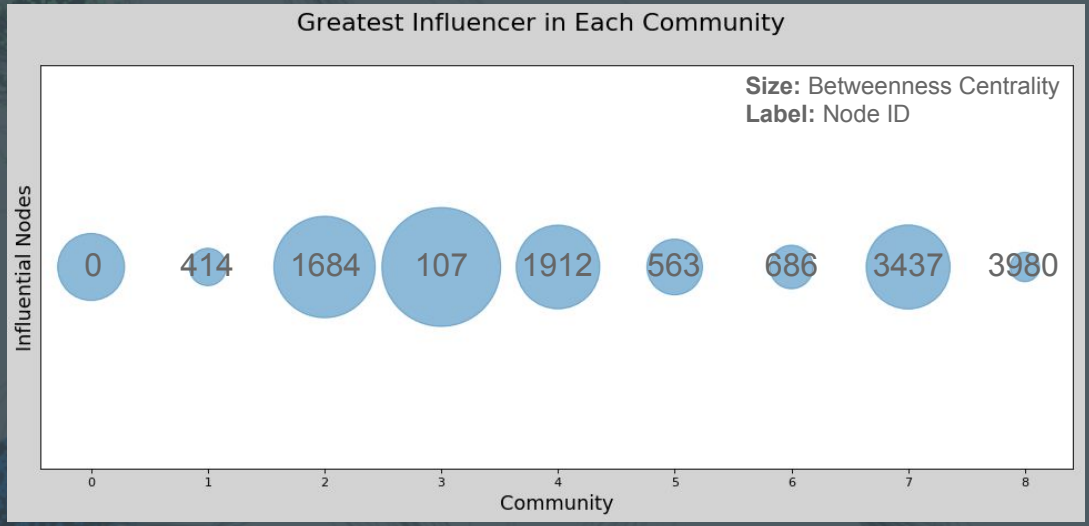
Understand Communities: Compare Summary Statistics



- The number of nodes in a community in comparison to the betweenness score tells us node size effect on overall connectivity scores.
- Our results suggest that the betweenness score is affected by, but far from reliant on, the overall community size.
- We see community 5 has a high connectivity, but not necessarily so in volume.
- Communities 0 and 8 are largely disconnected from the larger network, even with community 0 having a large community.

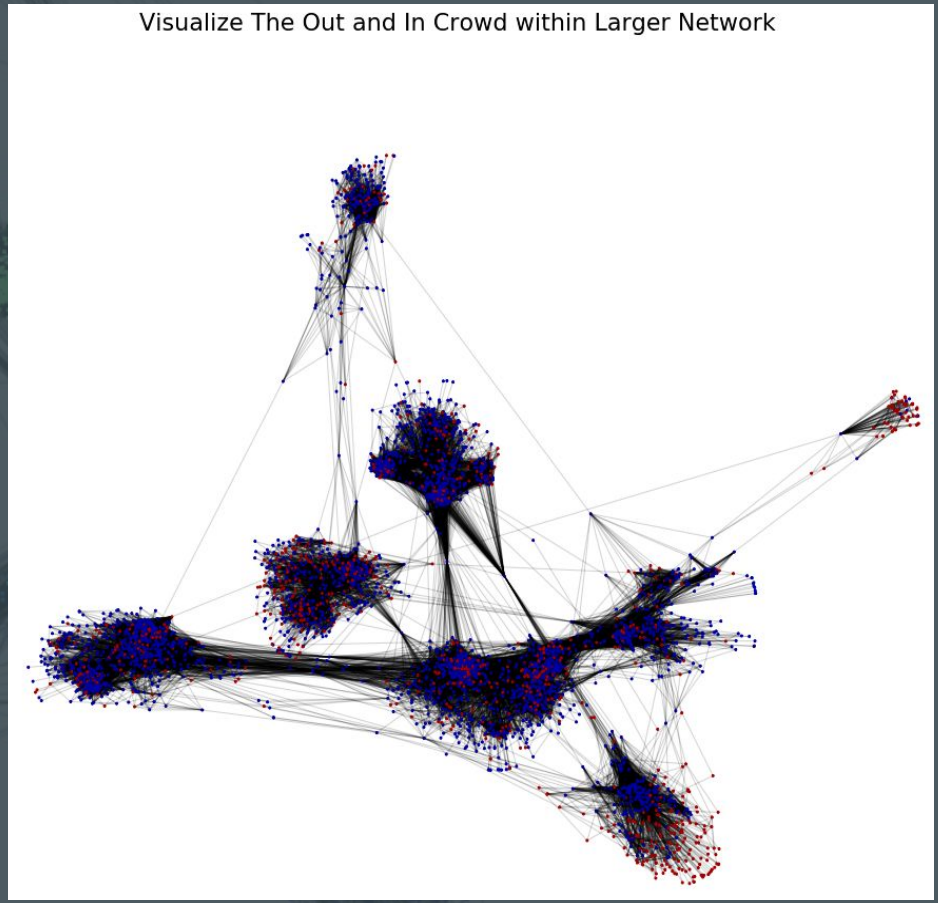
Understand Communities: Influencers

- A major insight of social network analysis is to find important people in the network and quantify exactly how influential they are.
- In thee structure of the entire network we care most about community leaders, therefore we determined the nodes with the highest betweenness score in each community
- With a list of the most influential people in each community we now have maximized leverage to influencing each community.



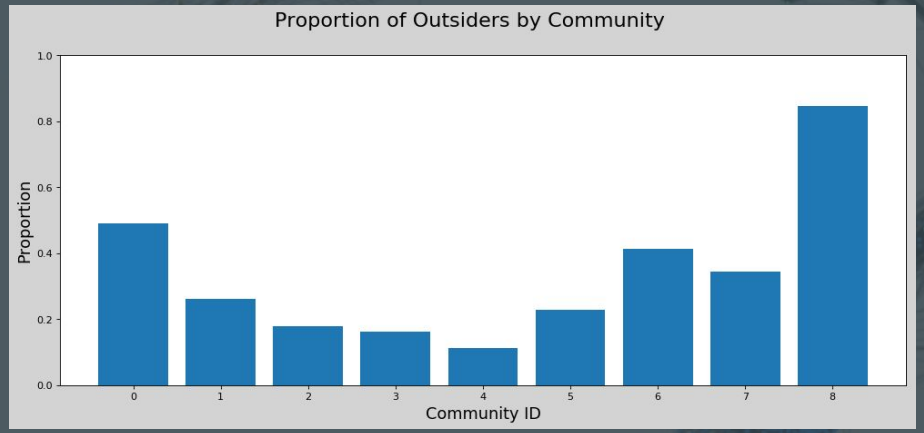
Understand Communities: Outsiders

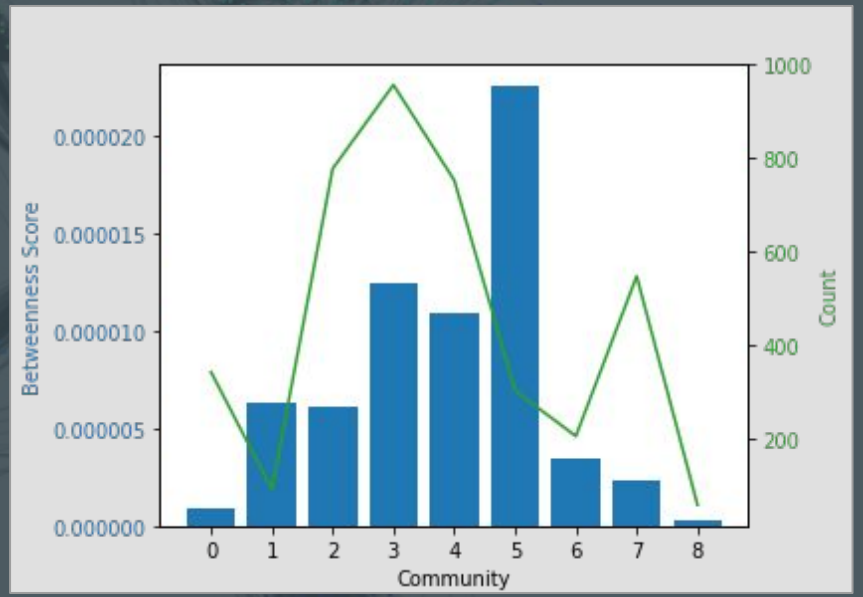
- Priority of understanding the whole community, not just the important people
- Outsider: Person with 10 or fewer connections
- Smaller communities are not well connected to other groups.



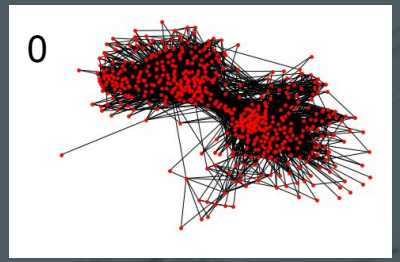
Understand Communities: Outsiders

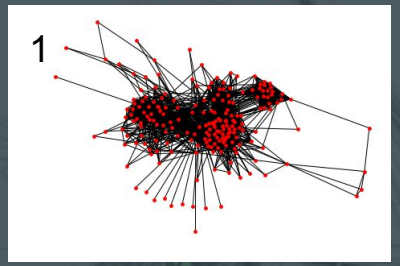
- In the left graph, the higher the proportion, the higher the number of outsiders in that community.
- In the right graph, we see the community size and the betweenness score

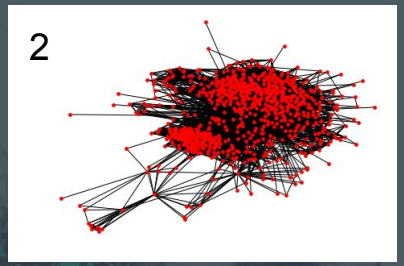


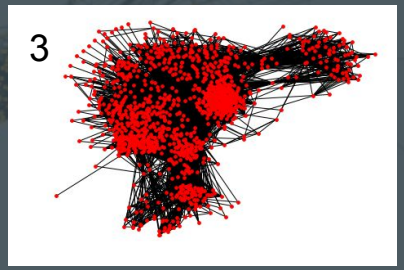


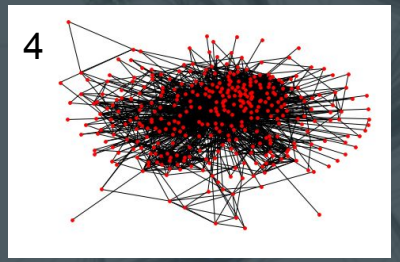
Specific Community Analysis: Subgraphs

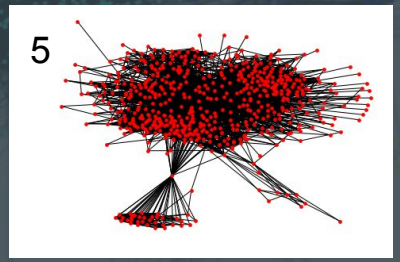


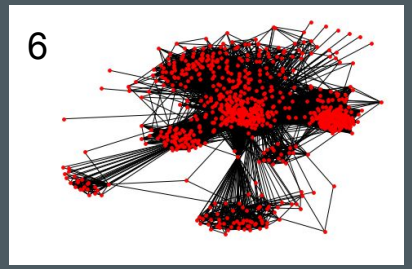


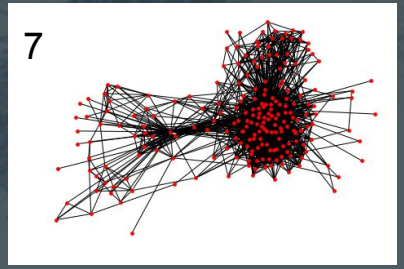


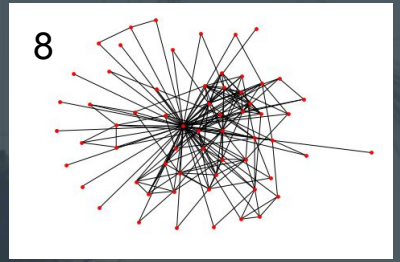












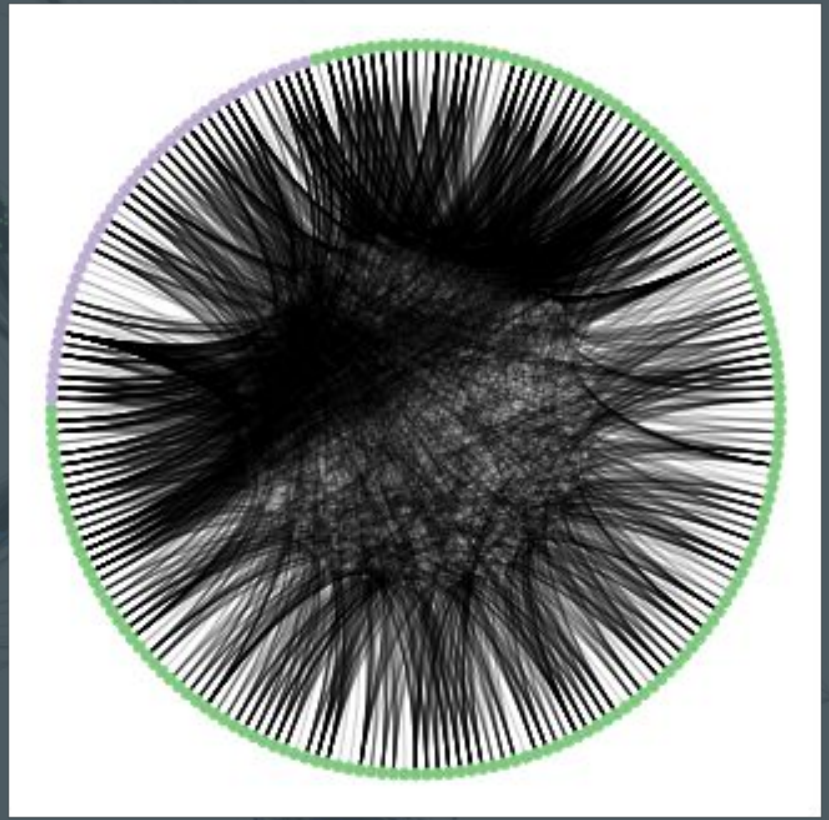
Specific Community Analysis: Circos Plot

Visualize Community 1

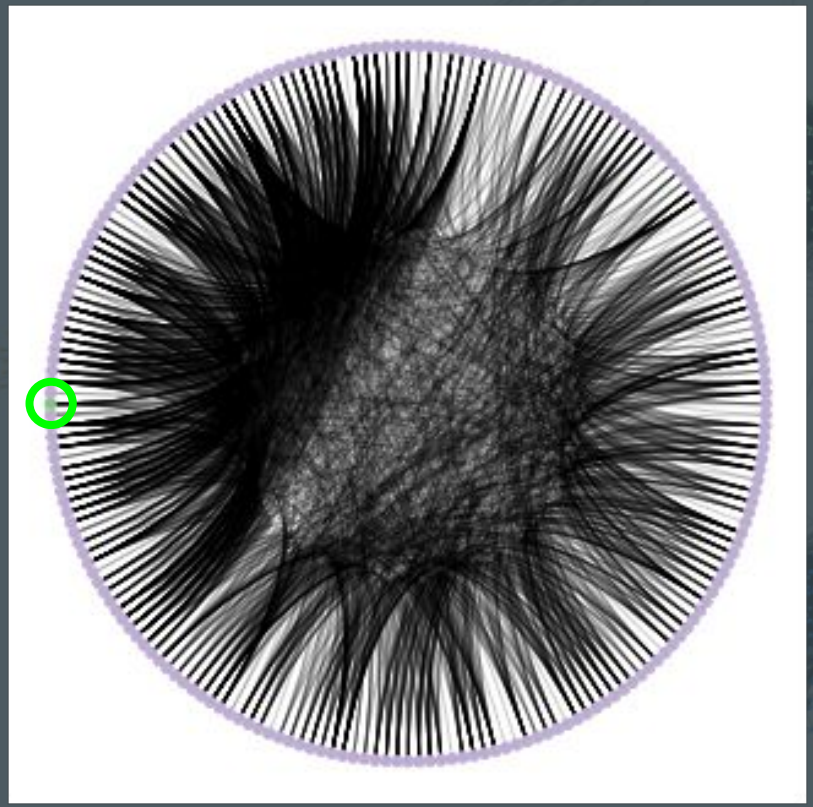
- Small, yet strong community that performs well in nearly every metric.

Circos Visualization

- Nodes distinguished by status:
  - Outsider -- purple
  - Normal -- Green



Specific Community Analysis: Circos Plot



Visualize Community 1

- Small, yet strong community that performs well in nearly every metric.

Circos Visualization

- Nodes distinguished by status:
  - Top Influencer -- green
  - Everyone Else -- purple

Friend Suggestion Engine: Open Triangles

Top Suggestions

1979, 2244	2104, 2324	2323, 2369
1985, 2607	1985, 2593	1917, 2233
2229, 2600	2233, 2414	2088, 2369

Summary: Understanding

Final Data Frame

- Community
- Node
- Degree Centrality
- Between Centrality
- Community Leader
- Community Outsider
- Suggested Friend

Summary: Value Production

Ingestion.

Understand the community.

Schooled in data visualization.

Human readable aggregate dataset describing the nodes involved in the network.

We have given shape and life to a simple list of friend pairs and built a community story.

Summary: Challenges & Opportunities

Computational expense of community building.

Building the communities required extensive processing and memory with an 88k edge dataset.

The second challenge was that of networkx. The package is still in its beginning years and many features are in the beta phase and contain little documentation and a couple bugs.

We garnered serious computational insight performing this analysis and many of the concepts are special and easily comprehensible.

Opportunities would surround gaining additional demographic and personal information on the nodes to better understand these communities and the nodes within.

References & Resources

Github Location to All Files & Scripts:

- https://github.com/showmalley/SeanOMalleyCodePortfolio/tree/master/Python%20--%20Network%20Analysis

Resources

- http://snap.stanford.edu/data/ego-Facebook.html
- https://en.wikipedia.org/wiki/Girvan%E2%80%93Newman_algorithm
- https://econsultancy.com/twitter-network-analysis-identifying-influencers-and-innovators/
- https://www.datacamp.com/courses/network-analysis-in-python-part-1
- https://www.datacamp.com/courses/network-analysis-in-python-part-2