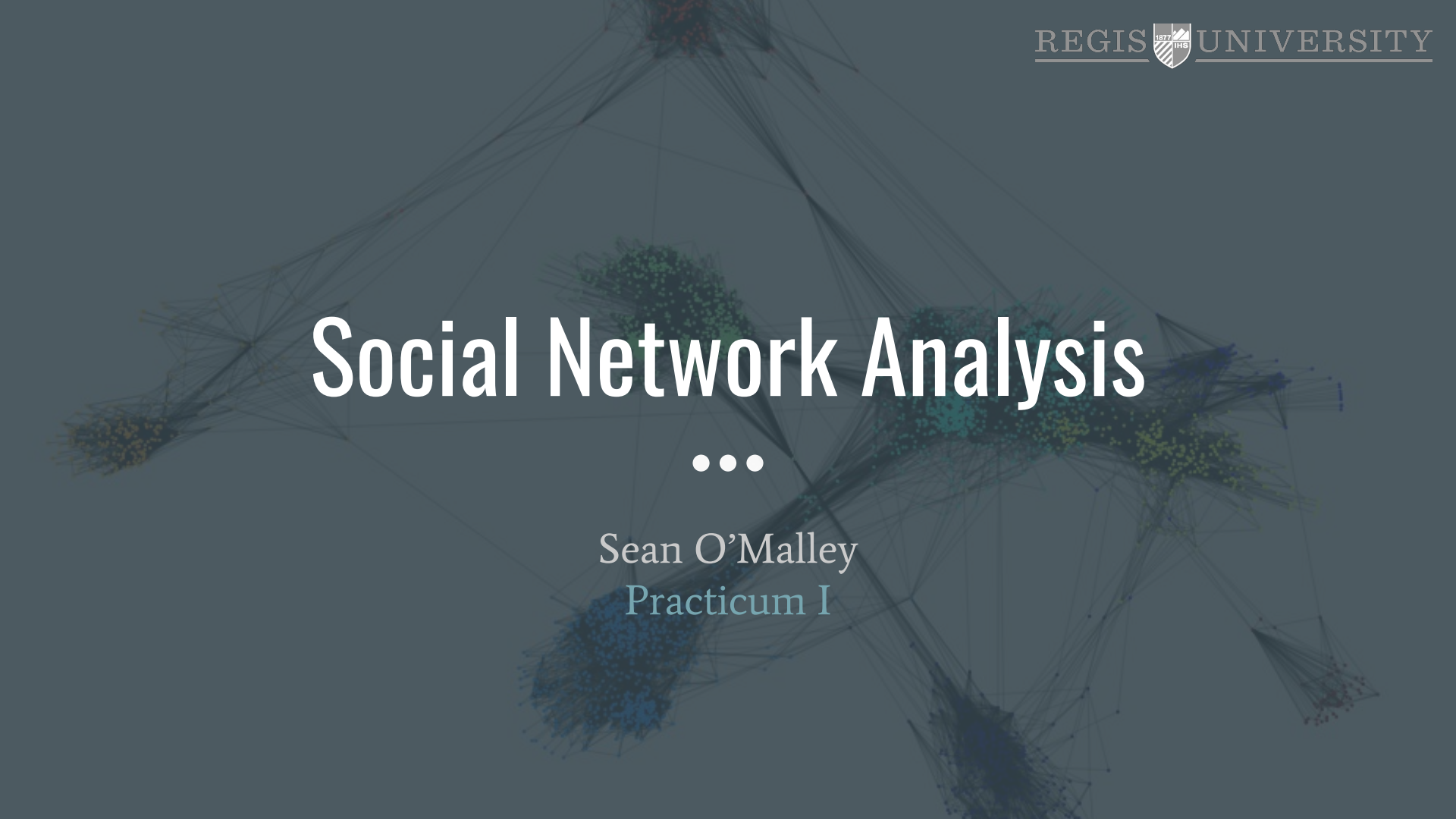


Social Network Analysis

Sean O'Malley

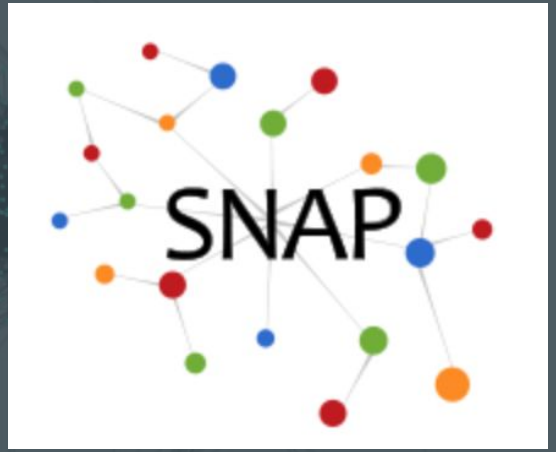
Practicum I

Abstract: Project

- The intent of this analysis is to be able to take a large network of undirected connections and make sense of that network, while also giving shape to the possible networks within.
- The analysis is framed as a test case for maximizing understanding of a community given minimal input.
- 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

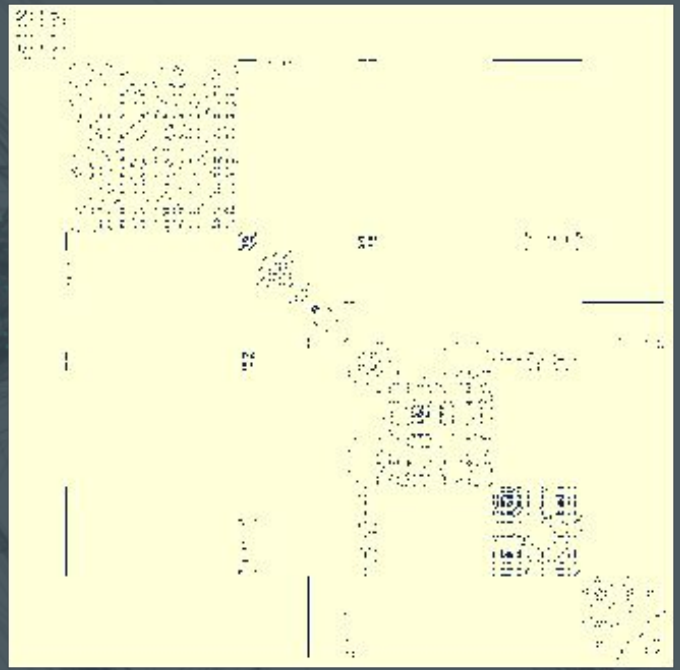
- The analysis used is network analysis, which comes from the concept of network (graph) theory.
- Network theory is the representation of symmetric relations between discrete objects.
- This analysis can be used to graph and understand various social and transportation networks.
- Primarily, network 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
- Cliques
  - Maximal Cliques
  - Triangles
- Friend Suggestion Engine
- Summary

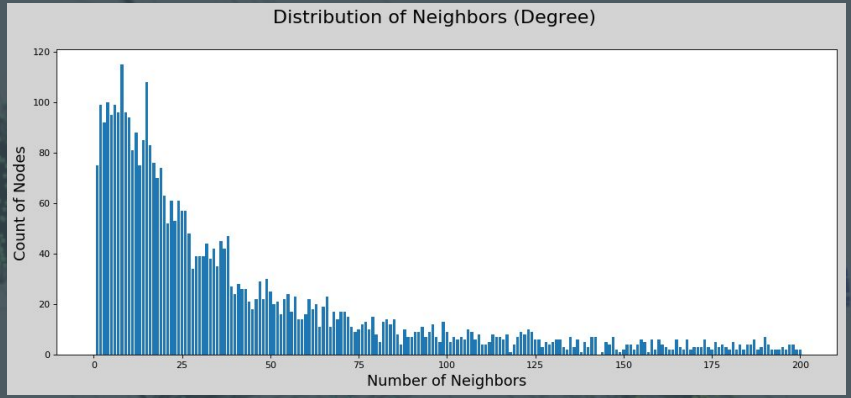
Exploratory Data Analysis: The Network

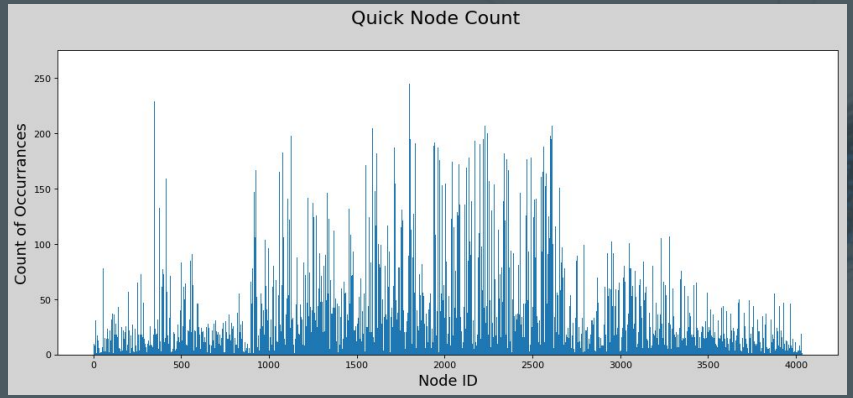
- In the matrix plot to the right, notice the high connectivity within certain grid segments and low connectivity with other segments, implying a natural community structure simply based upon ingest.
- A matrix plot returns the matrix form of the graph where each node is one column and one row, and an 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

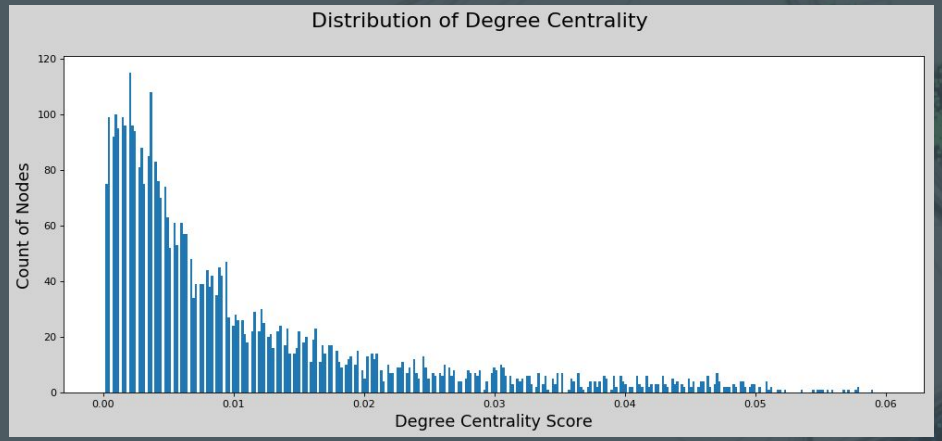
- The degree of a node in a network is the number of connections it has, while the degree distribution is the probability distribution of the degrees of nodes across the entire network.





- 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



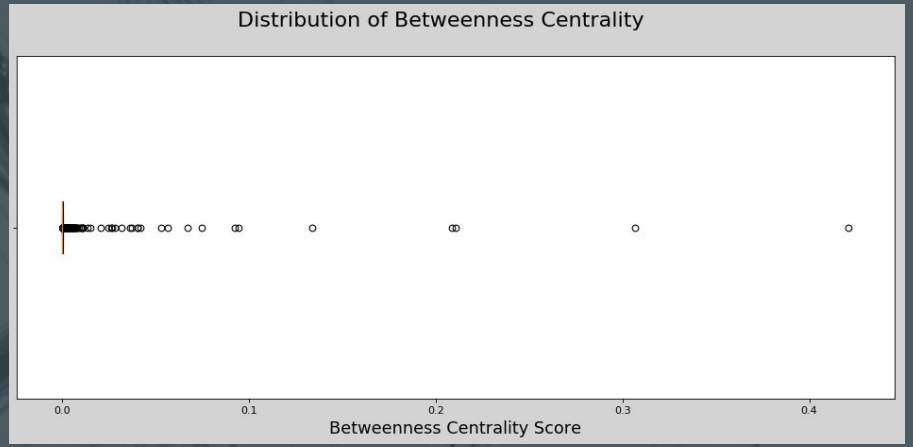
- The degree centrality takes the degree process a step further, by giving an importance score based on the number of links held by each node.
- Providing a measure of node connectivity in a simple equation

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

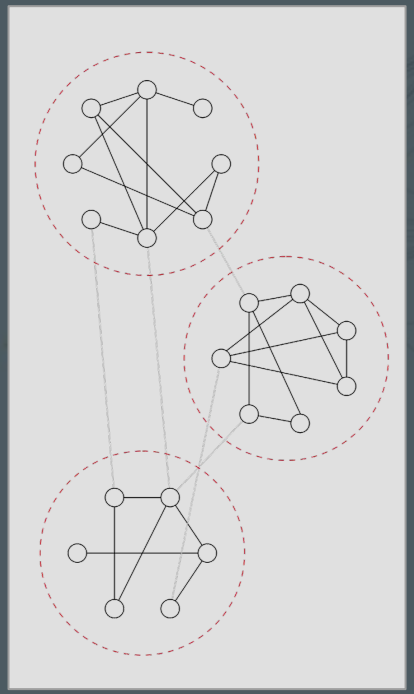
- The distribution to the left suggests that a majority of nodes are poorly connected in regards to the larger network

Exploratory Data Analysis: Betweenness Centrality

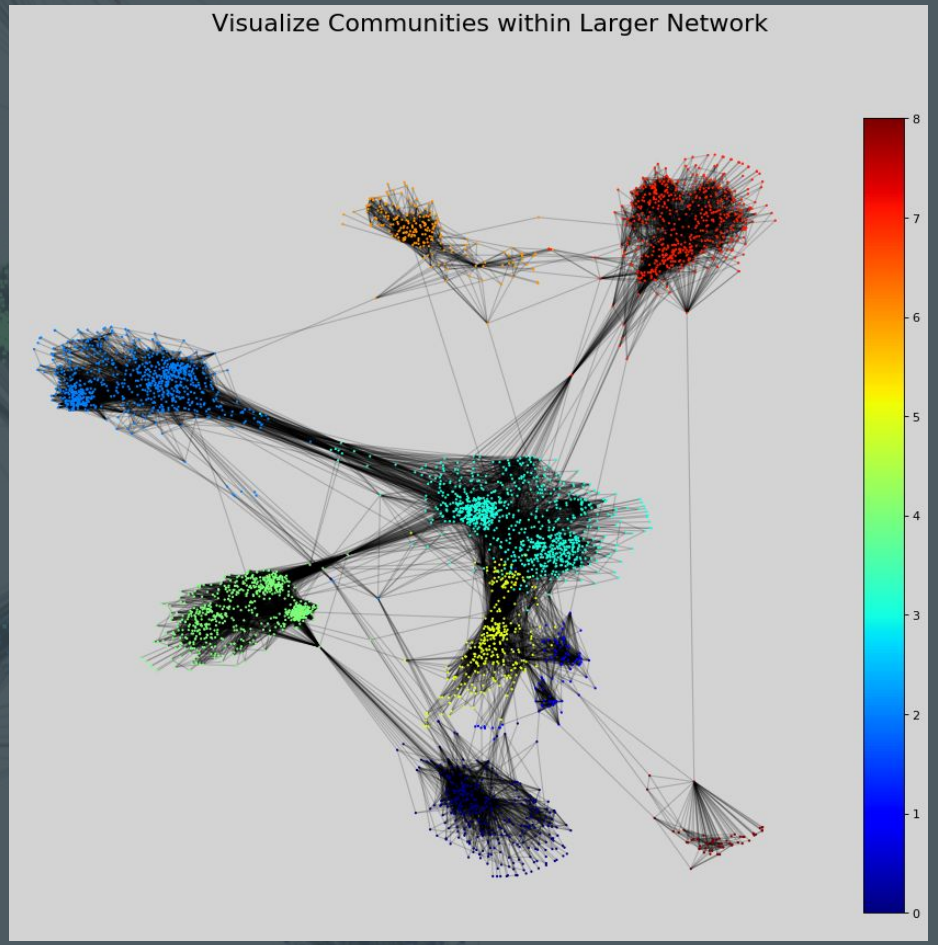
- Betweenness centrality is based on the concept of the shortest path, which states, for every pair of nodes in a connected graph there is a shortest path available between the those nodes.
- The betweenness centrality score is a quantification of the count of shortest paths that pass through a specific 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.



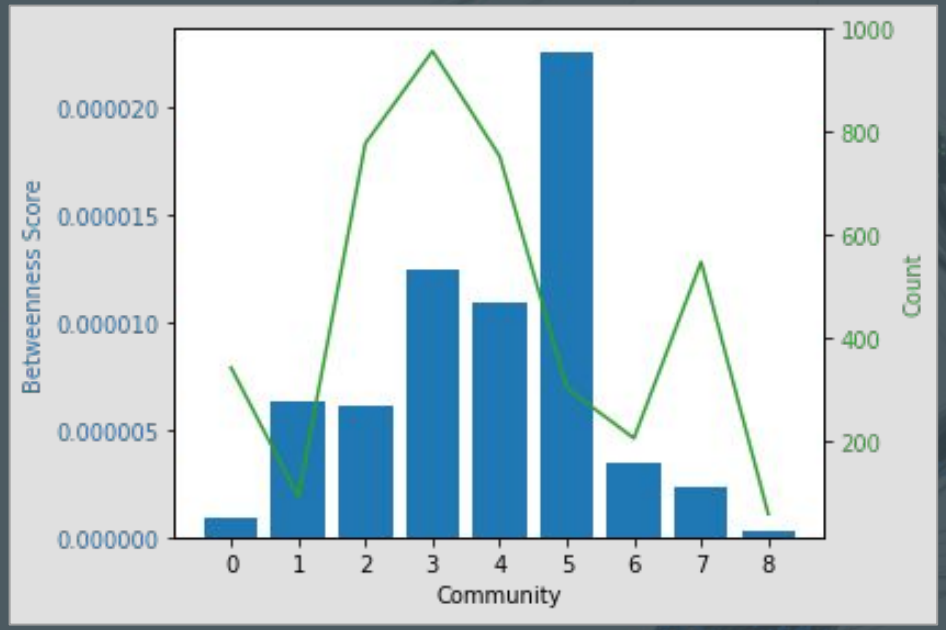
Find Communities: The Girvan-Newman Algorithm



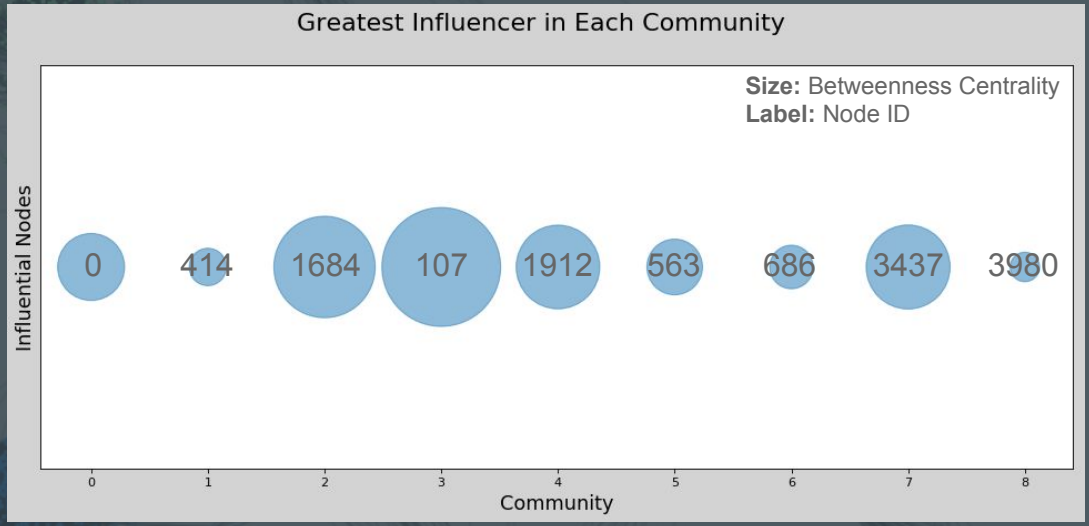
Find Communities: Visualize



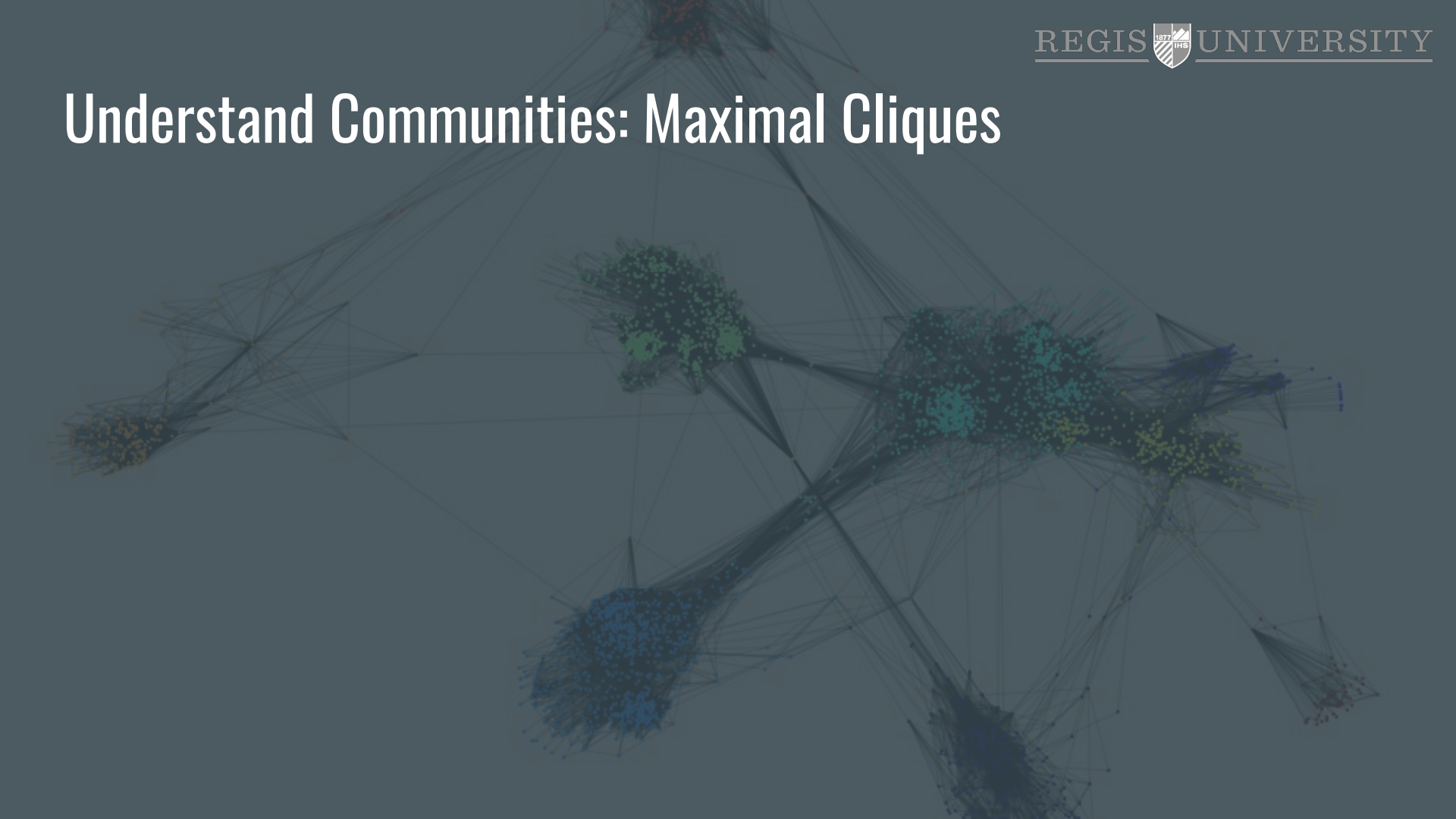
Understand Communities: Compare Summary Statistics



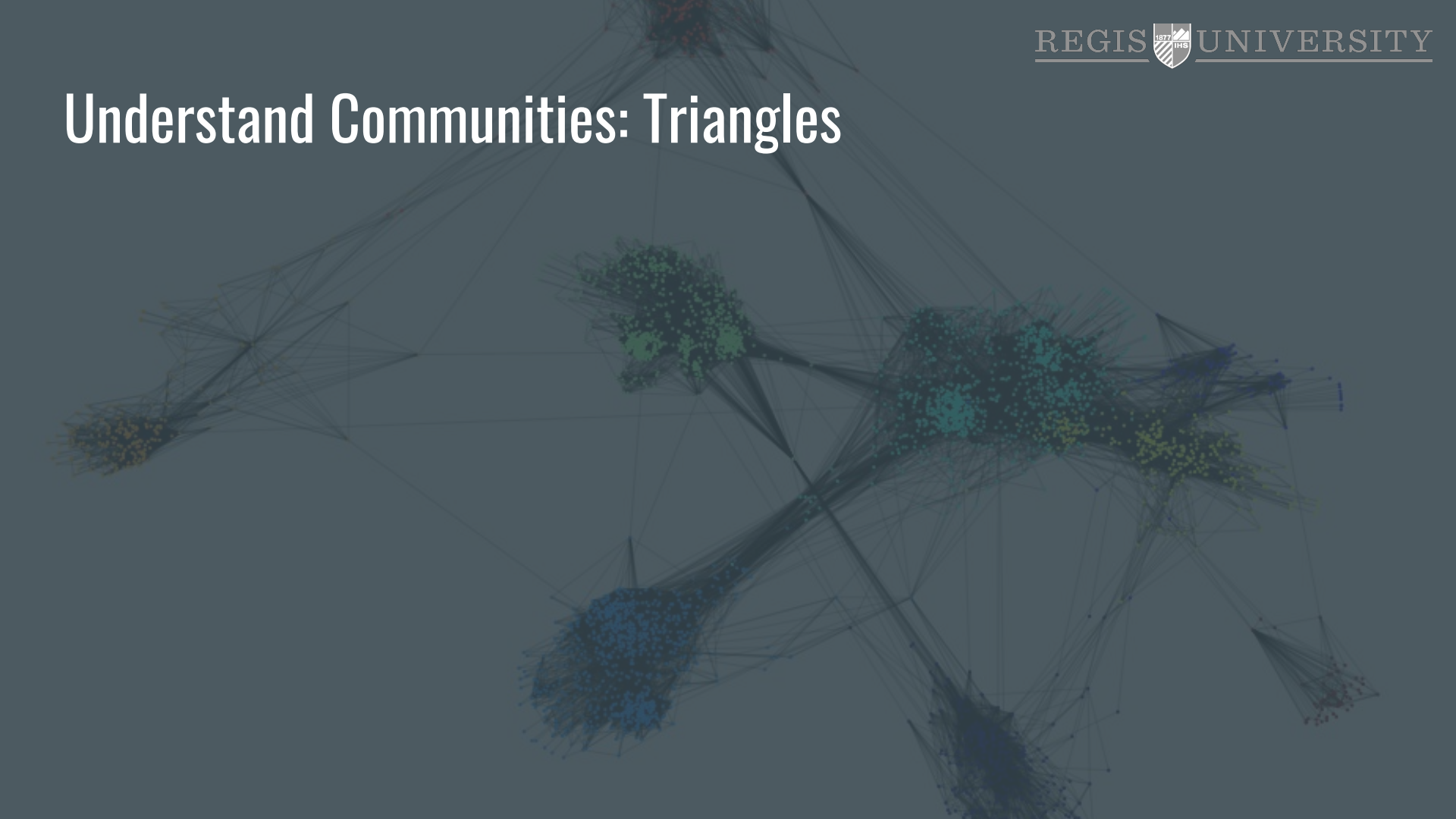
Understand Communities: Influencers



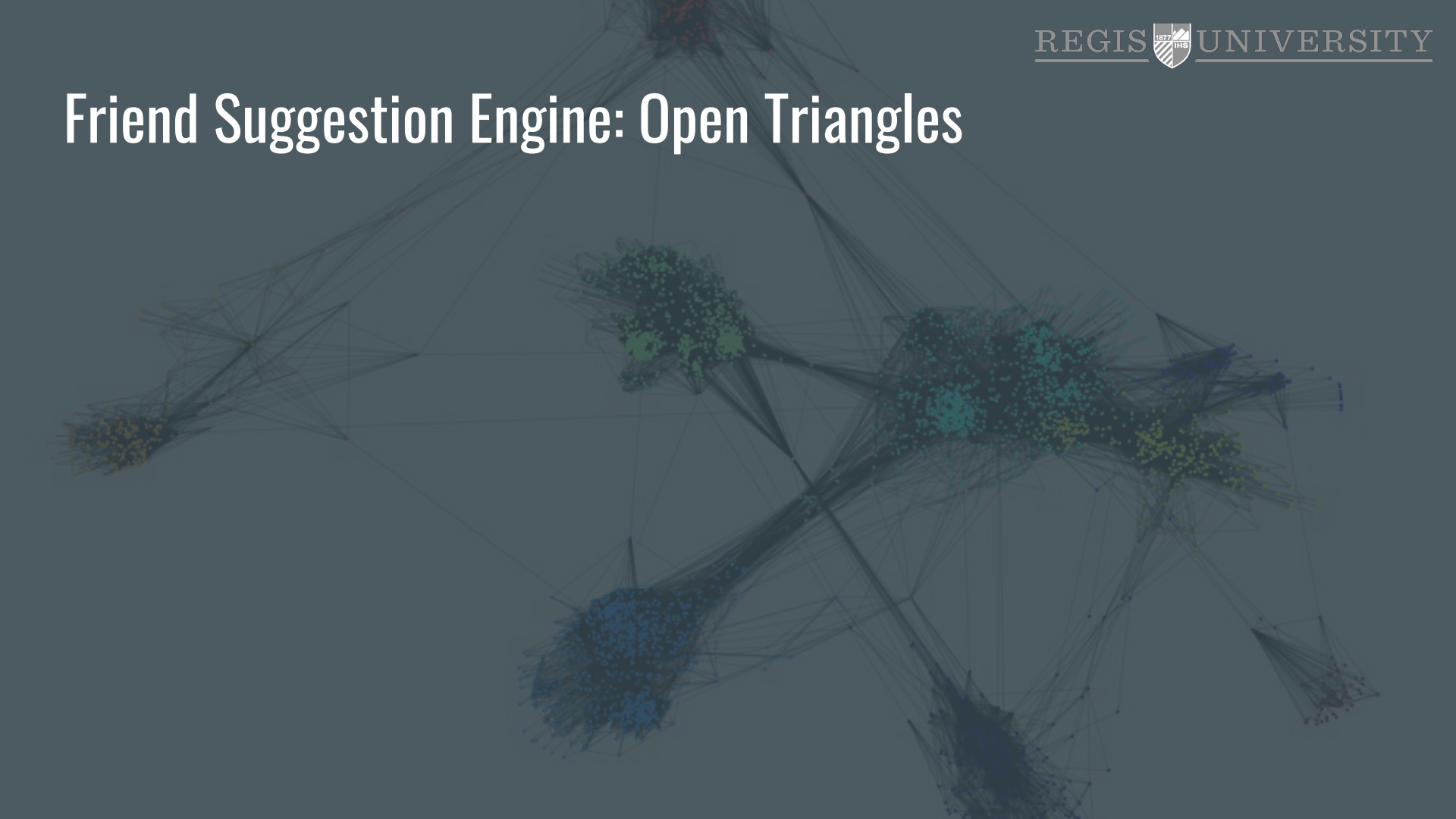
Understand Communities: Maximal Cliques



Understand Communities: Triangles

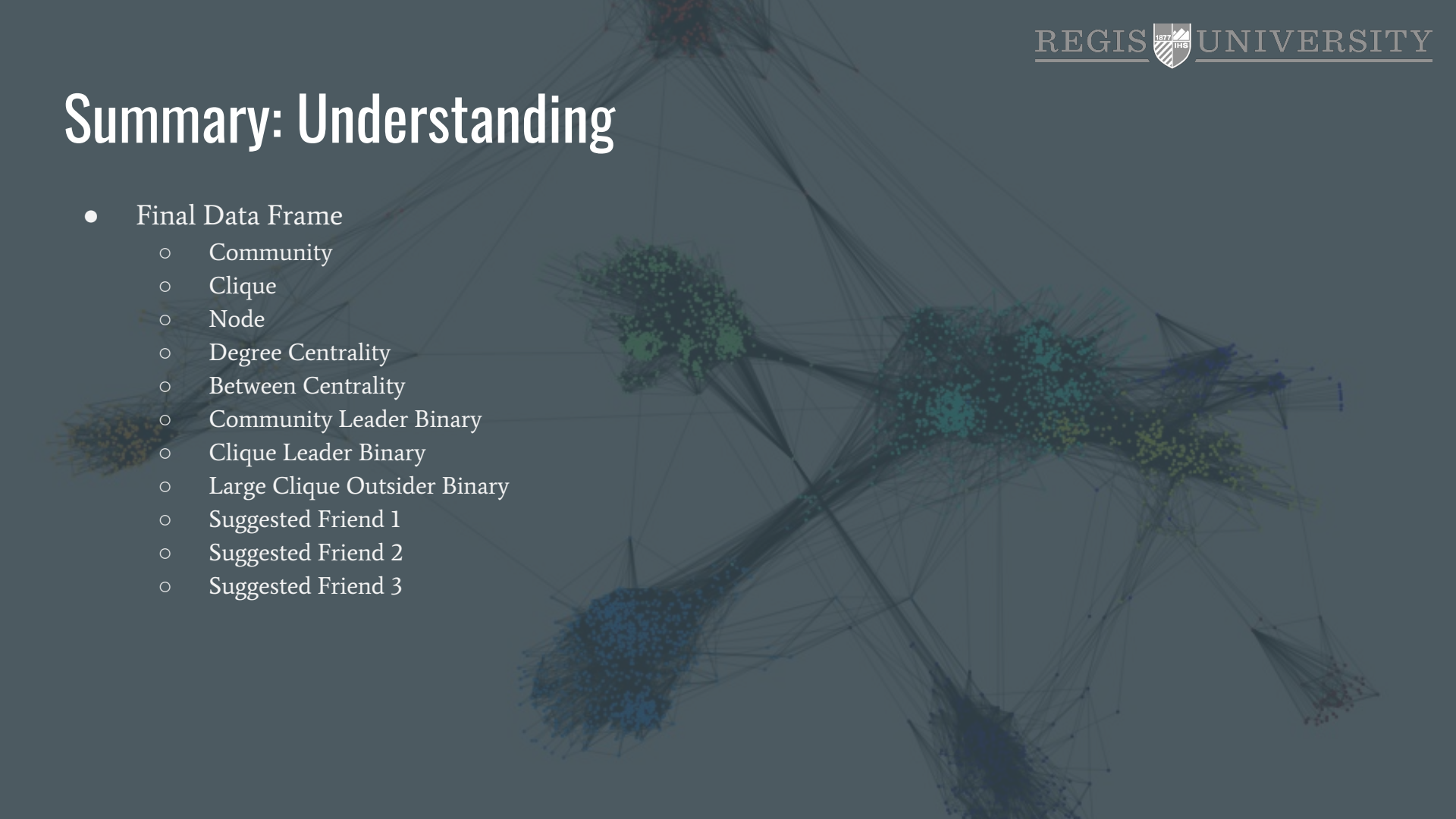


Friend Suggestion Engine: Open Triangles

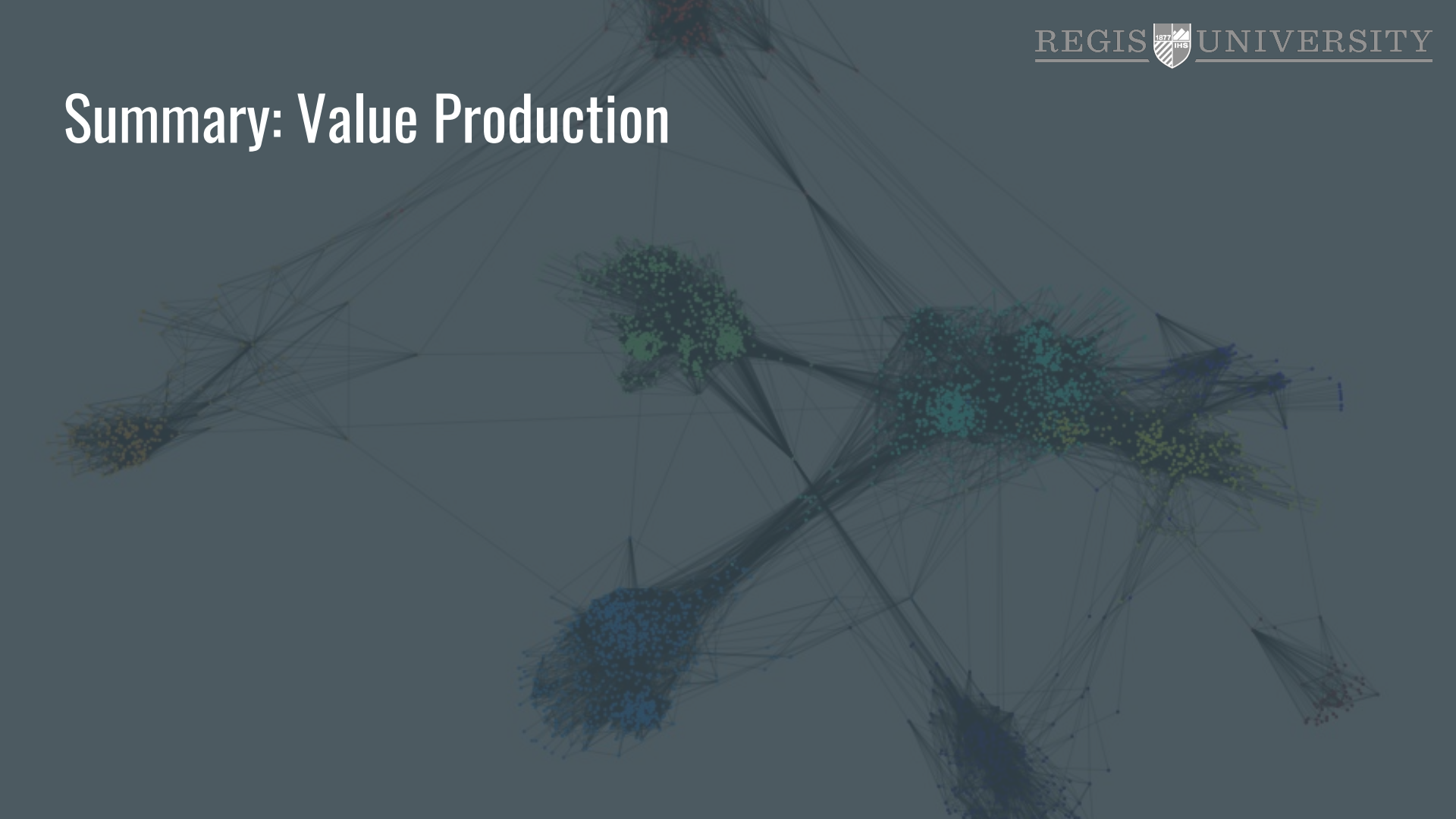


Summary: Understanding

- Final Data Frame
  - Community
  - Clique
  - Node
  - Degree Centrality
  - Between Centrality
  - Community Leader Binary
  - Clique Leader Binary
  - Large Clique Outsider Binary
  - Suggested Friend 1
  - Suggested Friend 2
  - Suggested Friend 3



Summary: Value Production



Summary: Challenges & Opportunities

References & Resources