Sunil Thapa
W0779269

To: Prof. Paulo Regis
From: Sunil Thapa
Data: October 31, 2025
Subject: Capstone Project Proposal – Homophily and Social Structure in the socfb-Auburn71 Network

This document outlines my proposal for the Comprehensive Network Analysis capstone project. I will conduct an analysis of the socfb-Auburn71 dataset [1], a large, real-world friendship network from Auburn University. This dataset is ideal for the project as it includes rich node attributes, which allow for an investigation that goes beyond simple network metrics. My project will focus on the principle of homophily to determine which of these real-world social attributes is the most dominant driver in the formation of online social communities.

Data Selection
The dataset I have selected is the socfb-Auburn71 network, which is part of the well-known "Facebook100" collection. It is publicly available from the Network Repository and was originally compiled by Traud et al. (2012). This graph represents an undirected, unweighted network of friendship ties at Auburn University. In this network, nodes represent university members (students, faculty, etc.), and an edge between them signifies a mutual "friendship" link. This dataset is particularly well-suited for a substantive research question as it is an attributed graph. Each node is labelled with categorical data, including the user's gender, class year, major, high school, and residence (dormitory). The network is large, with approximately 18,400 nodes and 973,900 edges.

Research Question
My project will investigate the sociological principle of homophily, which is the tendency for individuals to associate and form ties with similar others. My primary research question is: Which social attribute (class year, major, or residence) is the dominant driver of the social network's community structure at Auburn University? To answer this, I will also address two supporting inquiries. First, I will measure the overall strength of homophily across these different attributes to see which ones show the strongest "sorting" effect. Second, I will investigate whether the algorithmically-detected "emergent" social clusters align more closely with academic groupings (major), class-year cohorts, or physical location (residence).

Planned Methods
My methodology is a three-step process designed to move from a high-level statistical overview to a specific, structural comparison. First, I will calculate the Assortativity Coefficient [2] for each of the key attributes (major, class_year, residence). This will provide a single, powerful metric for each attribute, quantifying the baseline tendency for "like-to-like" connections. Second, I will identify the network's emergent social structure by applying the Louvain community detection algorithm [3]. This method optimizes for modularity and will partition the graph into de facto communities based purely on the friendship topology, without any knowledge of the attributes. Finally, I will use Normalized Mutual Information (NMI) [4] to directly compare the partitions. I will calculate the NMI score between the algorithmic Louvain communities and the "ground-truth"

attribute partitions (e.g., one partition for 'major', one for 'residence'). The attribute whose partition has the highest NMI score when compared to the Louvain communities will be the one that best explains the network's emergent social structure, thereby answering my primary research question.

Reference:

1. Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16), 4165-4180.
2. M. E. J. Newman (2003), Mixing patterns in networks. *Physical Review E,* 67 (2), https://doi.org/10.1103/PhysRevE.67.026126
3. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.
4. Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837-2854.