Report:
LinkedIn Network Analysis of Class of 2028

Prepared By:

Anshu Kumar

Freshman at Sitare University
Department of Computer Science and Engineering

15 April 2025

Overview

This report explores the LinkedIn network of **Sitare University's** Class of 2028. The analysis began with cleaning the raw student connection data—first manually, and then through a programmatic approach to ensure accuracy and consistency. A graph-based structure using dictionaries was created to represent each student and their connections, enabling the calculation of their degree (number of connections). To understand the connectivity within the network, random walks between pairs of students were performed, and cycle-free paths were extracted to identify meaningful routes. Finally, statistical estimates such as mean, median, and standard deviation were calculated to gain deeper insights into the network's overall structure and interaction patterns.

Data Collection and Cleaning Process

Initially, there were 131 CSV files, each representing the LinkedIn connection data of individual students.

- Duplicate CSV files were manually deleted from the folder.
- Files with read errors were replaced with corrected versions provided by the respective students.
- Some files in `.xlsx` format were converted to `.csv` for consistency.
- CSVs with incorrect file extensions like `..csv` were renamed to proper `.csv` format.
- After manual cleaning, the dataset was reduced to 126 clean and usable CSV files.

These cleaned files were then processed using Python code to:

- Read and combine data from all CSVs.
- Automatically remove missing or invalid entries.
- Standardize names and connections into a structured format for further analysis.

Graph Construction and Network Analysis

Once the cleaned data was ready, a graph-based structure was built to represent student connections.

- An adjacency list was generated, where each student was linked to all other students they are connected to. This helped in visualizing and navigating the entire network easily.
- The degree of each student was calculated by counting the number of unique connections they had. This metric represents how many peers each student is directly connected to in the network.
- To simulate how connections spread across the network, random walks were performed between randomly selected student pairs.
- From the paths generated by these random walks, cycle-free (non-repeating) routes were extracted to highlight clean and meaningful paths of interaction.

Statistical Estimates

This section summarizes the key statistics observed from 10 random walks and their pruned (cycle-free) paths in the student network.

Random Walks

- **Mean:** 595.70
- **Standard Deviation:** 585.89
- **Minimum:** 41
- **Maximum:** 1781
- **Median:** 405.0
- **Mode:** 1012

Pruned Paths

- **Mean:** 18.70
- **Standard Deviation:** 8.22
- **Minimum:** 6
- **Maximum:** 33
- **Median:** 16.5
- **Mode:** 23

Visualization of the Most Influential Nodes Based on Degree