SNA PROJECT UNIT 2

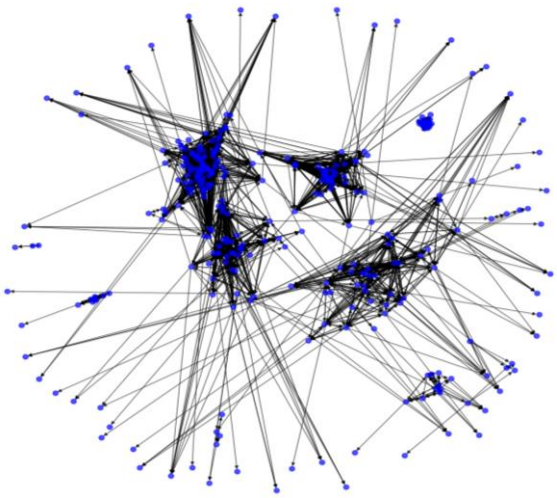
MARIANO MIRANDA SANCHEZ

DATA ENGENIREENG 8°B
UNIVERSIDAD POLITECNICA DE YUCATAN

Introduction

This document aims to show the results obtained from the social network analysis project, with the main focus being the search for the most influential nodes to maximize the spread of information or influence within the selected email dataset. The identification and use of influential nodes can be a significant advantage in terms of company competitiveness. In fields such as marketing, it can be highly useful for campaigns on various topics, whether related to health, technologies, ideas, etc.

The dataset used is an email network, with limited information available about its users, who belong to a departmental sales company, specifically department 1. The nodes, representing the company's users, are located at different points, with emails being sent at specific timestamps indicating their time of sending.



The image shows the nodes and their relationships with each other. Each node represents an individual in the network, and the edges represent the connections between them. At a glance, it is evident that some groups influence more than others, while the remaining groups have little relation or require an information bridge to communicate with one or many groups. The timestamps they possess vary according to the narrow range between nodes, indicating that the data covers short periods of time.

The network has a total of 309 existing nodes and 3031 links between them, with certain specific nodes having many more connections than the average, such as nodes 279, 234, 252, and 299. The nodes are strongly connected with each other, as evidenced by the Average Path and mean distance of 2.19, indicating that most nodes are closely connected. The Network Eccentricity ranges from 3 to 5, with very few nodes having a value of 5. Most nodes theoretically have an eccentricity between 3 and 4. Notably, the nodes with the maximum eccentricity often serve as information bridges between groups or are located at the network's periphery, such as nodes 37, 46, 105, 146, etc. We can deduce that each group likely has 2 to 3 highly influential nodes.

Model

An influence maximization model based on graph theory was used to select a subset of nodes that maximize the spread of influence and information between nodes. The model selects the most influential nodes in successive iterations. In each iteration, the potential impact of each unselected node is evaluated, and those that maximize the influence in each iteration are added to the list of influential nodes. This iterative approach allows for identifying the most suitable nodes.

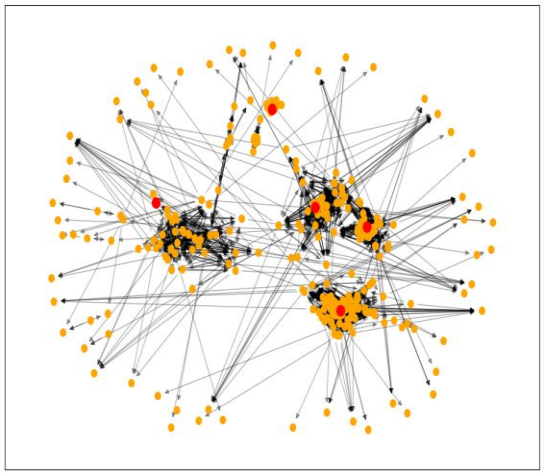
The model iterates a certain number of times with 'k', where 'k' is the selected node. It then evaluates the selected nodes against the unselected ones. The selected nodes are moved to a new list, and the influence is propagated through the network with a predetermined infection probability, calculating how many additional nodes are influenced by each candidate node's selection.

The influence propagation is simulated starting with the initial set of nodes. The model uses an infection probability to determine the spread to neighboring nodes (theoretically beginning with groups with a higher number of nodes until it spreads to the information bridges of peripheral nodes). The importance of the iterative process is to evaluate the potentially influenced neighbors, where the probability might be

around 10%, as an example, which can vary depending on the case.

Results

The model identified the nodes with the greatest influence for maximization: nodes 94, 78, 88, 251, and 103. These nodes are highlighted in red in the visualization, which will be shown later. The remaining nodes in the network, some being directly connected and closer to the influential nodes, are displayed in their usual colors.



The distribution is dense in certain areas, where most information flow is concentrated, with nodes highly interconnected, indicating that there are many communities within the network, while a few nodes are outside these groups. The selection of these influential nodes aims to reduce the time required to influence all nodes. These selected nodes have a large number of connections, facilitating the spread of influence across all groups. The presence of many influenced nodes close to the influential ones indicates that the network's coverage in propagation is maximized.

Having influential nodes in central areas of the groups maximizes the possibility of spreading influence to the rest of the nodes, making it easier and faster to influence multiple nodes simultaneously. The last nodes to be influenced are those at the network's edges.

The expectations for the project were initially low due to limited knowledge of the subject and the anticipated results. Starting the research, the expectations remained low due to the varying nature of different cases and possible outcomes. Upon obtaining and comparing the results with my expectations, there was satisfaction in seeing that the final outcome was much better and clearer, especially with a detailed graph showing the direction of influences established by the nodes during propagation.