Social Network Analysis Report

Valepe Srikrishna

April 2024

PART-I: TO FIND THE LEADER

We have a social network between students. Each student has to fill a maximum of 30 entries at random if they are impressed with the answer to their question. Now, we have this data of all the students and their impressions (maximum of 30), and we have to find the leader of this network with the help of this data.

Let's use the algorithm which made Google a billion dollar industry, called **the PageRank Algorithm**, and find out the leader.

First of all, we have stored all the students whose entry numbers are present in the column with the heading "Email Address". Then we created a graph and added these elements as nodes in the graph. Then, we added the directed edges to the graph according to the data presented in the Excel file. We have at most 30 outgoing edges from each node.

Now, create a dictionary and add all the 143 nodes as keys with their values initialized to zero, so we can keep modifying them later.

Now, perform a random walk starting from any random node. Whenever we pass through a node, we add a value of 1 to that particular key in the dictionary. In case we come across any dead end (i.e., no outgoing edges), we restart the random walk from the beginning. Now, run this process a million times and check all the values. The key with the maximum value is the leader of this network.

By writing a Python code, we found out that the entry number of the leader is **2023CSB1091**.

PART-II: TO FIND THE MISSING LINKS

To find the missing links, first we have to find the adjacency matrix. Then, we have a 143 $\times$ 143 matrix with values either 0 or 1. We consider all the zeroes and follow the procedure below for each zero.

1. Remove the column and row containing the chosen zero to obtain a 142×142 matrix. Call this matrix T.

2. Consider the column C excluding the zero; this matrix contains 142 elements. Similarly, consider the row B without the zero.

3. Now solve for $Tx = B$ and find x. x contains the coefficients of the linear combination.

4. Multiply the column matrix C with x to obtain a 1×1 matrix.

5. If the value in the 1×1 matrix is greater than 1, then replace the zero in the matrix by 1; otherwise keep it as zero.

Apply this procedure for all the zeroes and replace them if the above condition is satisfied.

PART-III: A NEW QUESTION

Question: Find out the *transitivity percentage*. Transitivity percentage is the ratio of the number of transitive cases present in the network to the total number of transitive cases possible (where each node has a maximum of 30 outgoing edges), multiplied by 100.

Let's consider a_1 as the number of transitive cases present. In the Python code, we run a loop and whenever we come across a transitive case, we add 1 to a_1.

By following the above procedure, we get the value of a_1 as 95988.

But we have repeated cases in the above procedure. For example, abc, bca, cab are one and the same. So, we divide the output by 3 to get 31996.

Let's consider a_2 as the total number of possible transitive cases. To find this, we follow a similar procedure: run a loop involving all possible cases and add 1 to a_2.

By doing this, we get $a_2 = 89760$.

$$\text{Transitivity percentage} = \frac{31996}{89760} \times 100 = 35.64$$

Therefore, transitivity percentage = 35.64.