# Minor Project

# Approximating the influential power of nodes in a network using Machine Learning

17103041 Archit Dhiman

17103050 Kashish Mittal

17103055 Agrim Dewan

17103113 Vivek Janjua

Under the guidance of Professor Mayank Gupta.

**Problem Statement -**

Identification of influential nodes in a graph is an important concept which finds many diverse applications in areas such as speeding up the spread of information, controlling spread of rumors and diseases etc. In a social network, nodes are often capable of exerting influence over the other nodes to which they are linked. An influenced node will take on the behavior or characteristic of a linked node. For example, a person who is friends with a number of people who visit a particular news website is more likely to start visiting the website as well. Finding nodes that are highly influential is of great interest to managers and analysts who work with social networks. Marketing managers may want to find influential people to offer them a discount or free product hoping that they will convince their friends to buy the product. Political operatives are also interested in finding these influential people to help them to spread their message.

Work on methods to identify influential nodes in a graph began during the 1950s when Shimbel proposed using the concept of Stress Centrality . He suggested that the centrality of a node should be the total number of shortest paths that go through it and can be used to measure it’s influence. Other methods include Degree Centrality,Eigenvector Centrality (considering the importance of neighbor nodes), Betweeness Centrality and Closeness Centrality. The limitation of these methods is that they require global level information of the graph and thus can’t be scaled to large sized graphs such as those in social networks.Another limitation of these methods is that they need information beforehand and ,therefore, are not able to deal with the dynamic nature of graphs.

Various methods have been utilized to calculate the influence of a node in a graph. Examples include the K-shell index (sometimes refered to as shell number), H-index,clustering coefficient,degree of node etc.

However, K-shell considers only the links between the residual nodes, whereas the links that connect to the exhausted nodes are entirely ignored. In some previous works, approximation of shell-index has been done considering shell-index as a measure of influential power, but the results are not so good.

In this project, we will be studying the process of influence with regard to network communties. Not much work is done concerned with the communities in a network and their structure. Communities are defined by the structure of the links.

The aim of this project is to device a method which estimates the influence of a node in a network by approximating the contribution of the various methods/parameters(devised so far, involved in estimating the influence and some newly defined), and using the local information only with respect to that node, using machine learning.

**Method -**

For building a machine learning model, the data required will be obtained from sources such as <http://snap.stanford.edu/>. SNAP is an initiative( with a library as well of the same name) by Stanford University for network analysis and graph mining. The dataset consists of the various types of networks such as social networks,collaboration networks,web graphs,communication networks etc where some of these are directed graphs while some are undirected. The representation of the networks is in the form of an edge list. Various other parameters such as clustering coefficient etc are also available for the network.

By conducting EDA ( Exploratory Data Analysis) on the dataset and using graph properties, features will be obtained which will be used to train a model to predict the influential power of the nodes and consequently identify most influential nodes within the graph.

This model can be used to demonstrate how a social network forms ,evolves and breaks for which a web interface is also included. For example, the removal of a high influence node from the graph can lead to a disconnected network. This type of phenomenon is important for analysis in various areas such as marketing.

The interface allows for a user to enter relevant information ( similar to a profile system) and get connected to various other people in the graph. The interface informs the user of their own influential power, the most influential person nearest to them and how to connect with that particular person.Initially the user will be connected to a lower influential node (node with degree one) within the graph. The user can send requests to other users to connect with them and increase their own influential power. The acceptance and rejection of a request by a user can be decided by viewing the influential power of the request sender(in the form of a rating system). The interface also shall provide ways to connect with higher influential people by suggesting “paths”(connections to other people) which can raise their rating which can further help in improving their chances of getting connected. The training model will adjust with the dynamic nature of the network such as varying number of users being added and removed etc.