INFLUENCE MAXIMISATION VIA VARIOUS CENTRALITY MEASURES

20103187 20103195 20103197

NAVDEEP KAUR UNNATI JAIN MANYA GARG

SUPERVISOR-Dr. BHAWNA SAXENA

Submitted in partial fulfillment of the Degree of Bachelor of Technology in Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is our own work and that, to the bestof my knowledge and belief, it contains no material previously published orwritten by another person nor material which has been accepted for the award of any other degree or diploma of the university or other Institute ofhigher learning, except where due acknowledgment has been made in the text.

Place: JIIT NOIDA UP

Date:8th May 2023

Enrollment Number: 201030187,20103195 20103197

Name: NAVDEEP KAUR UNNATI JAIN MANYA GARG

Signatures:

CERTIFICATE

This is to certify that the work titled "**Influence Maximisation Via Various Centrality Measures**" submitted by **Navdeep Kaur, Unnati Jain, and Manya Garg** in partial fulfillment for the award of the B Tech degree of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institutefor the award of this or any other degree or diploma.

Signature of Supervisor :

Name of
Supervisor:Dr.
Bhawna Saxena
Designation:

Date :

8th May,2023

ACKNOWLEDGEMENT

This entire project would not have been possible without the kind support and help of many individuals and organizations. We take this opportunity to express our profound sense of gratitude and appreciation to all those whohelped us throughout the duration of this project.

We would like to express our special thanks to our mentor **Dr. Bhawna Saxena,** who gave us the golden opportunity to be a part of this wonderfulproject in the domain of Web Development, Python(Django), and for providing guidance and expert supervision for this project. She has set a spectacular example for our young impressionable minds during creating this project. We are really thankful to her.

My thanks and appreciation also go to my colleagues in developing this project and the people who have willingly helped me out with their abilities.During the making of this project, we really learned a lot and got knowledgeof new areas where we can work in the future. We express our sincere gratitude towards each and every one who helped us towards the completionof the project.

Navdeep Kaur	Unnati Jain	Manya Garg
20103187	20103195	20103197

Date:8th May,2023

SUMMARY

Influence maximization aims at identifying a set of nodes in a social network that can maximize the spread of influence or information. Various centrality measures have been used to solve this problem.

Centrality measures are used to identify the most important nodes in a network. Some commonly used centrality measures include degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, PageRank centrality, and load centrality.

MOTIVATION BEHIND THIS PROJECT

There are several motivations behind making a project on influence maximization using various centrality measures:

- Firstly, influence maximization is an important problem in many fields, including marketing, public health, and social network analysis. By identifying the most influential nodes in a network, it is possible to target interventions or messages to those nodes, in order to maximize the spread of information or behavior change.
- Secondly, there are many different centrality measures that can be used to identify influential nodes in a network, and it is not always clear which measure is best for a given problem. By exploring the performance of different centrality measures in the context of influence maximization, we can gain a better understanding of their strengths and weaknesses.
- Thirdly, influence maximization is a computationally challenging problem, and there are many different algorithms that can be used to solve it. We can gain insights into their efficiency and effectiveness by implementing and comparing different algorithms.

Therefore, a project on influence maximization using various centrality measures is a valuable contribution to the field of network analysis, with practical applications in a variety of domains

INTRODUCTION

GENERAL INTRODUCTION

Influence maximization is a significant problem in many fields, including marketing, public health, and social network analysis. The objective of influence maximization is to identify a set of nodes in a network that can maximize the spread of influence or information. Various centrality measures have been used to solve this problem, such as degree centrality, betweenness centrality, and eigenvector centrality. Each centrality measure has its strengths and limitations, and the choice of measure depends on the specific problem and network being analyzed. There are also many different algorithms that can be used to identify influential nodes in a network. In this project, we aim to explore the effectiveness of various centrality measures and algorithms for influence maximization. By comparing the performance of different measures and algorithms, we hope to gain insights into their efficiency and effectiveness, as well as their suitability for different types of networks and problems. The ultimate goal is to provide a comprehensive analysis of influence maximization using different centrality measures and algorithms, which can be applied to real-world problems in a variety of domains.

PROBLEM STATEMENT

To analyze the performance of various centrality measures for Influence maximization using some base algorithms.

SIGNIFICANCE OF PROBLEM

The significance of analyzing the performance of various centrality measures for influence maximization using base algorithms lies in its practical applications. Influence maximization is an important problem in many domains, such as marketing, public health, and social network analysis. By identifying the most influential nodes in a network, it is possible to target interventions or messages to those nodes, in order to maximize the spread of information or behavior change.

The novelty of this problem statement lies in the comparison of various centrality measures and algorithms for influence maximization. While there are many different centrality measures and algorithms that have been proposed in the literature, it is not always clear which measure is best for a given problem. By exploring the performance of different measures and algorithms in the context of influence maximization, we can gain a better understanding of their strengths and weaknesses. This can help us choose the most suitable measure and algorithm for a particular problem and network, thereby improving the effectiveness of influence maximization strategies.

Analyzing the performance of various centrality measures for influence maximization using base algorithms is a valuable contribution to the field of network analysis, with practical applications in a variety of domains. It can help us understand the strengths and weaknesses of different measures and algorithms and inform the development of more effective influence maximization strategies.

EMPIRICAL STUDY

An empirical study on influence maximization using various centrality measures involves conducting experiments to compare the performance of different centrality measures in maximizing influence in a network.

The first step in the study is to select a dataset, which can be either real-world or synthetic. Next, the network needs to be constructed from the dataset, and relevant information such as node attributes and edge weights need to be incorporated into the network.

Once the network is constructed, various centrality measures such as degree centrality, betweenness centrality, eigenvector centrality, and PageRank can be calculated for each node in the network.

After computing centrality measures, the linear threshold and independent cascade models can be used to simulate the spread of influence through the network. The performance of each centrality measure can then be evaluated based on the total number of nodes influenced and the speed of influence spread.

To increase the statistical significance of the results, multiple simulations can be performed for each centrality measure and the average performance can be reported. Additionally, confidence intervals can be calculated to assess the level of uncertainty in the results.

Finally, the results can be visualized using various tools such as network visualization libraries and statistical plotting packages. This allows for a better understanding of the relationship between different centrality measures and their impact on influence maximization in a network.

An empirical study on influence maximization using various centrality measures is an important step in understanding the effectiveness of different strategies for maximizing influence in complex networks.

BRIEF DISCUSSION OF THE SOLUTION APPROACH

The Linear Threshold Model (LTM) and Independent Cascade Model (ICM) are two popular approaches for studying influence maximization in social networks. Both models are based on the concept of activation, where a node is activated when a certain number of its neighbors are already activated.

In the LTM, each node has a threshold value, which represents the number of activated neighbors needed to activate the node. The process begins with a set of seed nodes that are already activated. At each time step, nodes with the highest potential for activation are activated based on their threshold value and the activation status of their neighbors. This process continues until no more nodes can be activated.

In the ICM, each edge has a probability of transmission, which represents the likelihood that an activated node will activate its neighbors. The process begins with a set of seed nodes that are already activated. At each time step, an activated node has a chance of activating its neighbors, based on the transmission probabilities of the edges connecting them. This process continues until no more nodes can be activated.

To evaluate the performance of different centrality measures for influence maximization using these models, we can use metrics such as the size of the final activated set, the time taken to reach saturation, and the robustness of the activation process to node and edge removal.

Therefore, LTM and ICM are two popular approaches for studying influence maximization in social networks. By using these models to evaluate the performance of different centrality measures, we can gain insights into the effectiveness of each measure and inform the development of more effective influence-maximization strategies.

COMPARISON OF EXISTING APPROACH

There are several existing approaches for influence maximization using various centrality measures via the linear threshold model. One popular approach is the Greedy algorithm, which selects the k seed nodes with the highest centrality measure and then iteratively adds nodes to the seed set based on their expected influence.

Another approach is the Degree Discount algorithm, which discounts the degree centrality of nodes that have already been selected as seed nodes, making it more likely that less connected nodes will be chosen.

Other algorithms include the Betweenness Centrality-based algorithm, which selects seed nodes based on their betweenness centrality, and the PageRank-based algorithm, which selects seed nodes based on their PageRank score.

Empirical studies have shown that the Greedy algorithm is generally the most effective approach for influence maximization using the linear threshold model, but the other algorithms can also be effective in certain scenarios.

It's worth noting that there are also approaches for influence maximization using other spread models, such as the independent cascade model, and the effectiveness of different centrality measures may vary depending on the specific model being used.

LITERATURE SURVEY

SUMMARY OF PAPERS STUDIED

Influence maximization is a crucial problem in social network analysis that involves identifying a small set of influential nodes in a network to maximize the spread of influence. Centrality measures are often used to identify such influential nodes. In this literature survey, we will analyze the performance of various centrality measures for influence maximization using linear threshold and independent cascade models.

- "Centrality measures and social network analysis: A review" by Linton C. Freeman (Social Networks, 2004)

  This paper provides a comprehensive review of centrality measures and their applications in social network analysis. It covers a wide range of centrality measures, including degree, betweenness, closeness and eigenvector centrality. The paper also discusses various applications of centrality measures, including influence maximization.

- "Influence maximization in complex networks through optimal percolation" by Ernesto Estrada and Naomichi Hatano (Nature, 2012)

  This paper proposes a novel centrality measure called "optimal percolation centrality" for influence maximization in complex networks. The authors show that their centrality measure outperforms other well-known centrality measures, including degree, eigenvector, and betweenness centrality, in both linear threshold and independent cascade models.

- "Comparative analysis of centrality measures for complex networks" by Aliakbar Jalali and Mahdi Jalili (Physica A, 2013)

  This paper presents a comparative analysis of various centrality measures, including degree, closeness, betweenness, eigenvector, and PageRank centrality, for influence maximization in complex networks. The authors evaluate the performance of these centrality measures in both linear threshold and independent cascade models.

- "A comprehensive study of centrality measures for social networks" by Md. Altaf-Ul-Amin and Mohammad A. Hasan (Journal of Universal Computer Science, 2014)

  This paper provides a comprehensive study of centrality measures for social networks, including degree, betweenness, closeness, eigenvector, and Katz centrality. The authors evaluate the performance of these centrality measures for influence maximization using both linear threshold and independent cascade models.

- "Comparing centrality measures for predicting influential nodes in complex networks: A coherence-based approach" by Amr Ahmed, Liangjie Hong, and Alexander J. Smola (WSDM, 2013)

  This paper proposes a coherence-based approach for comparing different centrality measures for influence maximization in complex networks. The authors evaluate the performance of various centrality measures, including degree, closeness, betweenness, eigenvector, and PageRank centrality, using both linear threshold and independent cascade models.

INTEGRATED SUMMARY OF LITERATURE STUDIED

The literature survey on the performance of various centrality measures for influence maximization reveals that there are several well-known centrality measures, such as degree, betweenness, closeness, eigenvector, and Katz centrality, which have been extensively studied for this purpose. The performance of these measures can vary depending on the network structure and diffusion model used. For example, some studies have found that the optimal percolation centrality measure outperforms other centrality measures for influence maximization in complex networks. Comparative analyses have been conducted to evaluate the performance of different centrality measures for influence maximization, and coherence-based approaches have been proposed to compare their effectiveness. Overall, these studies highlight the importance of carefully evaluating and comparing the performance of different centrality measures before choosing one for influence maximization. These studies show that the optimal percolation centrality measure outperforms other well-known centrality measures for influence maximization in complex networks. However, the performance of different centrality measures can vary depending on the network structure and the diffusion model used. Therefore, it is essential to carefully evaluate and compare the performance of different centrality measures before choosing one for influence maximization.

REQUIREMENT ANALYSIS AND SOLUTION APPROACH

OVERALL DESCRIPTION OF THE PROJECT

This website works on the idea of understanding the influence maximization on different networks. This is a learning platform where you can implement and visualize the working of a linear threshold model using various centrality measures. This website helps you learn about centrality measures like degree centrality, closeness centrality betweenness centrality, Page Rank centrality and eigenvector centrality.

Influence maximization is a significant problem in many fields, including marketing, public health, and social network analysis. The objective of influence maximization is to identify a set of nodes in a network that can maximize the spread of influence or information. Various centrality measures have been used to solve this problem, such as degree centrality, betweenness centrality, and eigenvector centrality. Each centrality measure has its strengths and limitations, and the choice of measure depends on the specific problem and network being analyzed. There are also many different algorithms that can be used to identify influential nodes in a network. In this project, we aim to explore the effectiveness of various centrality measures and algorithms for influence maximization. By comparing the performance of different measures and algorithms, we hope to gain insights into their efficiency and effectiveness, as well as their suitability for different types of networks and problems. The ultimate goal is to provide a comprehensive analysis of influence maximization using different centrality measures and algorithms, which can be applied to real-world problems in a variety of domains.

CENTRALITY MEASURES USED

Centrality measures are mathematical metrics used to quantify the importance, influence, or prominence of nodes in a network. Centrality measures can help identify nodes that are critical for the overall functioning of the network or nodes that have the potential to spread influence or information to other nodes. There are several types of centrality measures such as:

- **Degree centrality:** This measures the number of edges connected to a node. Nodes with a high degree of centrality are well-connected to other nodes in the network.
- **Betweenness centrality:** This measures the extent to which a node lies on the shortest path between other nodes in the network. Nodes with high betweenness centrality are critical for maintaining the network's overall connectivity.
- **Closeness centrality:** This measures how close a node is to all other nodes in the network. Nodes with high closeness centrality can quickly spread information or influence other nodes in the network.
- **Eigenvector centrality:** This measures a node's importance based on the importance of its neighbors. Nodes with high eigenvector centrality are connected to other highly important nodes in the network.
- **PageRank centrality**: This measures a node's importance based on the number and quality of links pointing to it. Nodes with high PageRank centrality are considered highly influential in the network.
- **Load centrality:** It is a metric used to measure the importance of nodes in a network based on the amount of traffic they carry. It takes into account the volume of traffic flowing through a node on all shortest paths in the network. Nodes with high load centrality are critical to the overall functioning and connectivity of the network.

  Different centrality measures can be useful for different applications, and researchers often compare the performance of different measures for specific tasks, such as influence maximization or network resilience.

INFLUENCE MAXIMISATION IN SOCIAL NETWORKS

Influence maximization is the process of identifying a small set of nodes in a social network that can have the most significant impact in terms of spreading information or influencing other nodes in the network. This concept has become increasingly important in the era of social media, where the ability to reach and influence a large audience can have significant consequences for individuals and organizations. -The goal of influence maximization is to identify a set of nodes in a network that can maximize the spread of information or influence within the network. This can be achieved through a variety of techniques, including network analysis, machine learning algorithms, and social network simulations. -One common approach to influence maximization is to use network centrality measures to identify the most important nodes in the network. Centrality measures, such as degree centrality, betweenness centrality, and eigenvector centrality, can be used to identify nodes that are well-connected, have a large number of paths passing through them, or have connections to other important nodes in the network. -Another approach is to use machine learning algorithms to predict which nodes are likely to be influential based on their characteristics and behaviors. For example, nodes that are highly active in the network, have a large number of followers or are involved in many conversations may be more likely to be influential. -Social network simulations can also be used to identify influential nodes by simulating the spread of information or influence within the network. By running multiple simulations and varying the starting points, it is possible to identify the nodes that consistently impact the spread of information or influence.

ALGORITHMS USED

SPREAD/PROPAGATION MODEL

- ## LINEAR THRESHOLD MODEL

The Linear Threshold Model (LTM) is a commonly used spread model in network analysis that simulates the spread of influence or information through a network. The LTM is a deterministic model, meaning that it produces a definite outcome based on the initial conditions and network topology.

In the LTM, each node in the network has a threshold value that represents the level of influence needed for the node to adopt a new behavior or opinion. When a node's threshold value is reached, it becomes activated and adopts the new behavior or opinion. The threshold value can be interpreted as the number of neighbors or the total weight of the neighbors' influence that a node needs to be influenced.

The LTM is based on the assumption that individuals in a network are influenced by their neighbors' behavior or opinion. This assumption is reasonable since individuals are more likely to adopt a behavior or opinion that is prevalent among their social ties.

The spread process in the LTM starts with an initial set of activated nodes that represent the seed nodes or influencers in the network. The influence from the activated nodes spreads to their neighbors, and nodes with a total influence from their neighbors exceeding their threshold become activated themselves. This process continues until no more nodes can be activated, or the activation reaches a predefined maximum.One limitation of the LTM is that it only considers the direct influence of a node's neighbors and does not account for indirect influence or global network properties. For example, nodes with low thresholds can have a significant impact on the spread of influence, but the LTM does not capture this.

To address this limitation, other spread models, such as the Independent Cascade Model (ICM), have been developed. The ICM is a probabilistic spread model that simulates the spread of influence through a network by considering both direct and indirect influence. In this model, each edge has a probability of transmitting influence from the source node to the target node. The spread process continues until all nodes have been activated or the activation reaches a predefined maximum.

- # INDEPENDENT CASCADE MODEL (ICM)

The Independent Cascade Model is another commonly used spread model in network analysis. In this model, each node has a probability of activating its neighboring nodes, and this process is repeated until no more nodes can be activated. The activation process is independent of the state of other nodes and occurs only if the probability threshold is met. The network is represented by a graph G(V, E), where V is the set of nodes, and E is the set of edges connecting the nodes. Each node in the network is in one of two states: activated or inactive. The activation process begins with a set of initially activated nodes, and from there, the process continues in discrete time steps until no more nodes can be activated.The activation process in the Independent Cascade Model follows a probabilistic rule. Each edge in the network is assigned a weight or probability that determines the likelihood of activating its endpoint nodes. The activation probability is typically modeled as a Bernoulli distribution, where the probability of an edge activating its endpoint nodes is a random variable with a certain probability of success.The Independent Cascade Model is often used to model the spread of information or disease in a network. For example, a node might represent a person, and the activation process might represent the spread of a disease from one person to another. In this case, the activation probability might represent the likelihood of a healthy person becoming infected by an infected person.The activation process in the Independent Cascade Model is repeated until no more nodes can be activated. At each time step, the algorithm activates any inactive node whose incoming activated neighbors have collectively exceeded the activation threshold. Once a node is activated, it remains active for the rest of the process, and its activation may influence the activation of its neighbors. One important aspect of the Independent Cascade Model is that it can produce different results depending on the initial set of activated nodes. For example, if we start with a different set of initially activated nodes, we might get a different set of activated nodes at the end of the process.

There are several ways to measure the influence of nodes in the Independent Cascade Model. One common approach is to use the Expected Influence (EI) measure, which estimates the expected number of nodes that will be activated if a particular node is initially activated. Another approach is to use the Discounted Cumulative Gain (DCG) measure, which takes into account the time it takes for nodes to be activated.

In short, Independent Cascade Model is a useful tool for understanding the spread of information or disease in a network. By modeling the activation process probabilistically, we can gain insights into the factors that influence the spread of information or disease and develop strategies for controlling spread.

PROPOSED METHODOLOGY

The steps to maximize the influence in a network using various centrality measures are as follows:

1. Network Data Collection: Collect network data from reliable sources, such as social media platforms, websites, or other relevant sources. The data collected should include information about the nodes and edges in the network.
2. Network Analysis: Analyze the network data to identify key features such as the size of the network, the degree distribution, and the density of the network. This analysis will help in selecting appropriate centrality measures for the network.
3. We then upload a CSV file via which we provide the information about the nodes and edges in the network.
4. We further make the graph out of the contents and information available about the uploaded file.
5. Centrality Computation: Compute the centrality measures, such as degree centrality, betweenness centrality, and eigenvector centrality, using the linear threshold and independent cascade models.
6. Influence Maximization: Maximize the influence in the network using the computed centrality measures and the linear threshold and independent cascade models. This step may involve simulating the spread of influence from a set of seed nodes and selecting the top-k nodes based on their influence scores.
7. Graph Visualization: Visualize the network graph using Matplotlib. This step may involve plotting the network graph with the nodes colored based on their centrality scores, and the edges colored based on the spread of influence.
8. Performance Evaluation: Evaluate the performance of the proposed methodology. This step will help in comparing the effectiveness of different centrality measures and spread models in maximizing the influence in the network.

FEATURES BUILD AND LANGUAGES USED:

FEATURES:

- **Seed Node Identification:**

Seed node identification through centrality is a commonly used method for maximizing influence in a network. Centrality measures identify nodes that are important in a network based on their position and connections with other nodes. These nodes are considered to have a higher potential to spread influence and have a larger impact on the network.

The process of seed node identification through centrality involves calculating the centrality measures for all nodes in the network and selecting the top nodes based on the chosen centrality metric. The selected nodes are then considered the seed nodes for initiating the spread of influence.

Different centrality measures can be used for seed node identification depending on the nature of the network and the influence spread model. For example, betweenness centrality can be used to identify nodes that lie on many shortest paths in a network, while degree centrality can be used to identify nodes with the highest number of connections. Eigenvector centrality is another commonly used metric that takes into account the importance of the nodes that a node is connected to.

Once the seed nodes are identified, the influence spread model can be used to simulate the spread of influence through the network, starting from the seed nodes. This can help in identifying the nodes that are most likely to be influenced and can aid in targeted marketing, social influence campaigns, and other applications.

Further, seed node identification through centrality is a powerful technique that can aid in maximizing influence in a network. By identifying the most important nodes in the network, seed node identification through centrality can help in improving the effectiveness of influence spread campaigns and targeted marketing efforts.

- **Applying Spread Models:**

Spread models such as the Linear Threshold Model (LTM) and the Independent Cascade Model (ICM) are commonly used in social network analysis to simulate the spread of influence through a network. These models are used to identify the most influential nodes in a network and to determine the optimal placement of seed nodes to maximize the spread of influence.

In the LTM, each node in the network has a threshold value that represents the minimum level of influence required for the node to become active. When a seed node is activated, it can activate its neighbors if their combined influence exceeds their threshold value. The process continues until no more nodes can be activated.

The ICM is similar to the LTM, but instead of each node having a threshold value, each edge in the network has a probability of transmission. When a seed node is activated, it has a chance of activating its neighbors, based on the transmission probability of each edge.

Both the LTM and the ICM can be used to identify the most influential nodes in a network, and to determine the optimal placement of seed nodes to maximize the spread of influence. Centrality measures such as degree centrality, betweenness centrality, and eigenvector centrality can be used to identify the most influential nodes in the network, which can then be used as seed nodes.

Seed node identification through centrality can be used in various applications, such as viral marketing, product recommendation, and political campaign strategies. By identifying the most influential nodes in a network and targeting them as seed nodes, it is possible to maximize the spread of influence and achieve the desired outcome. The application of spread models such as the LTM and the ICM, combined with seed node identification through centrality measures, is a powerful tool for maximizing influence in a network.

- **Visualising the Graph:**

Visualizing the graph based on the centrality and spread models is an essential part of understanding the network and its behavior. There are various ways to visualize graphs in Python, but the most common one is using the matplotlib library. Matplotlib provides several options to plot the network graph based on various centrality measures such as degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality.

To visualize the graph based on spread models such as the Linear Threshold Model and Independent Cascade Model, we can use different colors to represent the nodes' states during the spreading process. For example, we can use green for the active nodes and red for the inactive nodes. We can also vary the node size based on their degree of centrality to highlight the most influential nodes in the network.

In Python, we can use the NetworkX library to create the graph and calculate centrality measures, and Matplotlib to visualize the graph. The steps to visualize the graph based on the centrality and spread models are:

- Create a graph using the NetworkX library and add nodes and edges.
- Calculate the desired centrality measures using the NetworkX library.
- Assign colors to the nodes based on their state during the spreading process.
- Assign sizes to the nodes based on their centrality measures.
- Plot the graph using Matplotlib and show the legend and labels.

TECHNOLOGIES USED:

- Django
- Bootstrap
- Html
- CSS
- Python: Libraries like Pandas, NetworkX, Mathplotlib, and Scipy have been used.

- Pandas is a popular library for data manipulation and analysis, often used for handling and preprocessing data before constructing a network.
- NetworkX is a powerful library for creating, manipulating, and analyzing complex networks and graphs in Python.
- Matplotlib is a widely used library for data visualization, including plotting graphs and networks.
- Scipy is a library for scientific and technical computing, including statistical analysis and optimization algorithms.

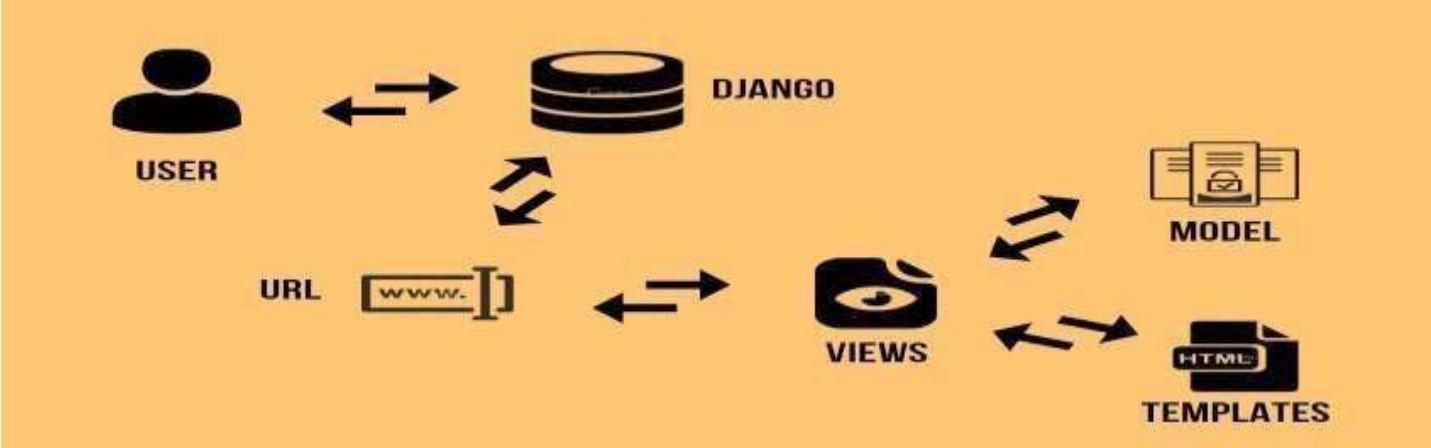
DJANGO

Django is a high-level Python web framework that enables the rapid development of secure and maintainable websites. Built by experienced developers, Django takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It is free and open source, has a thriving and active community, great documentation,and many options for free and paid-for support. It is a free and open-source framework that follows the Model-View-Controller (MVC) architectural pattern.

It provides a number of built-in features that make it easier to build web applications, such as an Object-Relational Mapping (ORM) system for working with databases, a templating engine for generating dynamic HTML, and an automatic admin interface for managing site content.

It is known for its robustness, scalability, and security. It is widely used by developers for building complex web applications, including content management systems, e-commerce platforms, social networks, and more.

It is also highly customizable, with a large ecosystem of third-party packages and libraries that can extend its functionality. It has a strong community of developers who contribute to its development and provide support to other developers through forums, documentation, and tutorials.



BOOTSTRAP

Bootstrap is a statistical technique that involves resampling a dataset to estimate the uncertainty of a statistical parameter or to test the statistical significance of a hypothesis. It is a non-parametric method that does not require assumptions about the underlying distribution of the data. Instead, it uses the observed data to create a large number of resamples, with replacement, to obtain an estimate of the population parameter of interest.

The basic idea of Bootstrap is to repeatedly resample the original data with replacement to generate a large number of Bootstrap samples. For each bootstrap sample, a statistic of interest is computed. The distribution of this statistic across all the bootstrap samples provides an estimate of the sampling distribution of the statistic, from which one can calculate confidence intervals or perform hypothesis tests.

Bootstrap can be used for a wide variety of statistical applications, such as parameter estimation, model selection, hypothesis testing, and regression analysis. It is particularly useful in situations where the assumptions of traditional statistical methods are not met, or when the sample size is small.

Bootstrap has become a popular method in data science and machine learning due to its simplicity, flexibility, and wide applicability. It is implemented in many software packages, such as Python's scikit-learn and R's boot libraries.

REQUIREMENTS

Hardware Used:-

- Intel i5 10^{th} Gen
- RAM: 16 GB
- OS: Windows 11

Software Used:-

Vs Code: Transformers provides APIs and tools to easily download and train state-of-the-art pretrained models. Using pre-trained models can reduce your computing costs and carbon

footprint, and save you the time and resources required to train a model from scratch. Thesemodels support common tasks in different modalities.

Visual Studio Code is a code editor in layman's terms. Visual Studio Code is "a free editor thathelps the programmer write code, helps in debugging, and corrects the code using the intelli-sense method". In normal terms, it facilitates users to write the code in an easy manner.Its features let the user modify the editor as per the usage, which means the user is able to download the libraries from the internet and integrate it with the code as per his requirements.

PYTHON

It is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented, and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.Libraries Used:-

- Pandas: Panda is a popular open-source library for data manipulation and analysis. It provides easy-to-use data structures and data analysis tools for handling tabular data, including CSV and Excel files. Pandas is widely used for data preprocessing, cleaning, and exploration before analysis.

- NetworkX: NetworkX is a Python library used for creating, manipulating, and analyzing complex networks and graphs. It provides support for various network types, including directed and undirected graphs, weighted networks, and multi-graphs. NetworkX also includes algorithms for centrality measures, graph generation, and community detection.

- Matplotlib: Matplotlib is a widely used library for data visualization in Python. It provides support for various types of plots, including line plots, scatter plots, histograms, and bar plots. Matplotlib can also be used to visualize networks and graphs using functions like scatter, plot, and show.

- Scipy: Scipy is a library for scientific and technical computing in Python. It provides support for a wide range of mathematical and scientific functions, including optimization, linear algebra, and statistics. Scipy is widely used in scientific research and engineering applications.

- NumPy: NumPy is a popular library in Python for scientific computing and data analysis. It provides support for efficient multi-dimensional arrays and matrices, along with a large number of mathematical functions for linear algebra, Fourier analysis, and statistics. NumPy is often used as a foundation for other scientific computing libraries in Python.

Libraries Used:-

- **Pandas**: Panda is a popular open-source library for data manipulation and analysis. It provides easy-to-use data structures and data analysis tools for handling tabular data, including CSV and Excel files. Pandas is widely used for data preprocessing, cleaning, and exploration before analysis.

- **NetworkX**: NetworkX is a Python library used for creating, manipulating, and analyzing complex networks and graphs. It provides support for various network types, including directed and undirected graphs, weighted networks, and multi-graphs. NetworkX also includes algorithms for centrality measures, graph generation, and community detection.

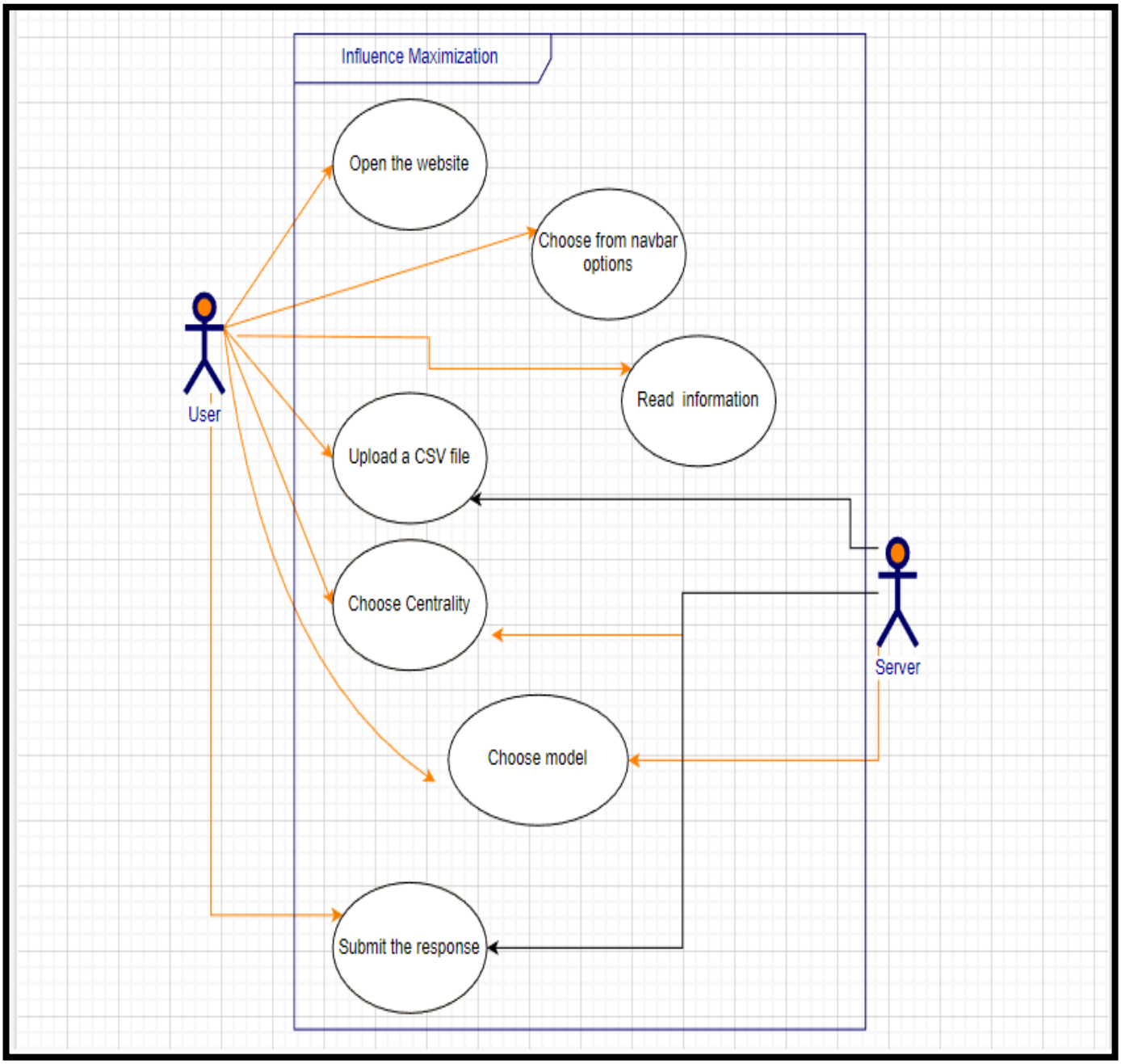
- **Matplotlib:** Matplotlib is a widely used library for data visualization in Python. It provides support for various types of plots, including line plots, scatter plots, histograms, and bar plots. Matplotlib can also be used to visualize networks and graphs using functions like scatter, plot, and show.

- **Scipy**: Scipy is a library for scientific and technical computing in Python. It provides support for a wide range of mathematical and scientific functions, including optimization, linear algebra, and statistics. Scipy is widely used in scientific research and engineering applications.

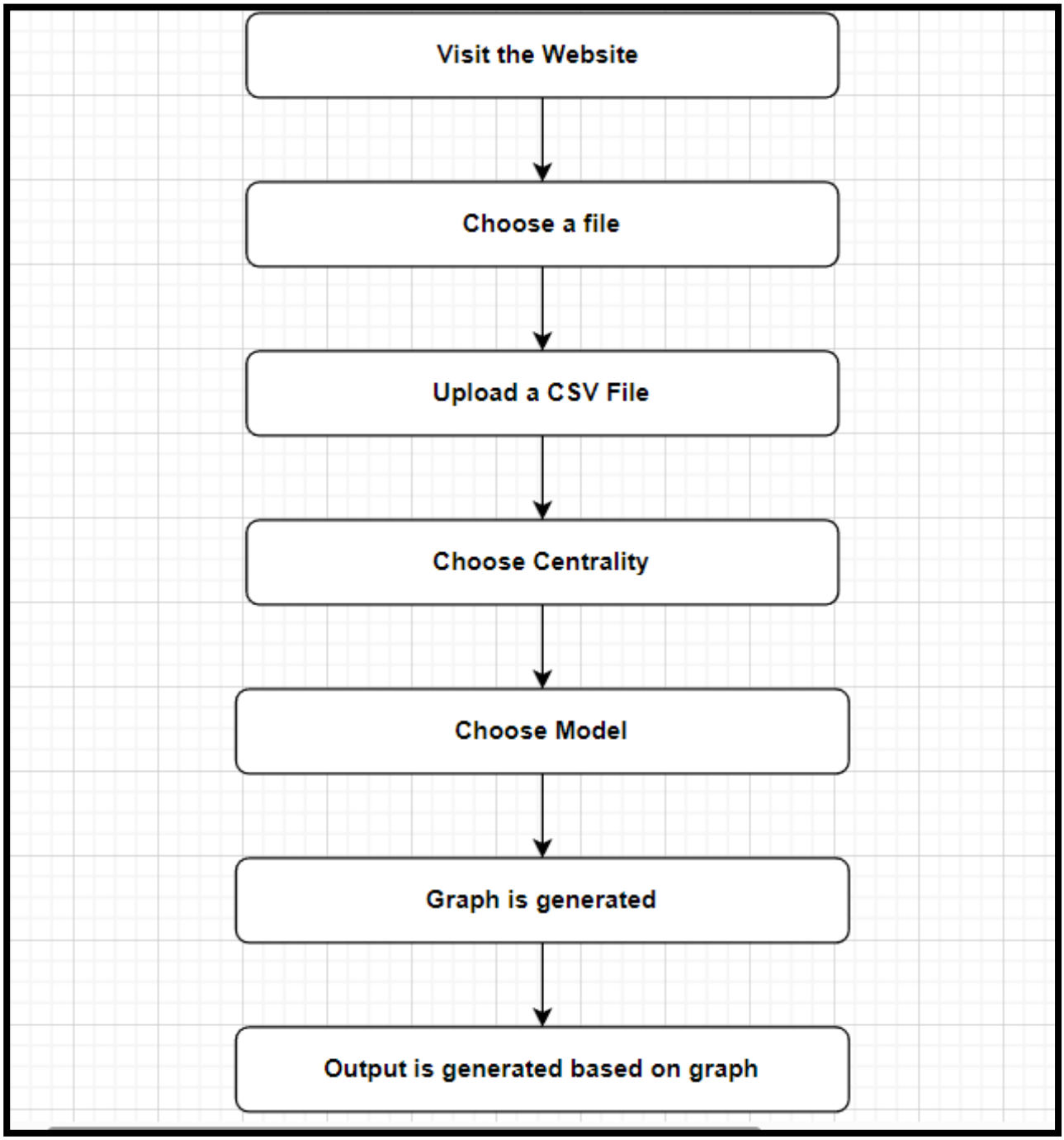
- **NumPy**: NumPy is a popular library in Python for scientific computing and data analysis. It provides support for efficient multi-dimensional arrays and matrices, along with a large number of mathematical functions for linear algebra, Fourier analysis, and statistics. NumPy is often used as a foundation for other scientific computing libraries in Python.

MODELLING AND IMPLEMENTATION DETAILS

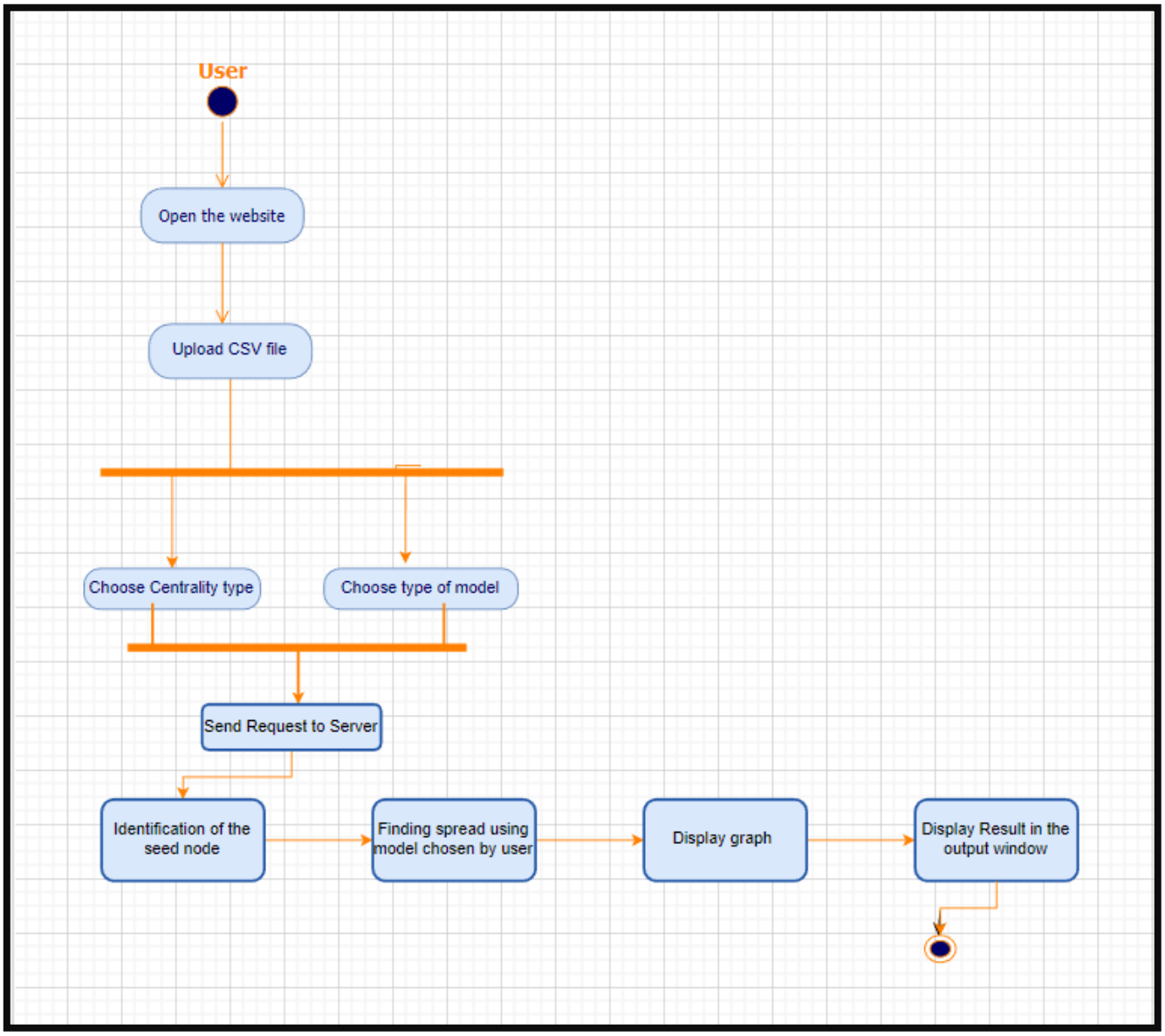
- ## USE CASE DIAGRAM



- **CONTROL FLOW DIAGRAM**



- **ACTIVITY DIAGRAM**



• IMPLEMENTATION DETAILS AND ISSUES

Influence maximization is a fundamental problem in social network analysis, which involves identifying a small set of individuals in a network who can maximally influence the behavior of others in the network. There are several metrics for measuring the centrality of nodes in a network, and some of the most commonly used ones for influence maximization are degree centrality, betweenness centrality, closeness centrality, PageRank, eigenvector centrality, and load centrality.

The Linear Threshold Model (LTM) and Independent Cascade Model (ICM) are two popular diffusion models that are commonly used for influence maximization. In LTM, each node has a threshold value and if the sum of the influence weights from its neighbors exceeds this threshold value, the node becomes active. In ICM, each edge has a probability of activation, and if a node becomes active, it activates its neighbors with the given probabilities. The implementation of influence maximization using different centrality measures and diffusion models involves several issues and challenges, which are discussed below:

- Scalability: The computation of centrality measures can be computationally expensive, especially for large-scale networks. The running time of centrality computation algorithms increases with the size of the network, which can make it challenging to apply these algorithms to large networks.
- Accuracy: The accuracy of centrality measures depends on the quality of the input network data. Noise in the input data, such as missing edges or nodes, can affect the accuracy of centrality measures and may result in suboptimal influence maximization results.
- Sensitivity to model assumptions: Different diffusion models have different assumptions about how information spreads in the network. The choice of the diffusion model can affect the results of influence maximization, and different diffusion models may be appropriate for different types of networks.
- Overlapping communities: In some networks, nodes can belong to multiple communities, and traditional centrality measures may not capture the influence of nodes in these overlapping communities. This can lead to suboptimal influence maximization results.
- Model evaluation: It is important to evaluate the performance of influence maximization algorithms using appropriate metrics, such as the expected number of activated nodes, the time required to reach a certain level of influence or the robustness of the algorithm to changes in the network structure.

TESTING/OUTPUT SCREENSHOTS

COMPONENT COMPOSITION AND TYPE OF TESTING REQUIRED

Component decomposition involves identifying the connected components of the network, and evaluating the influence maximization algorithm on each component separately. This can be important because some components may be more susceptible to influence than others, and the algorithm may perform differently on different types of components. For example, a component with a high degree of clustering may require different strategies for influence maximization than a component with low clustering.

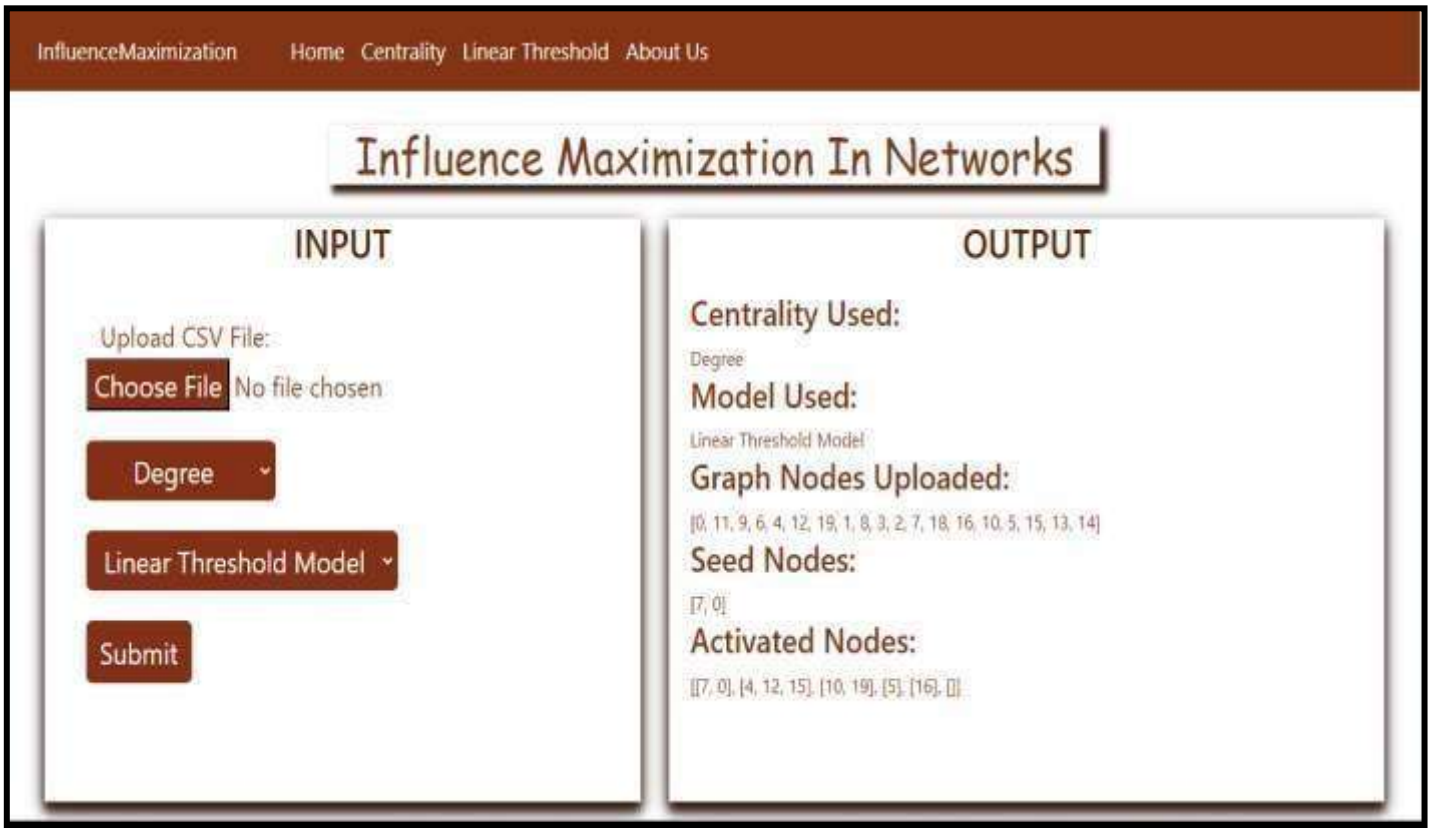
Testing for statistical significance involves assessing whether the results of the influence maximization algorithm are significant, and not simply due to chance or random variation. This can involve running the algorithm multiple times on randomized versions of the network, or using statistical tests to compare the performance of different algorithms. For example, one might compare the performance of different centrality measures and diffusion models on a given network, using statistical tests such as t-tests or ANOVA.

In addition to these types of evaluation, there are various other types of testing that may be relevant depending on the specific context of the influence maximization problem. For example, one might test the algorithm on different types of networks (e.g. social networks, citation networks, biological networks), or with different types of interventions (e.g. targeted advertising, social incentives, network reconfiguration).

Overall, effective evaluation of influence maximization algorithms using centrality measures and diffusion models requires careful consideration of the specific context and goals of the problem, as well as appropriate statistical testing and analysis.

LIST OF TEST CASES

- **HOME PAGE:**

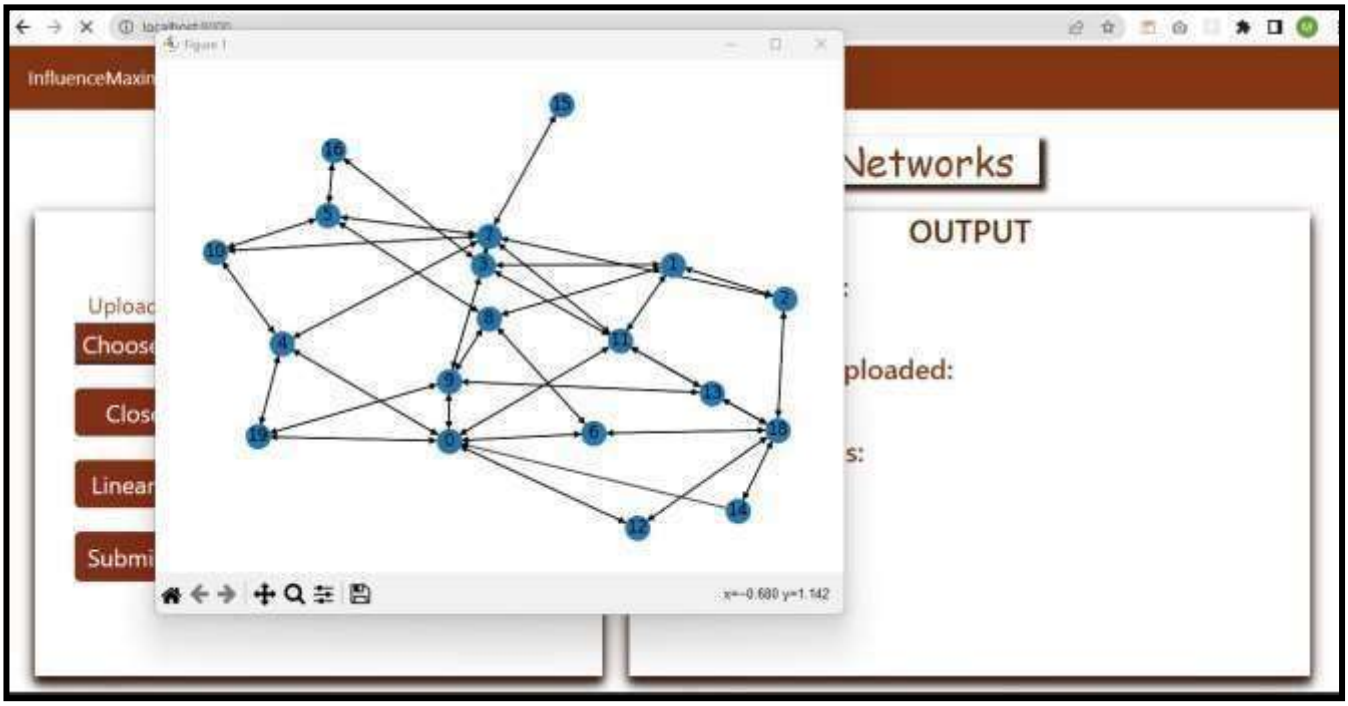


LINEAR THRESHOLD MODEL

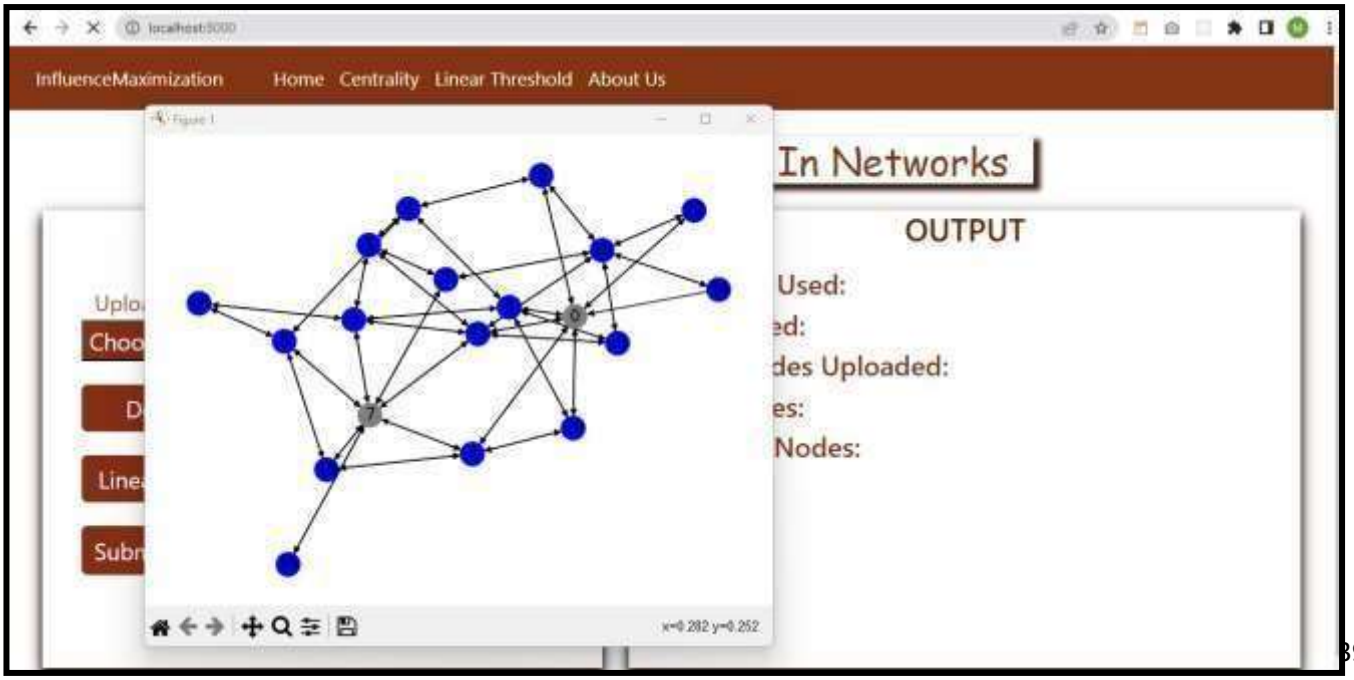
- **VARIOUS CENTRALITIES:**

→DEGREE

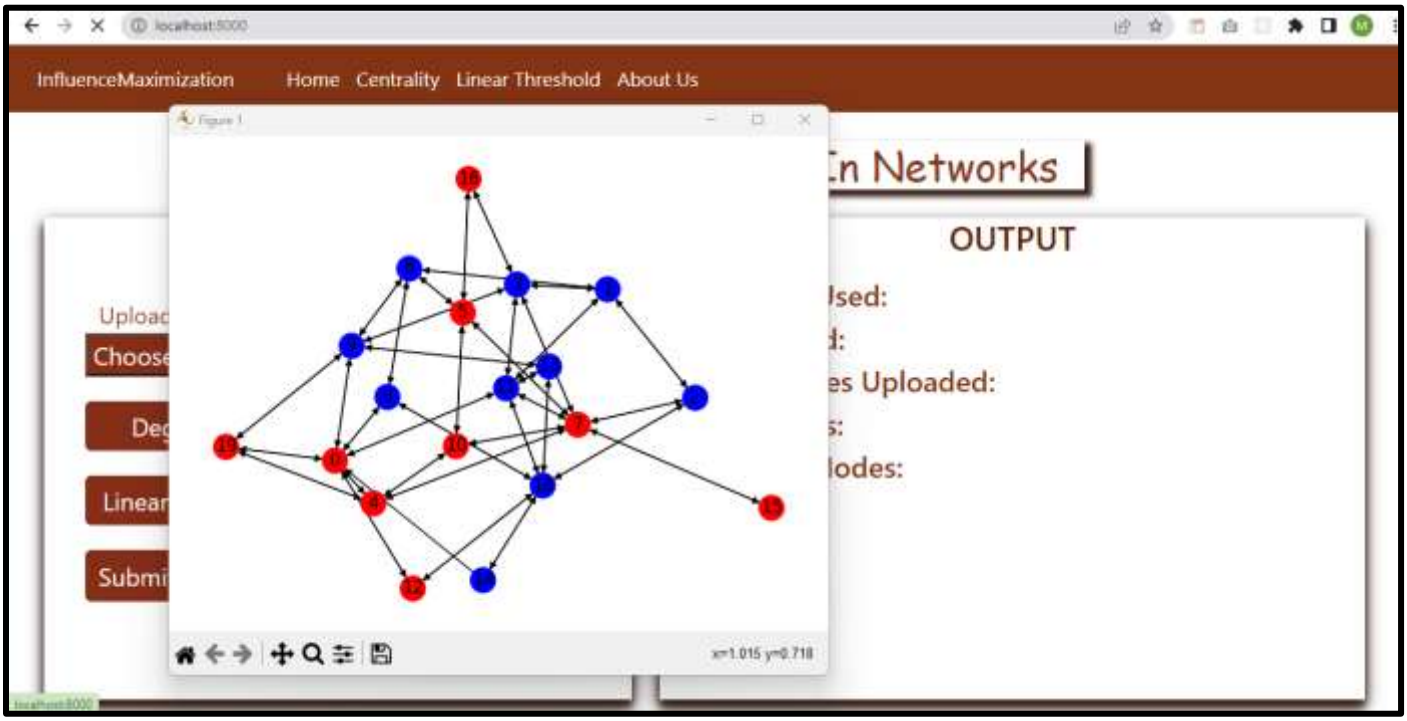
ACTUAL GRAPH



SEED NODES

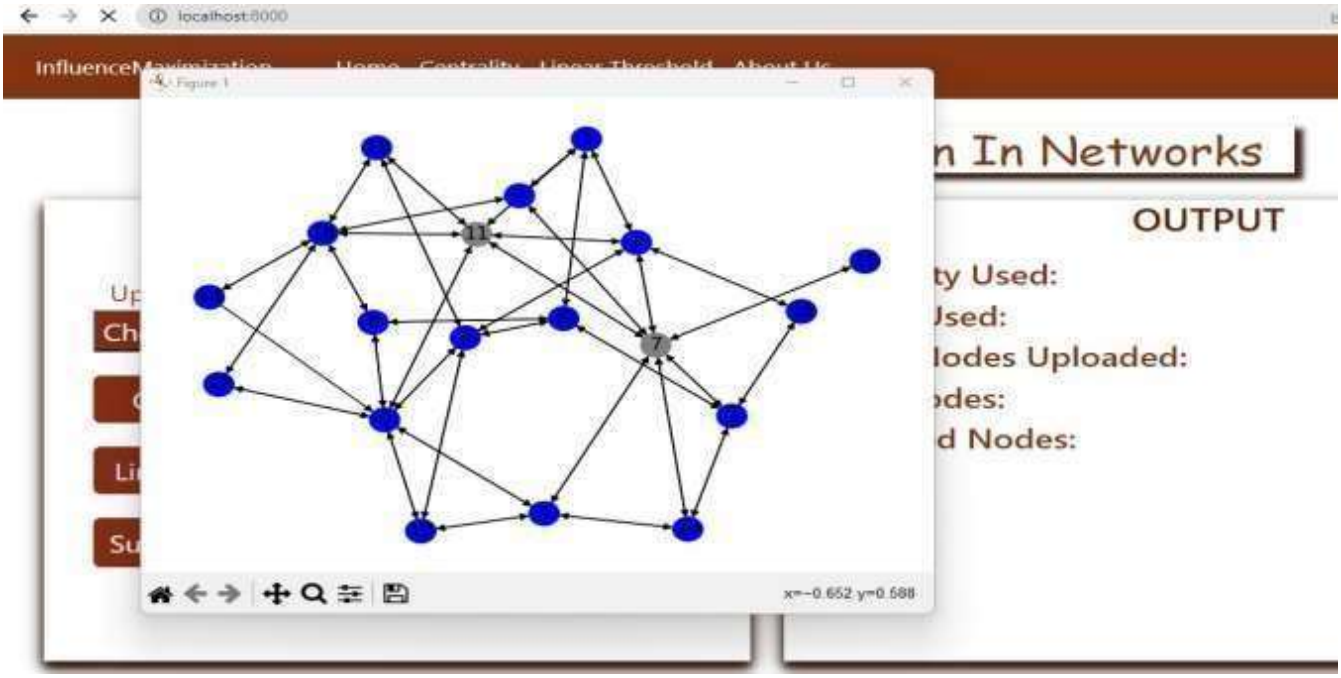


ACTIVATED NODES

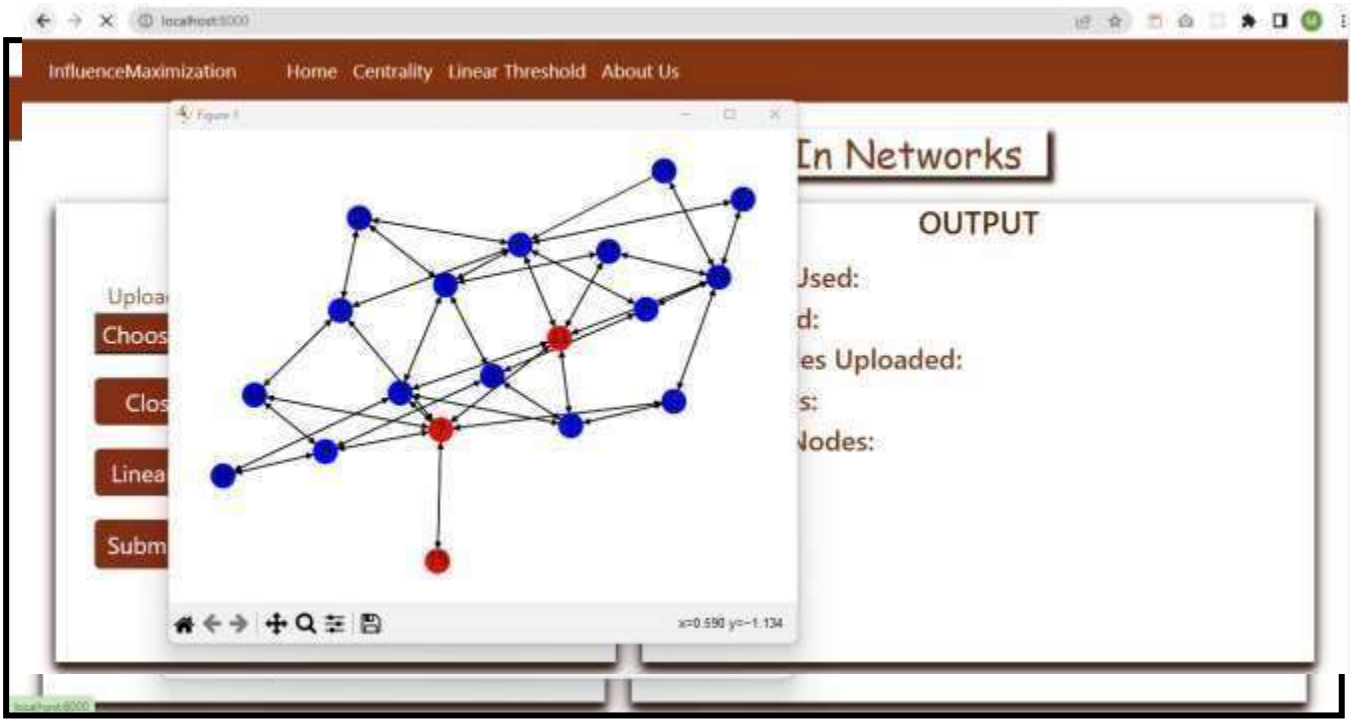


→CLOSENESS

SEED NODES



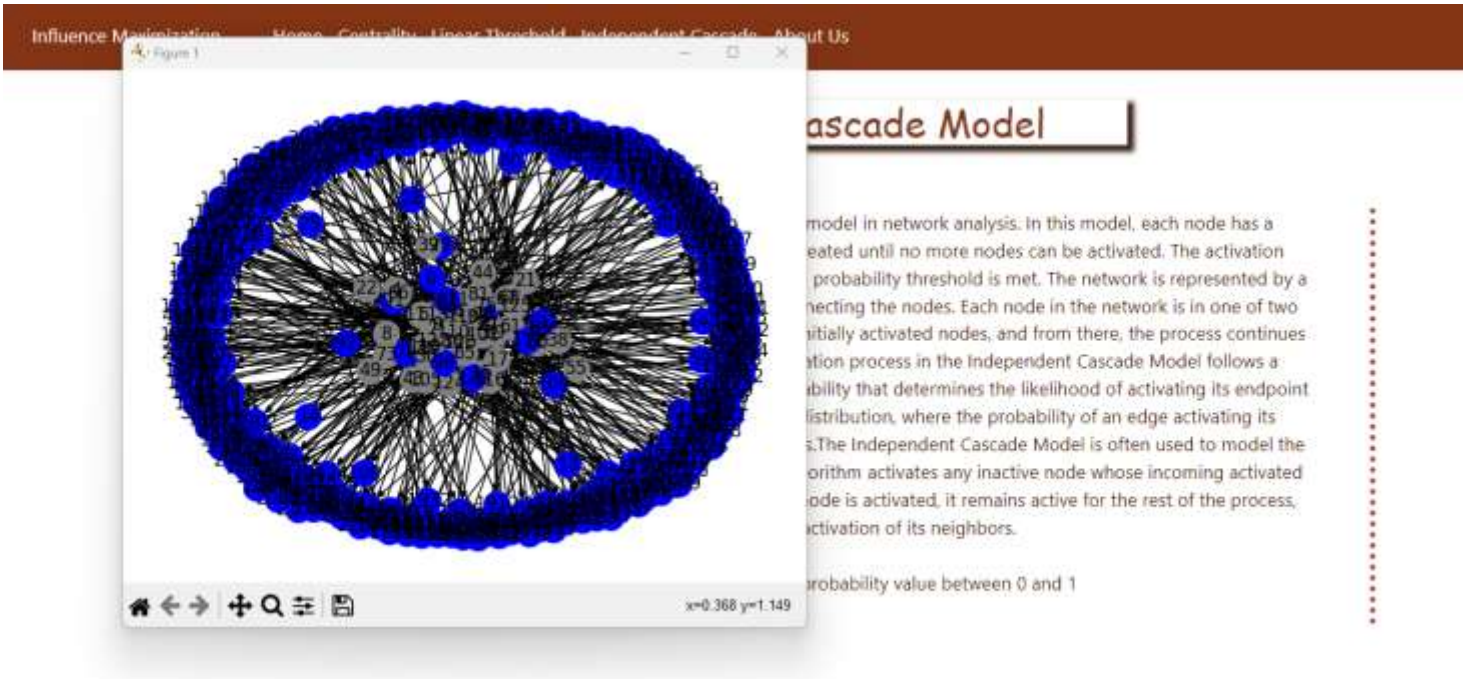
ACTIVATED NODES



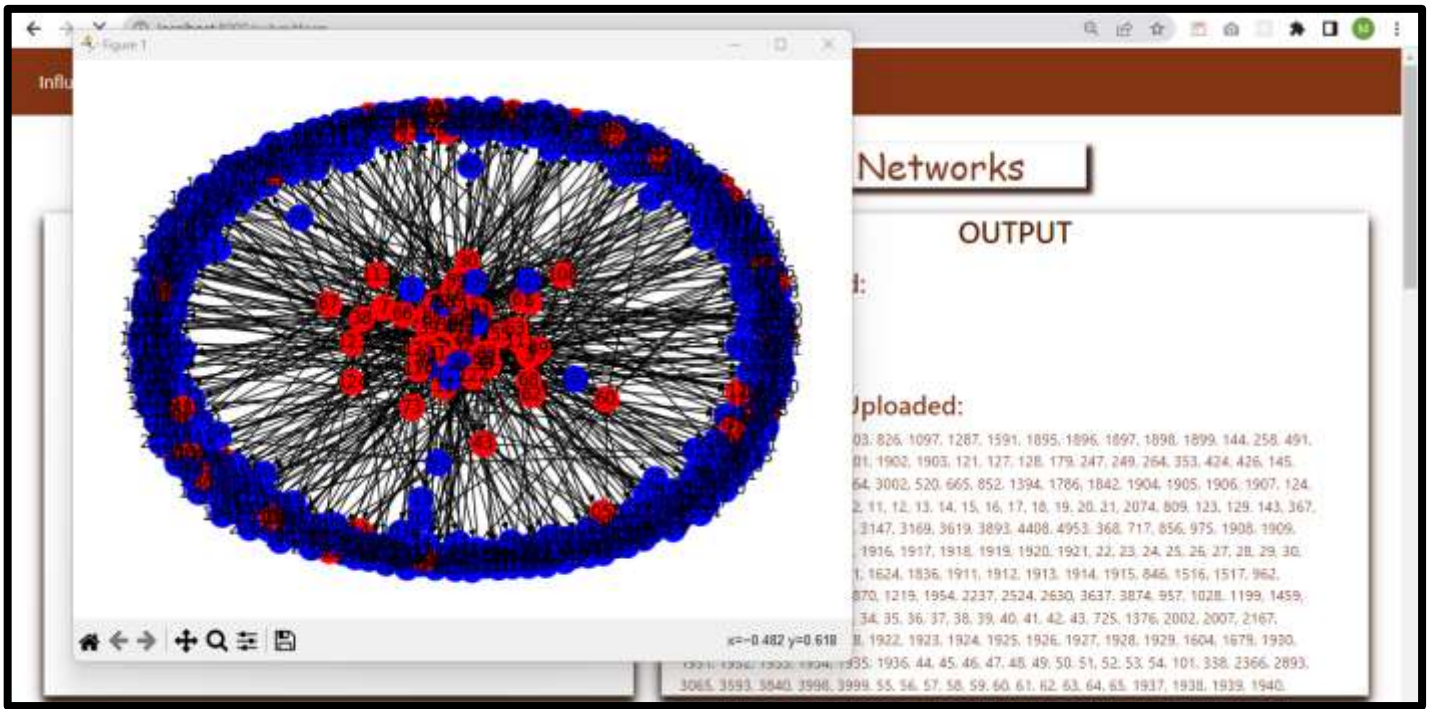
INDEPENDENT CASCADE MODEL

Edges=590
K=50

SEED NODES

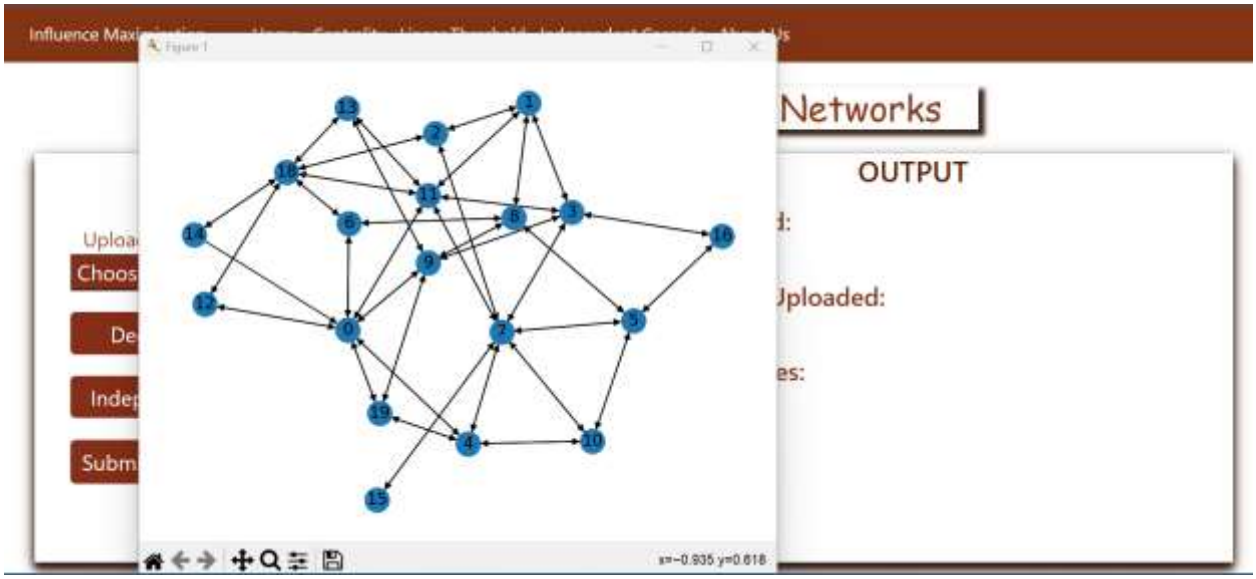


ACTIVATED NODES

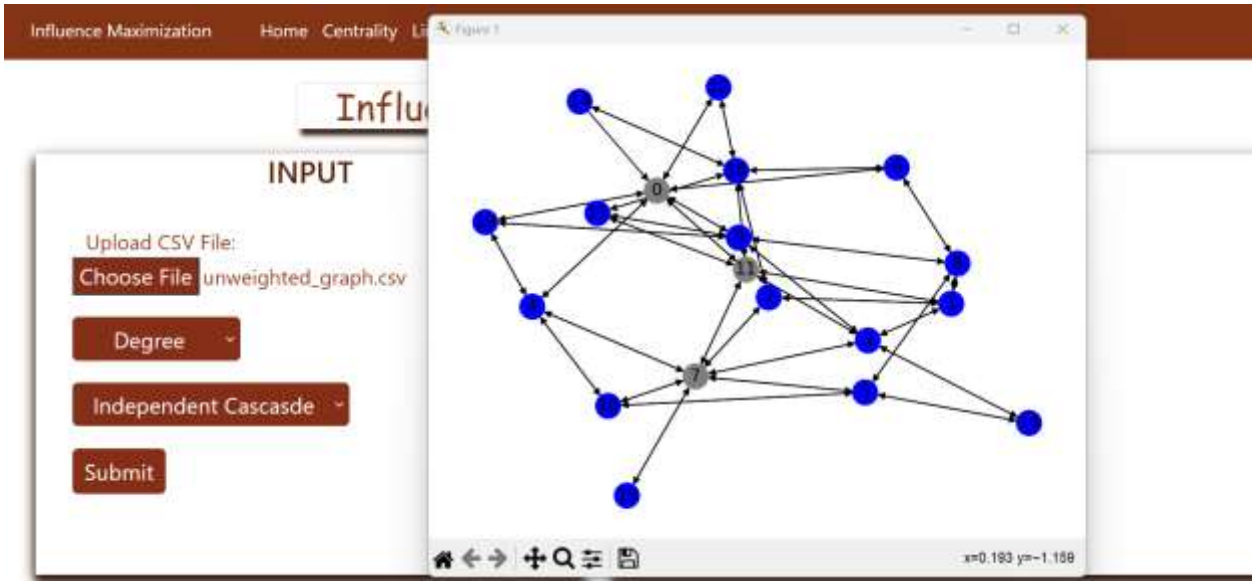


FOR A SMALLER NETWORK:

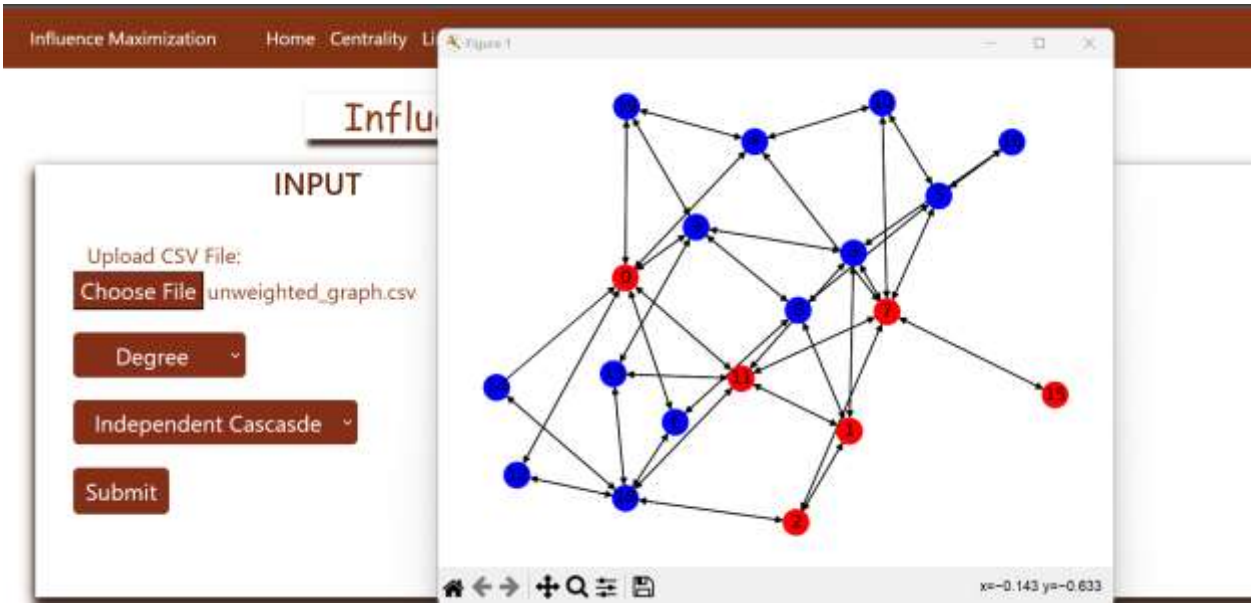
ACTUAL GRAPH



SEED NODES



ACTIVATED NODES



NAVBAR:


Influence Maximization
Home
Centrality
Linear Threshold
Independent Cascade
About Us
About Centrality
Centrality is a concept in network analysis that refers to the importance or influence of nodes within a network. It is a measure of how much a node is central or critical to the network structure, and it can provide insight into the flow of information, resources, and power within the network. Centrality measures can be used to identify key players or influential nodes within a network. For example, in a social network, a node with a high degree centrality may be a popular individual with many friends, while a node with a high betweenness centrality may be a mediator who connects different groups of people. Centrality measures can also be used to compare different networks and identify similarities or differences in their structure. By analyzing the centrality measures of different networks, researchers can gain insights into how networks evolve over time and how they respond to changes in their environment.
There are several types of centrality measures such as:
Degree Centrality
Closeness Centrality
Betweeness Centrality

Influence Maximization Home Centrality Linear Threshold Independent Cascade About Us

About Linear Threshold Model

The Linear Threshold Model (LTM) is a popular influence maximization model used in social network analysis. The LTM model starts by assigning a threshold value to each node in the network. Then, the influence or behavior is propagated through the network in a stepwise fashion. At each step, the model checks whether each node's threshold value has been met by the number of influenced neighbors. If a node's threshold has been met, the node becomes influenced and can in turn influence its neighbors in the next step. The process continues until no further nodes can be influenced, or until the model reaches a predefined stopping criterion.

In the LTM, each node in the network has a threshold value between 0 and 1 that represents the level of influence required from its neighbors to activate the node. When the sum of the weights of the active neighbors of a node exceeds its threshold value, the node becomes active and can influence its inactive neighbors.

The performance of these algorithms depends on the accuracy of the LTM in capturing the dynamics of influence propagation in the network. Several extensions to the LTM have been proposed, such as the weighted linear threshold model (WLTM), which allows nodes to have different degrees of influence on their neighbors, and the cascade model, which takes into account the temporal dynamics of influence propagation.

Influence Maximization Home Centrality Linear Threshold Independent Cascade About Us

About Independent Cascade Model

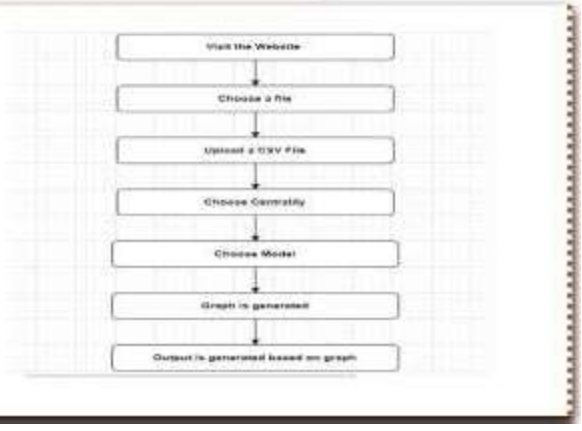
The Independent Cascade Model is another commonly used spread model in network analysis. In this model, each node has a probability of activating its neighboring nodes, and this process is repeated until no more nodes can be activated. The activation process is independent of the state of other nodes and occurs only if the probability threshold is met. The network is represented by a graph G(V, E), where V is the set of nodes, and E is the set of edges connecting the nodes. Each node in the network is in one of two states: activated or inactive. The activation process begins with a set of initially activated nodes, and from there, the process continues in discrete time steps until no more nodes can be activated.The activation process in the Independent Cascade Model follows a probabilistic rule. Each edge in the network is assigned a weight or probability that determines the likelihood of activating its endpoint nodes. The activation probability is typically modeled as a Bernoulli distribution, where the probability of an edge activating its endpoint nodes is a random variable with a certain probability of success.The Independent Cascade Model is often used to model the spread of information or disease in a network. At each time step, the algorithm activates any inactive node whose incoming activated neighbors have collectively exceeded the activation threshold. Once a node is activated, it remains active for the rest of the process, and its activation may influence the activation of its neighbors.

In the ICM, each node in the network has a probability value between 0 and 1

Influence Maximization Home Centrality Linear Threshold Independent Cascade About Us

About us

This website works on the idea of understanding the influence maximization on different networks. This is a learning platform where you can implement and visualise working of linear threshold model using various centrality measures . This website helps you learn about centrality measures like degree centrality , closeness centrality , betweenness centrality , Page Rank centrality and eigenvector centrality



LIST OF ALL TEST CASES

ERROR AND EXCPETION HANDLING

```
# make sure the seeds are in the graph
        for s in seeds:
            if s not in G.nodes():
                raise Exception("seed", s, "is not in graph")
```

```
# init thresholds
    for n in DG.nodes():
        if 'threshold' not in DG._node[n]:
            DG._node[n]['threshold'] = 0.5
        elif DG._node[n]['threshold'] > 1:
            raise Exception("node threshold:", DG._node[n]['threshold'], "cannot be larger than 1")

# init influences
    in_deg = DG.in_degree()
    for e in DG.edges():
        if 'influence' not in DG[e[0]][e[1]]:
            DG[e[0]][e[1]]['influence'] = 1.0 / in_deg[e[1]]
        elif DG[e[0]][e[1]]['influence'] > 1:
            raise Exception("edge influence:", DG[e[0]][e[1]]['influence'], "cannot be larger than 1")
```

```
# init activation probabilities
    for e in DG.edges():
        if 'act_prob' not in DG[e[0]][e[1]]:
            DG[e[0]][e[1]]['act_prob'] = 0.1
        elif DG[e[0]][e[1]]['act_prob'] > 1:
            raise Exception("edge activation probability:", \
            DG[e[0]][e[1]]['act_prob'], "cannot be larger than 1")
```

LIMITATIONS

1. It takes a long time to load the graph and data for bigger networks.
2. Takes only weighted graphs as inputs.
3. Cons of Linear Threshold Model:

   Limited to binary thresholds: The LT model assumes that each node in the network has a binary threshold, i.e. either it is influenced or not. This binary assumption may not be realistic in real-world scenarios, where individuals may have varying levels of susceptibility to influence.

   Assumption of fixed thresholds: The LT model assumes that each node has a fixed threshold, which does not change over time. However, in reality, a person's susceptibility to influence may change over time depending on various factors such as social context, mood, and personal experiences.

   Computationally expensive: The LT model involves multiple iterations of threshold updates for each node, which can be computationally expensive for large networks.

4. Cons of Independent Cascade Model:

   Limited to single activation: The IC model assumes that each node can only activate its neighbors once. This assumption may not be realistic in real-world scenarios, where individuals may be influenced by multiple sources or multiple times.

   Assumption of fixed probabilities: The IC model assumes that each edge has a fixed probability of influence, which does not change over time. However, in reality, the probability of influence may vary depending on the strength of the relationship or the context in which the influence occurs.

   May not always lead to optimal solutions: The IC model can sometimes lead to sub-optimal solutions in influence maximization, especially in scenarios where the network has a high degree of clustering.

FINDINGS

The choice of which centrality measure to use for influence maximization depends on the specific characteristics of the network and the goals of the analysis. Each of the centrality measures (degree, closeness, betweenness, PageRank, and eigenvector) has its own strengths and weaknesses, and the best choice depends on the specific context of the problem.

The LT model is generally considered to be more realistic as it allows for more nuanced and complex interactions between nodes, and can capture the idea that some nodes may have more influence or importance in the network than others. Additionally, the LT model is m

ore computationally efficient than the IC model, which can be important when working with large networks

The IC model assumes that each node in the network has a fixed probability of activating its neighboring nodes, while the LT model assumes that each node has a threshold value and will activate its neighbors if the sum of their weights exceeds this threshold.

CONCLUSION AND FUTURE SCOPE

Our project on influence maximization using degree, betweenness, closeness, PageRank, load, and eigenvector centrality measures, as well as the linear threshold and independent cascade models, has explored different ways to identify influential nodes in a network and simulate the spread of influence through the network.

The project has revealed that each centrality measure and diffusion model has its own strengths and limitations, and the choice of which measure and model to use depends on the specific characteristics of the network and the goals of the analysis. The project has also highlighted the importance of appropriate evaluation metrics to compare the performance of different centrality measures and diffusion models on the given network.

For future scope, researchers can consider exploring other centrality measures and diffusion models to improve the accuracy of influence maximization. They can also focus on developing new evaluation metrics to better capture the effectiveness of different approaches in identifying influential nodes in a network. Additionally, researchers can explore real-world applications of influence maximization, such as identifying key individuals in social networks for marketing or public health interventions. Finally, researchers can explore ways to optimize the performance of influence maximization algorithms to handle large-scale networks and improve the computational efficiency of the algorithms.

In the future, we would try to further implement our project for larger networks and explore more centrality measures and other models and algorithms.

REFERENCES

- https://towardsdatascience.com/graph-analytics-introduction-and-concepts-of-centrality-8f5543b55de3
- https://networkx.org/documentation/stable/reference/algorithms/centrality.html
- https://www.sciencedirect.com/science/article/abs/pii/S0020025520302395
- https://www.sciencedirect.com/science/article/abs/pii/S0950705117304975
- http://home.iitj.ac.in/~suman/articles/detail/how-to-code-independent-cascade-model-of-information-diffusion/
- https://bookdown.org/markhoff/social_network_analysis/centrality.html