Influence Maximization Using Community based Approach

MASTERS IN COMPUTER APPLICATION

DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY

University Of North Bengal

PROJECT REPORT

Submitted by
Arijit Kumar Kundu
Reg. no: 2082007030015
Charanraj Laya
Reg. no:2082007030005

Certificate

This is to certify that Arijit Kumar Kundu & Charanraj Laya have completed their project work titled “Influence Maximization Using Community based Approach”, under the direct supervision guidance of Mr. Anil Tudu, Department of Computer Science and Technology. I am satisfied with their work, which is being presented for the partial fulfillment of the degree of Master of Computer Application (MCA). North Bengal University.

Signature of the Guide,

Anil Tudu

(Assistant Professor)

Student's Declaration

We declare that this project titled "Influence Maximization Using Community based Approach", submitted as requirement for the award of degree of Master in Computer Application, does not contain any material previously submitted for a degree in any university, and that to the best of our knowledge, it does not contain any materials previously published or written by another person except where due reference is made in the text. We understand that the management of Department of Computer Science Technology, North Bengal University, has a zero-tolerance policy towards plagiarism.

_____

Signature

(Arijit Kumar Kundu)

_____

Signature

(Charanraj Laya)

Acknowledgment

We express our sincere thanks and gratitude to our guide Mr. ANIL TUDU, Assistant professor, Department of Computer Science and Technology, University of North Bengal for his guidance and suggestions. Without his disposition, spirit of accommodation, frankness and timely clarifications this project would have been incomplete in due time.

We also express our gratitude to all the faculty members, and our fellow mates who have helped us to carry out this work. Last but not the least, we thank our almighty God for His blessing showed on us during this period.

Project Team Members
Arijit Kumar Kundu
Charanraj Laya
MCA 4th Semester

INDEX

INTRODUCTION

A social network is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, memes spread, information circulation, friendship and acquaintance networks, business networks, knowledge networks, difficult working relationships, social networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines. These visualizations provide a means of qualitatively assessing networks by varying the visual representation of their nodes and edges to reflect attributes of interest.

Influence maximization is a problem in the field of computer science and social network analysis that seeks to identify the subset of nodes in a network that will maximize the spread of a particular influence or attribute over the network. This problem has applications in various fields, such as marketing, public health, and political campaigns, where it is important to identify the most influential individuals or groups in a network in order to maximize the reach of a message or the adoption of a particular behavior.

A community in a social network is a group of people who share common interests and interact with each other regularly through the platform. These communities can be formed around a wide range of topics, such as hobbies, professions, political views, or personal interests. Members of a community often communicate with each other by posting updates, sharing content, and commenting on each other's posts. Some social networks also allow members to join or create groups, which are specialized communities that focus on a specific topic or activity. Communities in social networks can be a great way for people to connect with others who have similar interests, share information and ideas, and support each other.

OBJECTIVE

Influence maximization refers to the problem of finding the most effective way to spread information or influence within a social network. The main objective of influence maximization is to identify the most influential individuals in a network and target them with a message or information in order to reach the largest possible audience.

There are several approaches that can be used to maximize influence in a social network, such as identifying the most central or influential nodes in the network, using targeted advertising or marketing campaigns, and using viral marketing techniques to encourage people to share the message with their own networks.

The ultimate goal of influence maximization is to use the social network to effectively disseminate information or influence the attitudes or behaviours of a large number of people in a cost-effective manner. This can be useful for a variety of purposes, such as promoting a product or service, raising awareness about a social issue, or influencing political opinions.

Proposed Model Description and methodology

Step-1: Louvain Community Detection:
Step-2: Locate core nodes using K Shell algorithm
Step-3: Calculate Extended H index of the core nodes

Louvain Community Detection :

Louvain is an unsupervised algorithm (does not require the input of the number of communities nor their sizes before execution) divided in 2 phases:

- Modularity Optimization and
- Community Aggregation

After the first step is completed, the second follows. Both will be executed until there are no more changes in the network and maximum modularity is achieved

$$M = \frac{1}{2m}\sum_{i,j}(A_{ij} - p_{ij})\delta(c_i, c_j) = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \quad (1)$$

Aij is the adjacency matrix entry representing the weight of the edge connecting nodes i and j, $ki = \sum j\ Aij$ is the degree of node i, ci is the community it belongs, δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise. $m = 1\ \sum ij\ Aij$ 2 is the sum of the weights of all edges in the graph.

Modularity Optimization:

Louvain will randomly order all nodes in the network in Modularity Optimization. Then, one by one, it will remove and insert each node in a different community C until no significant increase in modularity (input parameter) is verified:

$$\Delta M = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (2)$$

Let Σin be the sum of the weights of the links inside C, Σtot the sum of the weights of all links to nodes in C, ki the sum of the weights of all links incident in node i, ki,in the sum of the weights of links from node i to nodes in the community C and m is the sum of the weights of all edges in the graph.

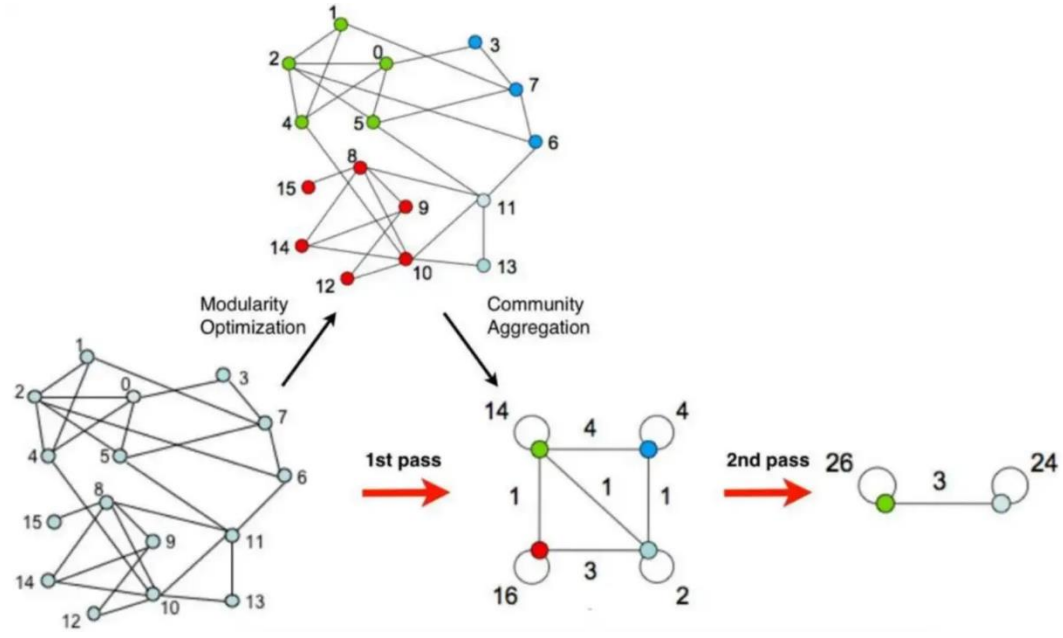
One way to further improve the performance of the algorithm is by simplifying (2) and calculating ΔMm instead of the complete expression:

$$\Delta M = \frac{k_{i,in}}{m} - \frac{2\sum_{tot} k_i}{(2m)^2} \Leftrightarrow \Delta Mm = k_{i,in} - \frac{\sum_{tot} k_i}{2m} \quad (3)$$

While ki, in and Σtot need to be calculated for each trial community, $ki/(2m)$ is specific of the node that is being analyzed. This way, the latter expression is only recalculated when a different node is considered in Modularity Optimization.

Community Aggregation:

After finishing the first step, all nodes belonging to the same community are merged into a single giant node. Links connecting giant nodes are the sum of the ones previously connecting nodes from the same different communities. This step also generates self-loops which are the sum of all links inside a given community, before being collapsed into one node (Figure 1).



Thus, by clustering communities of communities after the first pass, it inherently considers the existence of a hierarchical organization in the network. Pseudocode in Algorithm 1.

Algorithm 1 Louvain

Require:
$G^0 = (V^0, E^0)$: initial undirected graph. V^0 is the initial set of vertices. E^0 is the initial set of edges;
θ: modularity improvement threshold.

Ensure:
M: resulting module;
Mod: resulting modularity;
δMod: modularity variation using Equation 4;
k_i: sum of the weight of all edges connecting node i;
$k_{i,j}$: weight of the edge connecting nodes i and j.

$m = \sum k_{i,j}, (i,j) \in E^0$
$k = 0$ // Iteration number.
repeat
  // Attributing a different community to each node.
  for all $i \in V^k$ **do**
    $M_i^k = \{i\}$
  end for
  Compute $Mod_{new} = Mod(M)$ using Equation 3
  repeat
    $Mod = Mod_{new}$
    Randomize the order of vertices.
    for all $i \in V^k$ **do**
      $best\_community = M_i^k$
      $best\_increase = 0$
      for all $M' \in C^k$ **do**
        $M_i^k = M_i^k \backslash \{i\}$
        ${\Sigma_{tot}}^{M_i^k} = \sum_{\alpha \in M_i^k} k_\alpha - k_i$; ${\Sigma_{tot}}^{M'^k_i} = \sum_{\alpha \in M'^k_i} k_\alpha + k_i$;
        $\Sigma_{in} = {\Sigma_{tot}}^{M'^k_i} - \sum k_{i,j}, (i,j) \in E^k, i \in C_i^k \; and \; j \notin C_i^k$
        $k_{i,in} = \sum_{\alpha \in M'^k_i} k_{i,\alpha}$
        If $\delta Mod_{M_i^k \to M'^k_i} > best\_increase$ **then**
          $best_{increase} = \delta Mod_{M_i^k \to M'^k_i}$
          $best_{com} = M'^k_i$
          $M'^k_i = M'^k_i \cup \{i\}$
        else
          $M_i^k = M_i^k \cup \{i\}$
        end else
      end for
    end for
    Compute $Mod_{new} = Mod(M)$ using Equation 3
  until No vertex movement **or** $Mod_{new} - Mod < \theta$
  // Calculate updated modularity
  Compute $Mod_{new} = Mod(M)$ using Equation 3
  If $Mod_{new} - Mod < \theta$ **then**
    break;
  end if
  $Mod = Mod_{new}$
  // Merge communities into a new graph
  $V^{k+1} \leftarrow C^k$
  $E^{k+1} \leftarrow e(C_u^k, C_v^k)$
  $G^{k+1} = (V^{k+1}, E^{k+1})$
  $k = k + 1$
until break

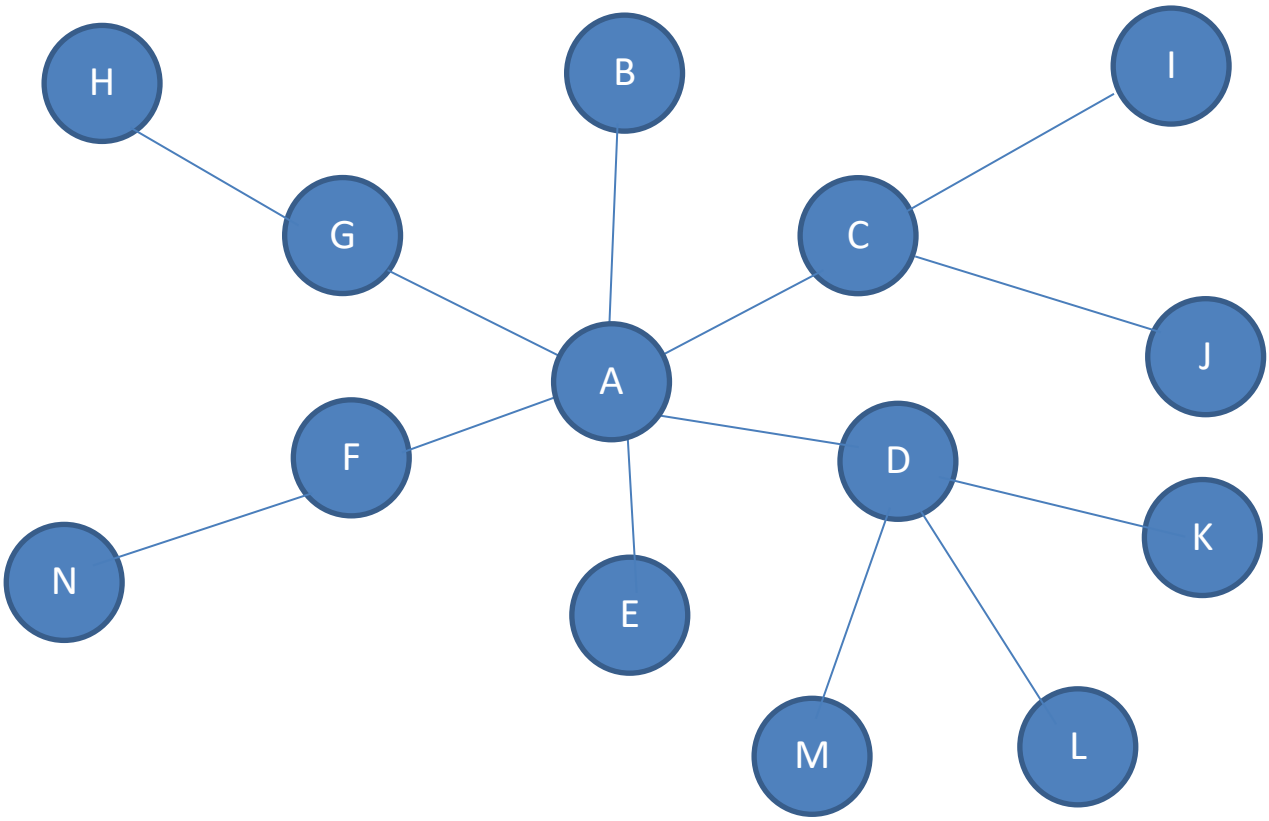
H index:

The Hirsch index or h-index is defined as the maximum value of h such that the given author/journal has published at least h papers that have each been cited at least h times. The index is designed to improve upon simpler measures such as the total number of citations or publications. The index works best when comparing scholars working in the same field, since citation conventions differ widely among different fields

$$h\text{-index}(f) = \max\{ i \in \mathbb{N} : f(i) >= i$$

Extended H Index:

The extended h-index, also known as the h-index 2, is a variant of the h-index that takes into account the number of citations received by each paper published by a researcher. The extended h-index is calculated by counting the number of papers that a researcher has published and then finding the highest number h such that h of those papers have been cited at least h times. However, instead of simply counting the number of citations for each paper, the extended h-index weights the citations according to the number of times each paper has been cited.



Extended H index of (A) = H index of (**A**) + H index of (**B**) + H index of (C) + H index of (**D**) + H index of (**E**) + H index of (**F**) + H index of (**G**)

Independent Cascade Model:

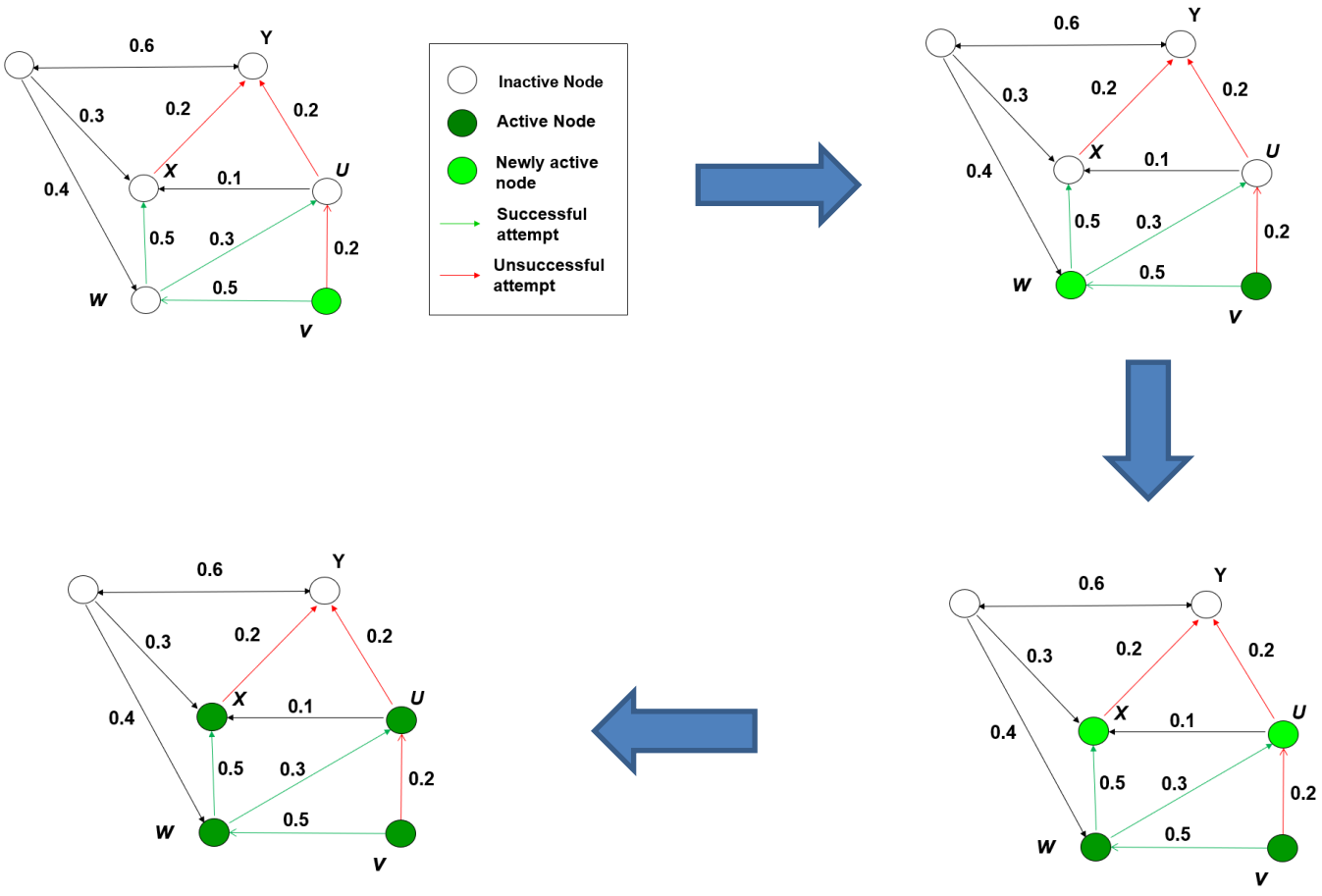
A Independent Cascade Model (ICM) is a stochastic information diffusion model where the information flows over the network through Cascade. Nodes can have two states, (i) Active: It means the node already influenced by the information in diffusion. (ii) Inactive: node unaware of the information or not influenced.

The process runs in discrete steps. At the beginning of ICM process, few nodes are given the information known as seed nodes. Upon receiving the information these nodes become active. In each discrete step, an active node tries to influence one of its inactive neighbors. In spite of its success, the same node will never get another chance to activate the same inactive neighbor. The success depends on the propagation probability of their tie. Propagation Probability of a tie is the probability by which one can influence the other node. In reality, Propagation Probability is relation dependent, i.e., each edge will have different value. However, for the experimental purpose, it is often considered to be same for all ties.

The process terminates when no further nodes became activated from inactive state. We can understand through a simple explanation:

- When node v becomes active, it has a single chance of activating each currently inactive neighbor w.
- The activation attempt succeeds with probability p_{vw}.
- The deterministic model is a special case of IC model. In this case, $p_{vw}=1$ for all (v,w).

Example

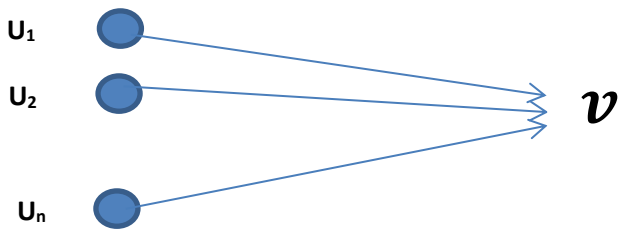


If a node v has k (coming) neighbors $u_1, u_2, \ldots, u_k$,then following k possible events are independent:

(1) u_1 makes v active with probability $p_{u_1 v}$

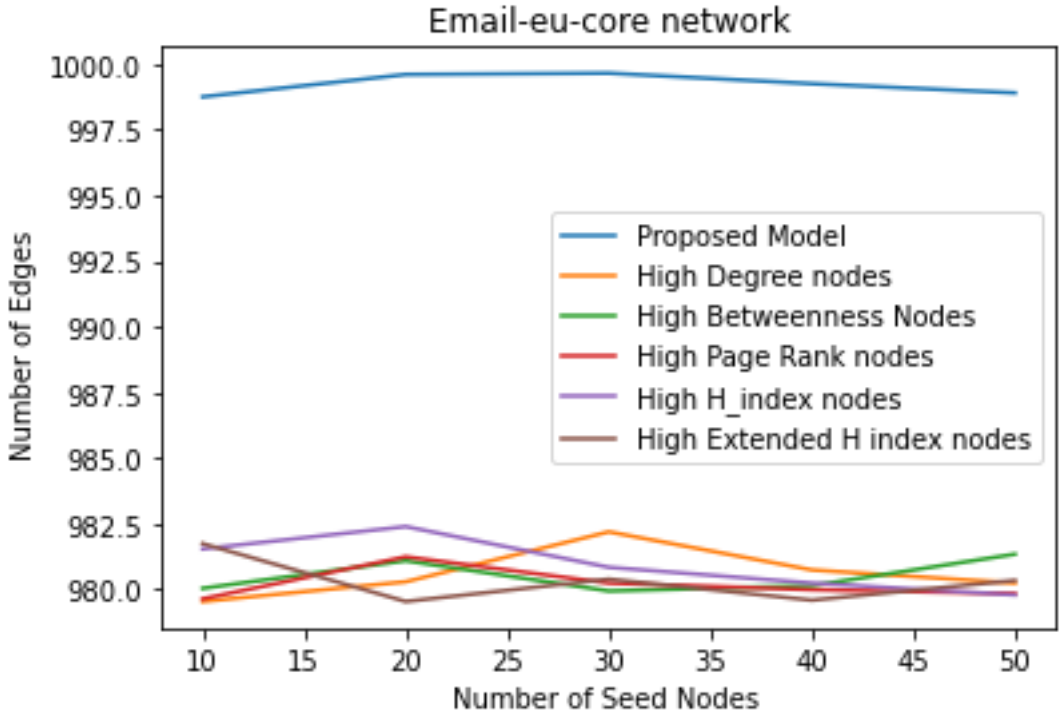
$\vdots$

(k) u_k makes v active with probability $p_{u_k v}$

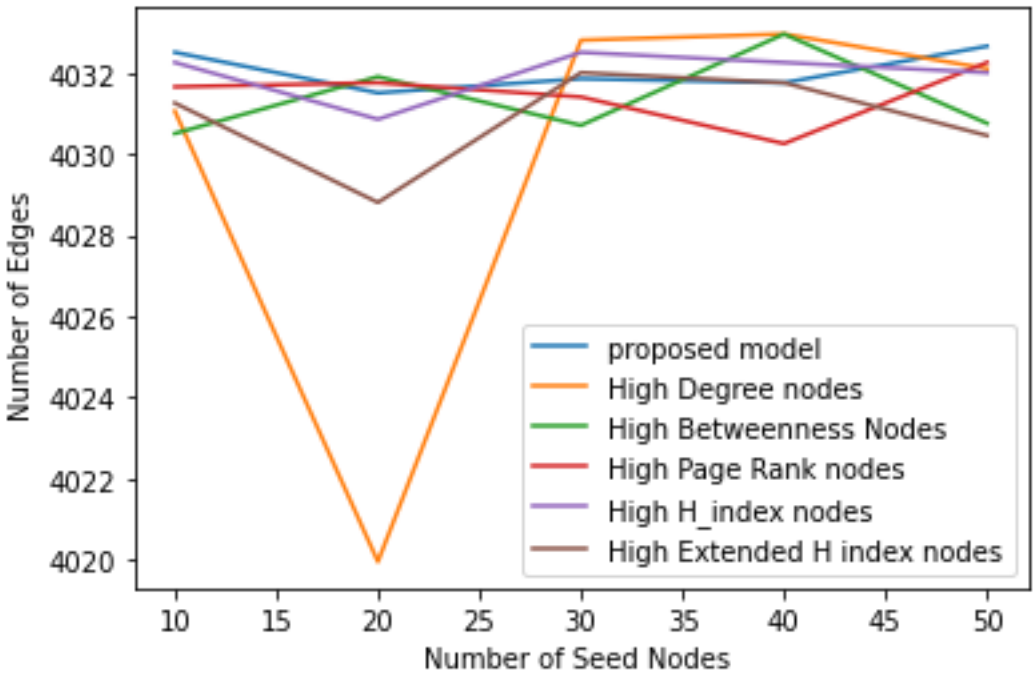


Datasets that are used in the proposed Model: Facebook dataset, Email-eu-core network

The proposed model then compared with High **degree** nodes, High **betweenness** nodes, high **page rank** nodes, high **H index** nodes and high **Extended H index** nodes. The comparison graph looks like:



Facebook Dataset



Discussion

From the upper two graphs we can observe that the proposed model's graph tends to go higher than the rest of the methods so we can it is highly effective for propagating influence over a network. But we can't say this is the best model as several other models exist that isn't considered here. Some potential directions for improvement include:

1. **Improved algorithms:** Researchers are working on developing more advanced algorithms and techniques for identifying the most influential nodes in a network and targeting them with a message or information. These algorithms may be able to better capture the complex dynamics of social networks and more accurately predict how information will spread.
2. **Real-time analysis:** As social media platforms become increasingly fast-paced, there is a need for models that can analyse and predict the spread of information in real-time. Developing models that can quickly adapt to changing circumstances and identify the most influential nodes in real-time could be a key area of focus.
3. **Multi-layered networks:** Many social networks are not just simple networks, but are actually made up of multiple layers or types of connections. Future models may be able to more accurately capture and analyse the dynamics of these multi-layered networks.
4. **Integration with other technologies:** Influence maximization models may be integrated with other technologies, such as artificial intelligence or virtual reality, to create more sophisticated and immersive social media experiences

Conclusion

The accuracy and effectiveness of an influence maximization model depends on the quality of the data used to build it, as well as the algorithms and techniques employed. There are many different approaches that can be used to maximize influence in a social network, and the most appropriate approach will depend on the specific goals and needs of the model.

As social media platforms and the ways in which people use them evolve, influence maximization models will also need to evolve in order to remain relevant and effective. Further research and development in this field is likely to lead to the development of more advanced and sophisticated models that are better able to capture the complexity and dynamics of social networks.

Reference

https://towardsdatascience.com/louvain-algorithm-93fde589f58c

https://beckerguides.wustl.edu/authors/hindex

https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank_numpy.html

https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.betweenness_centrality.html

https://networkx.org/documentation/stable/reference/classes/generated/networkx.Graph.degree.html

http://home.iitj.ac.in/~suman/articles/detail/how-to-code-independent-cascade-model-of-information-diffusion/#:~:text=Independent%20Cascade%20Model%20(ICM)%20is,the%20information%20or%20not%20influenced.

https://en.wikipedia.org/wiki/H-index

https://en.wikipedia.org/wiki/Social_network

https://snap.stanford.edu/data/email-Eu-core.html